

Ivanova, Daniela (2025) *Machine learning for analogue media damage restoration.* PhD thesis.

https://theses.gla.ac.uk/85461/

Copyright and moral rights for this work are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This work cannot be reproduced or quoted extensively from without first obtaining permission from the author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Enlighten: Theses
https://theses.gla.ac.uk/
research-enlighten@glasgow.ac.uk

Machine Learning for Analogue Media Damage Restoration

Daniela Ivanova

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Computing Science
College of Science and Engineering
University of Glasgow



March 2025

Abstract

Analogue media degradation presents a unique challenge for digital restoration, requiring the disentanglement of physical damage from intended content and style. This thesis investigates machine learning approaches for detecting and segmenting damage across diverse media types, addressing a critical bottleneck in restoration workflows.

We first focus on the constrained setting of a single type of analogue media, and develop a statistical framework for modeling film damage. Through perceptual studies, we demonstrate that our approach generates training data indistinguishable from authentic damage. Using this data to train supervised models yields significant improvements in damage detection performance, as validated on our benchmark dataset of professionally restored high-resolution film scans.

Expanding beyond film, we introduce ARTeFACT, the first comprehensive dataset for analogue media damage detection comprising over 11,000 pixel-accurate annotations across 15 damage categories and 10 diverse media types. Systematic evaluation reveals that state-of-the-art supervised segmentation methods, including foundation models like Segment Anything, fail to generalize across different media and struggle to disambiguate damage from visually similar content features.

To address these limitations, we investigate the semantic understanding capabilities of text-to-image diffusion models. We develop a novel unsupervised , zero-shot semantic segmentation framework leveraging self-attention mechanisms in Stable Diffusion, achieving state-of-the-art performance on standard benchmarks. Through matrix exponentiation of attention maps, we provide a principled control mechanism for segmentation granularity based on random walk theory. Such mechanism is crucial for enabling segmentation of artefacts at varying granularity.

Our findings demonstrate that effective analogue media damage detection requires moving beyond rudimentary pattern recognition toward semantic understanding of content-damage relationships. This work establishes a methodological foundation for automated restoration systems that can support heritage preservation at scale while respecting the unique characteristics of different analogue media. The challenges presented by damage detection further illuminate fundamental questions about representation learning and visual-semantic disentanglement that are significant for advancing machine learning beyond pattern recognition toward more meaningful visual understanding.

Contents

Abstract **i**

Acknowledgments **xiv**

Declaration **xv**

List of papers **xvi**

List of datasets **xvii**

Note on the format of this thesis **xviii**

1 Introduction **1**

1.1 Restoration: Dual Perspectives 2
1.2 Research Challenges 3
1.3 Restoring Analogue Damage: A Computational Approach 4
1.4 Motivation 5
1.4.1 Conservationist 5
1.4.2 Practical 5
1.4.3 Personal 6
1.5 Thesis statement 6
1.5.1 Research Questions 6
1.5.2 Scope 7
1.6 Contributions 7
1.7 Thesis Walkthrough 8

2 Background **9**

2.1 Image Restoration: Cultural Dimensions 10
2.1.1 The Medium is the Message 10
2.1.2 Taxonomies of Damage 12
2.1.3 Perception of Content vs. Medium 13
2.1.4 The Digital Advantage 13

2.2 Image Restoration: A Technical Perspective 14
2.2.1 Historical context 14
2.2.2 Mathematical formulation 15
2.2.3 DSP: Techniques and limitations 16
2.2.4 Machine Learning Advancements 19
2.3 Analogue Media Restoration in the Digital Domain 22
2.3.1 Understanding Analogue Damage 22
2.3.2 Existing Approaches to Analogue Media Restoration 23
2.3.3 Limitations 25
2.3.4 Decomposing Analogue Restoration: Detection and Inpainting 26
2.3.5 Related Detection Tasks 27
2.3.6 Metrics 29
2.3.7 The data challenge 29
2.4 Preliminary work in film damage restoration 32
2.4.1 Data challenges 32
2.4.2 Comparative evaluation 32
2.4.3 Takeaways 34

3 Simulating analogue film damage to analyse and improve artefact restoration on high-resolution scans 36
3.1 Summary 37
3.2 Abstract 38
3.3 Introduction 39
3.4 Background and related work 41
3.4.1 Film damage 41
3.4.2 Inpainting 42
3.4.3 Processing of high-resolution data 43
3.4.4 Deep learning for analogue film restoration 43
3.5 Evaluation data for damage restoration 44
3.6 Modelling analogue film damage 44
3.6.1 Capturing real analogue film artefacts 45
3.6.2 Analysing artefact statistics 48
3.6.3 Synthetic artefacts. 48
3.6.4 Generating damage overlays 49
3.6.5 Human validation of generated artefact damage 51
3.7 Experiments 52
3.7.1 Comparison against existing damage simulations 53
3.7.2 Comparison of restoration models 54
3.7.3 Restoration results and discussion 59

3.8 Conclusions . . . 60
3.9 Afterword . . . 61
3.10 Supplementary . . . 62

4 State-of-the-Art Fails in the Art of Damage Detection 66
4.1 Summary . . . 66
4.2 Abstract . . . 69
4.3 Introduction . . . 70
4.4 Related work . . . 71
4.5 Damaged Analogue Media Dataset . . . 72
4.5.1 Annotation methodology . . . 72
4.5.2 Damage Types . . . 73
4.5.3 Image-level Annotations . . . 74
4.6 Benchmark . . . 75
4.6.1 Zero-shot Segmentation with SAM . . . 75
4.6.2 Supervised Semantic Segmentation . . . 76
4.6.3 Diffusion-based segmentation . . . 77
4.6.4 Results . . . 78
4.7 Conclusion . . . 81
4.8 Afterword . . . 82
4.9 Supplementary . . . 83
4.9.1 Data collection & curation . . . 83
4.9.2 Image-level caterogorisation and textual descriptions . . . 85
4.9.3 Qualitative assessment . . . 86
4.9.4 Zero-shot generative segmentation . . . 88

5 Unsupervised Segmentation by Diffusing, Walking and Cutting 102
5.1 Summary . . . 102
5.2 Abstract . . . 104
5.3 Introduction . . . 104
5.4 Related Work . . . 106
5.4.1 Unsupervised Segmentation via Generative Models . . . 106
5.4.2 Normalised cuts . . . 107
5.5 Methods . . . 108
5.5.1 Extracting diffusion features . . . 109
5.5.2 Self-attention defines a random walk . . . 109
5.5.3 Adjacency matrices from feature similarity . . . 110
5.5.4 Applying normalised cuts (NCut) . . . 110
5.5.5 Adapting MAN-C . . . 110

5.5.6 Segmentation aggregation and upsampling . . . 111
5.6 Experiments . . . 111
5.6.1 Random Walk Baseline . . . 112
5.6.2 Self-Attention adjacency . . . 113
5.6.3 Hyperparameter-free NCut . . . 114
5.6.4 Power Walks . . . 115
5.6.5 Comparison with state-of-the-art . . . 116
5.7 Conclusions . . . 118
5.8 Afterword . . . 119
5.9 Supplementary . . . 120
5.9.1 A note on evaluation strategies . . . 120
5.9.2 Ablation over timestep for feature extraction . . . 121
5.9.3 Additional qualitative examples . . . 121

6 Conclusion **126**
6.1 Summary of Contributions . . . 126
6.2 Evolution of Computational Damage Detection . . . 128
6.2.1 DSP Limitations: Single-image, explicit modeling . . . 128
6.2.2 Machine Learning Advances: Intra-medium generalization . . . 128
6.2.3 Cross-Media Challenges & The limits of visual learning . . . 130
6.2.4 Semantic Understanding: The promise of multiple modalities . . . 132
6.3 Outlook . . . 134
6.4 Limitations . . . 136
6.4.1 Data quality & biases . . . 136
6.4.2 Processing real-world data . . . 137
6.4.3 Statistical modeling . . . 137
6.4.4 Evaluation . . . 137
6.4.5 Damage as a semantic concept . . . 138
6.5 Future Directions . . . 138
6.5.1 Damage-Specific Semantic Segmentation via Text . . . 138
6.5.2 Technical advancements . . . 139
6.5.3 Theoretical advancements . . . 140
6.5.4 Cross-Disciplinary Applications . . . 140
6.6 Practical Applications . . . 141
6.6.1 Accelerating Conservation Workflows . . . 141
6.6.2 Interactive Access to Cultural Heritage . . . 142
6.6.3 Digital Surrogates with Enhanced Metadata . . . 142
6.6.4 Democratizing Access to Restoration . . . 142
6.7 Final Remarks . . . 143

List of Tables

2.1 Analogue media damage datasets predating this thesis. Most are small, qualitative, or lack annotations. Only two are made publicly available by the respective authors. 31

2.2 SSIM Scores obtained by a simple MSE baseline, "Bringing Old Photos Back to Life" [204], and our preliminary perceptual loss approach [94]. 34

3.1 Statistics of annotated artefacts – we report the total count and average area for each artefact class. See Figure 3.6 for the full distributions of areas. 49

3.2 Summary of participant scores (in %) based on responses to the Real vs Synthetic Damage survey. A score of 50% is optimal since it indicates that our synthetic damage is on average indistinguishable from real damage. 52

3.3 Comparison between the predicted segmentations for the evaluated models trained on different simulated damage. 54

3.4 Quantitative evaluation of restoration results on our high resolution dataset of authentic film damage. 55

4.1 Results from benchmarking SAM at zero-shot damage segmentation across various prompt modes. N corresponds to average number of prompts (i.e. points, bounding boxes, text) required as input. Point grid was evaluated with 64 points, sampled uniformly [108]. 76

4.2 Results across both Material and Content splits, for all baselines, trained on the task of binary damage segmentation. Best F1 scores for each test class in bold. Last row is a stratified split over Material x Content. 79

4.3 Results across both Material and Content splits, for all baselines, trained on the task of multiclass damage segmentation. Best F1 scores for each test class in bold. Last row is a stratified split over Material x Content. 79

4.4 Results across both Material and Content splits diffusion based text-guided (DiffEdit) and unsupervised (DiffSeg) binary damage segmentation. 80

4.5 Cosine distances between the CLIP embeddings of image descriptions generated by LLaVA and our descriptions, averaged across Material and Content classes. . 86

4.6 Weighted average metrics per damage types across Material categories for each model at the task of multiclass semantic segmentation of damage. 87
4.7 Weighted average metrics per damage types across Content categories for each model at the task of multiclass semantic segmentation of damage. 87
4.8 Results across both Material and Content splits for DiffEdit binary mask generation in both center cropped and full resolution settings. 89

5.1 Quantitative results on COCO-Stuff-27 between Normalised Cuts performed on adjacency matrices build as dot product vs cosine distance between patch self-attention across three different NCut thresholds. 113
5.2 Comparison of our automatic threshold methods across different attention resolution combinations on Coco-Stuff-27. MAN-C NCut consistently outperforms both Scaled MinCut and state-of-the-art DiffSeg. Random Walk baseline has NCut threshold set to 0.3. 115
5.3 Comparison for a range exponents (random walk steps k) applied to the self-attention transition matrix P before constructing the adjacency matrix A. We employ Scaled MinCut for automatic threshold selection, considering all available resolutions. 116
5.4 Quantitative comparison of mIoU between our approach and current state-of-the-art on COCO-Stuff-27, Cityscapes and ADE20K. Best in bold, second best underscored. 117
5.5 Comparison with DiffCut (mIoU) on COCO-Stuff-27. 117
5.6 We calculate metrics on predictions from Coco-Stuff-27 validation set under three different settings: globally averaged, per-image averaged, and per-image averaged with an oracle which merges oversegmented regions; our approach excels in all three. 121
5.7 Ablation over noising step and importance of DDIM Inversion (mIoU) on Coco-Stuff-27. 121

List of Figures

1.1 Vermeer's *The Love Letter* was crudely cut from its frame and stolen. Upon recovery, an international expert committee at Rijksmuseum spent a year restoring it [110]. 1

2.1 Ashley-Smith's framework for classification of a conservation object's possible alterations. Reproduced from [142]. 10
2.2 From left to right: the original version of the fresco; a deteriorated version; the restored version. Reproduced from [147]. 11
2.3 Gaze scanpaths for the Neutral Infill restoration (left) and full restoration (right) of the painting. Dots represent fixations; lines represent saccades. Reproduced from [130]. 13
2.4 Knowing exactly how flat the primary mirror of the Hubble Space Telescope (HST) was allowed scientists to model the resulting spherical aberration and invert it. 15
2.5 Common digital degradations shown in the top row can be modeled within the $y = H(f) + n$ framework. Example restorations via standard methods shown in bottom row. These degradations affect either the light path before capture (optical blur, motion blur) or the digital signal itself (spatial quantization, additive noise), and can be modeled with well-understood operators. They do not affect the medium itself pre-digitisation. Image from NASA Flickr Archive. 17
2.6 Example of frequency domain filtering in digital image restoration. The transfer function in the frequency domain (middle column) is applied to remove periodic noise patterns from the original images (left column), resulting in the restored images (right column). 18
2.7 Degradations solved by machine learning in the top row can also be modeled within the $y = H(f) + n$ framework. Example restorations shown in bottom row. These degradations are much more complex, but they still occur in the digital domain. Image from NASA Flickr Archive. 21

2.8 Example inpaintings with LaMa [192] and Stable Diffusion Inpainting [167] across several different types of media with damage masks (binarized for inpainting) from our dataset detailed in Chapter 4. 25
2.9 Comparison of MVTec AD Dataset images and defects with ARTeFACT images and defects. 28
2.10 Qualitative comparison between our preliminary end-to-end restoration approach [94] and "Bringing Old Photos Back to Life" [204]. 33

3.1 Input and ground truth from our authentic artefact damage dataset, along with restorations from some of the models we evaluated, presented at full resolution. Ours is the first public dataset allowing accurate quantitative evaluation of film damage restoration. 38
3.2 Overview of our proposed approach to generating realistic damage overlays showing analogue film artefacts (Section 3.6). We extract statistics from 12 000 manually-annotated artefacts, and use these to build a probabilistic model of artefact damage that can be applied to clean film scans. The resulting damage overlays are highly realistic, and suitable for training image restoration models. 40
3.3 Number of artefacts per 256×256 pixel patch for each artefact type across all scans, and fitted Gamma distributions. 45
3.4 Examples of extracted and padded artefacts. The number above each is its area in μm^2 (square microns). 46
3.5 An example of an annotated scan. Full resolution (4944 by 3396 pixels) is shown at the top, along with zoomed-in regions around examples of each type of artefact below. Overlaid annotations for dust (green), dirt (blue), long hairs (yellow), short hairs (magenta), and scratches (red). 47
3.6 Artefact size histograms in pixels, and fitted Gamma distributions. While average sizes for each type of artefact vary, the sizes of artefacts are distributed in a similar way. 48
3.7 Visual comparison between the true spatial distribution of short hairs (by kernel density estimation on annotation centroids) **(a)**, and Perlin noise **(b)**, which we use as an approximation when sampling new overlays. Each plot shows the distribution over a full 35mm frame, with brighter colors corresponding to higher density. 49
3.8 Example image pairs (positives and negatives) shown to perceptual study participants at five different crop resolutions (in pixels). 50
3.9 High resolution image, generated damage overlay and resulting synthetically damaged image. 51

3.10 Overview of (a) training and (b) evaluation workflows for restoration, segmentation and inpainting models. We retrain existing restoration models on our synthetic data where relevant, and train an additional segmentation model to adapt state-of-the-art inpainting methods to the film artefact restoration task. 52
3.12 Input and ground truth from our authentic artefact damage dataset, along with chosen restorations. 56
3.12 Input and ground truth from our authentic artefact damage dataset, along with chosen restorations. 57
3.12 Input and ground truth from our authentic artefact damage dataset, along with chosen restorations. 58
3.13 Input and ground truth from our authentic artefact damage dataset, along with chosen restorations. 62
3.13 Input and ground truth from our authentic artefact damage dataset, along with chosen restorations. 63
3.13 Input and ground truth from our authentic artefact damage dataset, along with chosen restorations. 64
3.13 Spatial frequencies aggregated across all 10 scans, broken down by artefact type ((a), (b), (c), (d), (e)) and for all artefact types ((f)). 65

4.1 Examples from our dataset of damaged artwork, categorised by Material (rows 1 and 2) and Content (row 3). Annotation colours correspond to different types of damage. Note the diversity of media and content, and pixel-accurate damage masks. 69
4.2 Damage types found in our dataset, demonstrating the variety of shape, scale and severity of damage. 73
4.3 Overview of the prevalence and severity of different damage types in the Dataset. 74
4.4 Qualitative results for SAM at zero-shot damage segmentation. Top row shows initial segments from N prompts as predicted by SAM (unique colour per prompt), bottom row shows the segments after being assigned binary class (Clean or Damaged) via an oracle. 76
4.5 Qualitative comparison for supervised binary (top row) and multiclass (bottom row) damage segmentation. 80
4.6 Example segmentation, class and text annotations from our dataset. 84
4.7 CLIP embeddings of the images in the dataset, dimensionality reduced via t-SNE. Each colour corresponds to a material (a) or content (b) category as provided in our dataset. 85
4.8 Relative damage type frequency per material (a) and content (b) class. Lighter colours indicate higher frequency. 85

4.9 Qualitative comparison for binary damage segmentation across Material categories. . . . 90
4.9 Qualitative comparison for binary damage segmentation across Material categories. . . . 91
4.10 Qualitative comparison for binary damage segmentation across Content categories. 92
4.11 Qualitative comparison for binary damage segmentation across Material categories. . . . 93
4.11 Qualitative comparison for multiclass damage segmentation across Material categories. . . . 94
4.12 Qualitative comparison for binary damage segmentation across Content categories. 95
4.13 Qualitative comparison for diffusion-based damage segmentation between DiffSeg and DiffEdit. . . . 96
4.13 Qualitative comparison for diffusion-based damage segmentation between DiffSeg and DiffEdit. . . . 97
4.13 Qualitative comparison for diffusion-based damage segmentation between DiffSeg and DiffEdit. . . . 98
4.14 Qualitative comparison for text-guided damage segmentation using DiffEdit over center-cropped images and in full resolution. . . . 99
4.14 Qualitative comparison for text-guided damage segmentation using DiffEdit over center-cropped images and in full resolution. . . . 100
4.14 Qualitative comparison for text-guided damage segmentation using DiffEdit over center-cropped images and in full resolution. . . . 101

5.1 We perform unsupervised segmentation by applying Normalised Cuts [181] to self-attention features from Stable Diffusion [167] in a hyperparameter-free setting, and achieve superior performance on Coco-Stuff-27 [27] compared to trained and zero-shot methods. . . . 104
5.2 Qualitative comparison on COCO-Stuff-27 between our simple Random Walk approach for different NCut cost thresholds (columns 2-4) and DiffSeg for different Kullback-Leibler Divergence thresholds (columns 5-7). . . . 112
5.3 Self-attention PDF (aggregated) (a) and corresponding dot product (b) and cosine similarity (c) adjacency values for a single patch. . . . 113
5.4 Qualitative comparison on COCO-Stuff-27 between NCut over a dot-product Adjacency matrix (columns 2-4) and over a cosine similarity Adjacency matrix (columns 5-7) across a range of NCut thresholds. . . . 114
5.5 Qualitative comparison of segmentation results for our two automatic thresholding approaches: Scaled MinCut (columns 2-5) and NCut (columns 6-9) across different self-attention resolution levels. . . . 115

5.6 Power Walks: Self-attention extracted from Stable Diffusion for an arbitrary patch marked by red dot (row 1) framed as a random walk transition probability matrix and raised to the k^{th} power, allows us to build adjacency matrices which increasingly capture higher-order patch-similarities (row 2). The resulting unsupervised segmentations (row 3) produced by our hyperparameter-free NCut method demonstrate that as long-range relationships are enhanced, the granularity of segmentation decreases, while still extracting semantically-meaningful segments. 116
5.7 Qualitative comparison of segmentation results on Coco-Stuff-27 [27] Validation across multiple methods. All approaches are unsupervised; PICIE and STEGO are fit on the training set of Coco-Stuff-27, while EmerDiff, DiffSeg, DiffCut and Ours are all zero-shot and rely only on features extracted from Stable Diffusion. Our approach, in contrast to DiffSeg, DiffCut and EmerDiff, is also hyperparameter-free. 122
5.8 Qualitative comparison of segmentation results on Cityscapes [49] Validation across multiple methods. All approaches are unsupervised; PICIE and STEGO are fit on the training set of Cityscapes, while EmerDiff, DiffSeg, DiffCut and Ours are all zero-shot and rely only on features extracted from Stable Diffusion. Our approach, in contrast to DiffSeg, DiffCut and EmerDiff, is also hyperparameter-free. 123
5.9 Examples of predictions used in the traditional evaluation setting (Hungarian matching) from the zero-shot diffusion feature-based approaches (columns 2, 4, and 6), and the corresponding merged predictions (columns 3, 5, and 7), designed to account for the models segmenting more granular classes compared to what is given in the ground truth. Notice that predictions from our model can be even more precise (baseball players example, row 3) or more accurate (tennis net covering the player's legs, row 1). 123
5.10 Segmentations of various types of anomalous objects from the MVTec Dataset [17]. 124
5.11 Segmentations of Brain MRI scans from the BraTS 2017 Dataset [136]. 124
5.12 Segmentations of varying granularities on different types of damaged analogue media [93]. 124
5.13 Segmentations of varying granularities of in-the-wild images captured by a smartphone: overlay (left), input (middle), and segmentation (right). 125

6.1 Edge case: the best performing damage segmentation model from Chapter 3, applied to a film scan in the wild. While they may *look* similar, Cooper's whiskers are *not* damage! 129

6.2 SAM requires an infeasible amount of bounding boxes (a), effectively asking the user to do the heavy lifting in damage detection to produce accurate segmentations (b). In contrast, DiffEdit, leveraging the visual-semantic features of StableDiffusion, can detect damage via a single prompt. However, it also detects salient features (faces) and objects which *look* like damage, like the string the cherub is holding. 130
6.3 Text-to-image generative models are able to generate convincing "ancient" frescoes, without explicitly mentioning the presence of damage. The prompt for this image is "Ancient roman fresco of a man presenting a modern laptop, closeup photograph, pompeii –ar 3:2 –v 5.1 –q 2". Reproduced from "A well-preserved Roman fresco recently unearthed in Pompeii". 131
6.4 Self-attention activations extracted from Stable Diffusion for example damaged images, with Principle Component Analysis applied for dimensionality reduction. Showing the first 3 components, matched to RGB channels. 132
6.5 Applying the approach developed in Chapter 5 to examples from the ARTeFACT dataset from Chapter 4. 133
6.6 OVAM applied to various real images of cats, showing raw and optimized attention attribution maps for the <*cat*> token. First column has single cat instances, second column shows multiple cat instances, and third column shows cats in stylised depictions (mosaic and painting). 134
6.7 OVAM applied to examples from the ARTeFACT dataset from Chapter 4 (random sample per material class). Ground truth binarized masks in first column; unoptimized <*damaged*> token OVAM [132] attribution maps in second column; the remaining columns show the token attribution maps optimized for a given material class, applied across all classes in the dataset. Best viewed zoomed-in. 135
6.8 We can extend the approach described in Chapter 5 by using the OVAM attribution maps as bias vectors in an extension to the Normalized Cuts algorithm. . . 136

Acknowledgments

This thesis arose at the intersection of my passions for visual media and machine learning, and people across all aspects of my life contributed to its realization. Some provided direct academic guidance, others—the equally valuable gifts of emotional support, perspective, and encouragement. Many did both! So many, in fact, it is impossible to name all.

I was fortunate to have an entire constellation of supervisors and mentors. Without Paul Siebert's enthusiasm, I would have not pursued this PhD; without Paul Henderson's expertise and diligence, I likely would have not completed it! Special thanks to my primary supervisor, John Williamson, whose influence spans my entire academic journey—from delivering the first Computing Science lecture I attended eight years ago to stepping in as my supervisor four years ago. You have been an unwavering source of knowledge, inspiration, and guidance. Thank you to Gerardo Aragon-Camarasa who, though not officially my supervisor, always made the effort to make me feel included within the CVAS research group. Thank you to Marco Aversa for being a good friend, an excellent collaborator and an exemplar academic. Thank you to Fani Deligianni and Nicolas Pugeault for their invaluable feedback throughout the years. I extend my thanks to the numerous academics and friends in SoCS, and in particular to Melonie and Özgü, who brightened the final stretch of this journey.

Thank you to UK Research and Innovation and the British Taxpayer for providing funding for the first 3.5 years of my PhD under grant number EP/R513222/1, and to Gethin Norman for ensuring sufficient GTA hours to allow me to support myself thereafter. Thank you to Dmitri, Arild, Stefan and Sebastían, for believing in this work and trusting me with their data and/or expertise.

Thank you to my parents Katerina and Stefan, and to my brothers Tsonko and Stoyan, for simply being there for me; to Plamena for always making the effort to see me during my sporadic visits home; to Maggie for being a cornerstone presence in my life despite geographical distance; to notable Glaswegians Yana, Martin and Vasilena for keeping me social and cheering me on. Thank you to my cats Cooper and Bogart, and dogs Hector and Richard, for providing cuddles when they were needed most.

Last but not least, I give my utmost gratitude to Wesley, for his love and patience, and for making me take a break every now and then; and to his family, for treating me like their own.

This thesis would not have been possible without all of you.

Declaration

I hereby declare that this thesis is the result of the my own work, except where explicitly indicated otherwise in the text.

With regards to chapters in journal format (Chapters 3 4 and 5), the sections where the authors' versions of publications have been included, have been prefaced with a disclaimer stating the contribution of all named authors to the work.

I declare that this thesis does not contain work submitted for any other degree at the University of Glasgow or any other institution.

List of papers

The following is a list of publications which the work in this thesis directly contributed to:

- Ivanova, D., Siebert, J. P. and Williamson, J. (2022). *Perceptual Loss based Approach for Analogue Film Restoration*. In: 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, pp. 126-135.
- Ivanova, D., Williamson, J. and Henderson, P. (2023). *Simulating analogue film damage to analyse and improve artefact restoration on high-resolution scans*. Computer Graphics Forum, Vol. 42, No. 2, pp. 133-148.
- Ivanova, D., Aversa, M., Henderson, P., and Williamson, J. (2024). *State-of-the-Art Fails in the Art of Damage Detection*. In: Vision for Art VISART VII Workshop at the European Conference of Computer Vision (ECCV).
- Ivanova, D., Aversa, M., Henderson, P., and Williamson, J. (2025). *ARTeFACT: Benchmarking Segmentation Models on Diverse Analogue Media Damage*. In: Proceedings of the Winter Conference on Applications of Computer Vision (WACV), pp. 7439-7449.
- Ivanova, D., Aversa, M., Henderson, P. and Williamson, J. (2025). *Unsupervised Segmentation by Diffusing, Walking and Cutting*. arXiv preprint arXiv:2412.04678. (under review)

In addition, I have also contributed to the following publications:

- Aversa, M., Nobis, G., Hägele, M., Standvoss, K., Chirica, M., Murray-Smith, R., Alaa, A.M., Ruff, L., Ivanova, D., Samek, W. and Klauschen, F. (2023). *Diffinfinite: Large mask-image synthesis via parallel random patch diffusion in histopathology*. In: Advances in Neural Information Processing Systems (NeurIPS), Vol. 36, pp. 78126-78141.
- Henderson, P., de Almeida, M., Ivanova, D. and Anciukevičius, T. (2024). *Sampling 3D Gaussian scenes in seconds with latent diffusion models*. arXiv preprint arXiv:2406.13099. (under review)
- de Almeida, M., Ivanova, D., Williamson, J. and Henderson, P. (2025). *Pixel-Motion: Image-to-4D Prediction for Camera-Controlled Video Synthesis*. (under review)
- Shi, T., de Almeida, M., Ivanova, D., Pugeault, N. and Henderson, P. (2025). *SplatPortrait: Generalizing Talking Heads with Gaussian Splatting*. (under review)

List of datasets

The following is a list of datasets collected, curated and annotated as part of the work in this thesis:

- 35mm Authentically Damaged and Expertly Restored Film Scans
- Synthetically Damaged Documerica Film Scans
- ARTeFACT: Diverse Analogue Damage Across Various Types of Media

Note on the format of this thesis

This thesis contains three chapters that are in journal format: Chapters 3, 4, and 5. These have been included in the thesis with the same format as the rest of the thesis, and numbering of pages, references etc. are in sequence with the rest of the thesis.

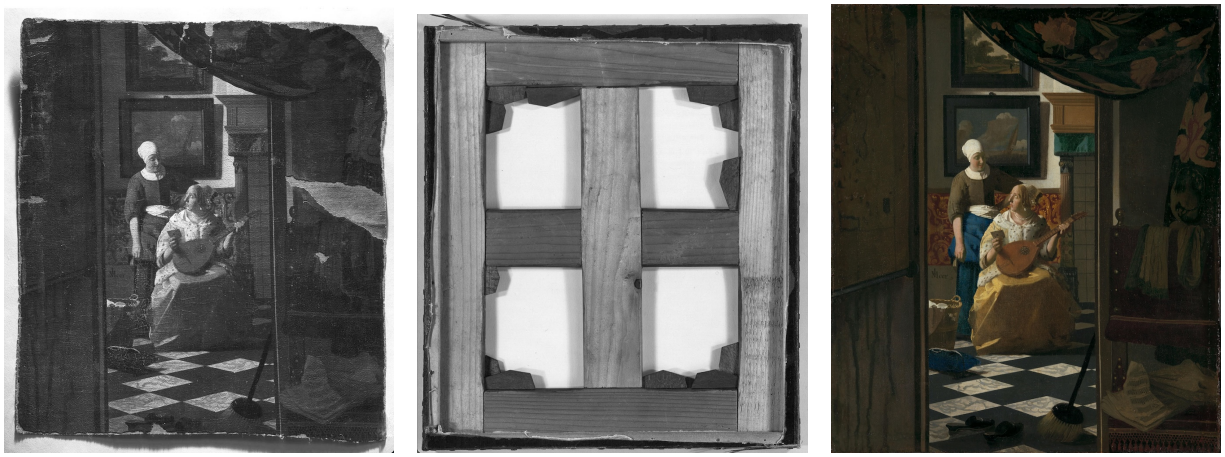
Chapter 1

Introduction

We know that behind every image revealed there is another image more faithful to reality, and in the back of that image there is another, and yet another behind the last one, and so on, up to the true image of that absolute, mysterious reality that no one will ever see.

MICHELANGELO ANTONIONI, 1995

Restoration (noun), "is the action of returning something to a former (...) condition" [165]. Yet the *former condition* in *image restoration*, especially for *analogue media*, is inherently uncertain. No imaging process yields a perfect, lossless record of reality: digital images suffer from sensor and compression artefacts, while analogue ones are shaped by artistic intent, provenance, and the material vulnerability of their medium. Which of these layers to reveal has long been debated in the arts and humanities. This thesis instead explores, formalises, and tackles analogue media restoration through the lens of machine learning instead.



The Love Letter—damaged | The chassis of the painting | The expert restoration

Figure 1.1: Vermeer's *The Love Letter* was crudely cut from its frame and stolen. Upon recovery, an international expert committee at Rijksmuseum spent a year restoring it [110].

This pursuit finds a reflection in the history of Vermeer's "The Love Letter." Vermeer meticulously captured reality through precise perspective, likely using a camera obscura [188] to translate his artistic vision onto canvas. This embodiment of his intention remained largely intact until 1971 [195], when it suffered brutal damage: slashed with a potato peeler during theft, as an attempt to draw attention to the 1971 Bangladeshi genocide [95]. Before restoration proceeded, the damaged masterpiece was documented—somewhat ironically—via grayscale photography.

Machine learning is now capable of recolorising the grayscale photograph [236, 111, 6, 174]. This thesis instead aims to apply machine learning to identify the physical damage which occurred to analogue media—such as the canvas—itself. "The Love Letter"'s provenance exemplifies the fundamental challenge: distinguishing between the artist's intention (depicted image), physical damage to the medium (tears on the canvas), and artifacts of the digitization process (e.g. loss of colour information).

When we examine a digital photograph of "The Love Letter", we see a flattened representation that combines the artist's original depiction on canvas, knife damage from theft, and any artifacts from the photographic process itself, i.e. loss of colour information. Machine learning must somehow separate these layers—distinguishing the slashes from the artist's brushstrokes, identifying what is damage versus the true image—all without the clear mathematical models or extensive training data that typically enable restoration in a computational setting.

1.1 Restoration: Dual Perspectives

John Berger described an image as "a sight which has been recreated or reproduced" [16], highlighting how every imaging process inherently transforms reality [1]. Conservation science and computational image restoration represent two parallel approaches that both strive toward recovering a "true image," but with different understandings of what constitutes an image, damage, and the restoration process itself.

In conservation science, an *image* is inseparable from its physical substrate. A painting is not merely its visual content but also its canvas, pigments, and varnish; a photograph includes both its visual representation and its physical emulsion on paper or film. *Damage* refers to mechanical or chemical alterations to this material substrate—cracks in paint, tears in canvas, scratches on film—that compromise both the artifact's appearance and its physical integrity. *Restoration* involves stabilizing the physical medium and recovering the original appearance through material interventions. Conservators analyze degradation mechanisms, document condition histories, and make careful interventions using specialized tools and techniques. Their work is governed by principles of minimum intervention, reversibility, and material authenticity, treating each artifact as unique with its own historical and cultural context.

In computational imaging, an *image* is a two-dimensional array of intensity values, ab-

[1] We depict what we see, or imagine. [16]

stracted from any physical medium. *Damage* represents deviations from an ideal signal, typically modeled as known transformations, e.g. blurring, noise addition, compression artifacts, loss of colour information. Computational image *restoration*, developed within digital signal processing and inherited by machine learning, treats restoration as a signal recovery problem. It models degradation as mathematical operations on the image signal and seeks to invert these operations algorithmically. With optical systems and digital sensors, degradations follow well-understood physical principles: lens distortions, sensor noise, and sampling effects can be described parametrically. However, it is intractable to analytically model all the ways someone might cut canvas with a potato peeler, let alone the countless other ways physical media deteriorate through mechanical damage, chemical degradation, or biological processes. So how could we bridge this gap?

1.2 Research Challenges

Traditional digital signal processing (DSP) approaches to restoration directly model degradations to invert them. These models require manual adjustment for each image and cannot adequately capture the complex physical processes that damage analogue media. DSP excels at well-defined degradations like blur or noise but struggles with spatially varying, non-uniform damage patterns characteristic of physical media.

Machine learning methods improved upon DSP by learning from large datasets rather than requiring explicit mathematical models. By observing many examples, neural networks can estimate generalizable functions that work across unseen data without per-image parameter tuning. These approaches have revolutionized digital image restoration tasks like super-resolution, colorization, and denoising.

Applying computational approaches to analogue media damage restoration presents fundamental challenges in either setting:

1. **Intractable degradation modeling**: DSP requires explicit mathematical models of degradation, but the complex physical processes that damage analogue media cannot be easily explicitly parameterized in a way which generalises across different media.

2. **Training data scarcity**: Machine learning requires many examples, but paired damaged/undamaged analogue media is rarely available naturally, and purposefully damaging cultural heritage to create such pairs is ethically unacceptable, impractical, and wasteful. Where it does exist, digitised damaged media is not usefully labeled.

3. **Simulation paradox**: To generate synthetic training data, we must model the damage processes—bringing us back to the initial challenge of modeling complex physical phenomena.

Another consequence of computational image restoration focusing primarily on sensor, transmission or display degradations is that they are mainly applied to natural images. Analogue media, however, encompass a much broader range of representations. Even within a single medium like canvas, artistic styles vary dramatically from Titian's realism to Pollock's abstraction. Generalizing across diverse analogue media types introduces an even higher order of stylistic complexity, challenging traditional restoration approaches that assume consistent image characteristics.

Unlike visually cohesive concepts that neural networks excel at recognizing—a cat remains identifiably feline across variations in pose, lighting, or artistic style—damage lacks visual consistency. Its manifestation depends on medium-specific physical properties, varying dramatically between film emulsion, canvas, parchment, or mosaic. This variation extends across multiple dimensions: visual appearance, scale, spatial density, opacity, and severity. It can be intra-media (film exhibiting scratches, dust, and light leaks simultaneously) and inter-media (cracks in paintings, peels in mosaics, tears in tapestries), yet both conceptually represent "damage".

1.3 Restoring Analogue Damage: A Computational Approach

Despite these challenges, recent advances in machine learning (particularly foundation models with emergent semantic understanding, as this thesis will later explore) still offer potential for bridging the gap between conservation science and computational restoration. We first confirm the limitations of supervised learning empirically, before shifting the focus from modeling specific degradation processes to disentangling the representations of content and damage.

The work on this thesis follows three stages:

1. **Statistical modeling of damage**: Developing detailed statistical models of how intra-medium damage manifests, using 35mm film as a case study, to enable realistic synthetic data generation for supervised learning.

2. **Cross-media generalization assessment**: Evaluating whether supervised approaches can generalise across different media types, image contents and damage manifestations.

3. **Semantic disentanglement**: Leveraging the multimodal representations learned by text-to-image diffusion models to perform semantic segmentation in images, with the outlook to apply it to damage segmentation.

This progression mirrors the broader evolution of computational imaging: from explicit mathematical models to data-driven approaches to discovering and utilising emergent behaviours in foundation vision models. However, it applies this evolution to the unique challenges of analogue media, opening new avenues for digital conservation of cultural heritage.

1.4 Motivation

1.4.1 Conservationist

Prior to digital formats, images existed as physical artifacts (e.g. cave paintings, mosaics, frescoes, photographs) inherently susceptible to material decay. Each medium exhibits specific vulnerabilities: canvas tears and flakes, film emulsion scratches and dust, mosaics crack and peel. The volume of damaged analogue media in cultural heritage collections worldwide creates an urgent need for automated restoration methods. Their digitization alone is extremely laborious. The following manual damage identification by conservators, while expert, is also prohibitively time-consuming [95] and expensive [186, 76, 48] for large-scale preservation efforts. Library, museum, and archive collections contain millions of damaged items, from fragile manuscripts to deteriorating film negatives, many at risk of permanent loss without timely intervention [186]. Digitization offers an advantage which aligns with traditional conservation ethics: non-destructive restoration [9]. Converting analogue media to the digital domain allows manipulation and restoration while preserving the original artifact [68]. Automated damage detection systems could dramatically accelerate preservation workflows, enabling triage of at-risk materials and supporting conservators in thorough document assessment at a reduced cost [68]. However, the visual diversity and complex interaction of damage with artistic intent make this a significant technical challenge requiring advanced solutions.

1.4.2 Practical

Despite digital media becoming ubiquitous in the 21st century, people still create using analogue media. An example of this is the recent resurgence of analogue film photography. After filing for bankruptcy in 2012, Kodak marked a 21% increase in profit in 2019, and have continued to expand their film production since [77]. The newly released Pentax 17 represents the first new production 35mm film camera from one of the major brands in nearly 20 years [151]. Even still, improper handling could result to damage in newly-shot film emulsions, which persists in digital scans, and can only be restored manually or via expensive and inaccessible hardware.

Simultaneously, major filmmakers continue to champion analog film despite digital dominance. Christopher Nolan shot “Oppenheimer” (2023) entirely on 65mm film, following his work on “Dunkirk” (2017) and “Tenet” (2020). Quentin Tarantino’s “Once Upon a Time in Hollywood” (2019), Yorgos Lanthimos’ “Poor Things” (2023), and Robert Eggers’ “Nosferatu” (2024) were also captured on celluloid [182]. Denis Villeneuve’s “Dune” is particularly interesting, as it was shot on digital, transferred to 35mm film, and then scanned back to digital [3]. These directors value film’s distinctive color rendition, highlight handling, and organic grain structure—qualities digital cinematography emulates but cannot perfectly replicate, supported by Kodak’s renewed commitment to film manufacturing [77].

1.4.3 Personal

Beyond academic curiosity and broader cultural motivations for automating analogue media restoration, this research was also shaped by my direct experience as a film photographer. Over more than a decade working with 35mm and medium-format film, I have gained practical expertise in shooting, developing, and scanning, as well as a detailed understanding of how different emulsions behave under digitisation and how common artefacts such as dust, dirt, hairs, and scratches present across materials. This background informed the design of datasets, the credibility of manual annotations, and the evaluation of restoration results, ensuring they reflect both the physical characteristics and aesthetic qualities valued by practitioners and conservators alike.

More specifically, the precursor to this project (i.e. my Honours Thesis work, reviewed in Section 2.4) originated from a personal incident in 2019, when several rolls of negatives processed at a commercial lab were returned with significant damage. These rolls contained the last photographs I would take of family and friends before the COVID-19 pandemic. Faced with the prospect of manually restoring each image individually—an incredibly time-consuming and tedious process—I instead saw an opportunity to apply machine learning techniques to automate the restoration process [2]. Hence, what began as a practical solution to a personal problem expanded into a broader investigation of how computational approaches might address the complex challenge of analogue media damage.

1.5 Thesis statement

This thesis investigates whether machine learning techniques can effectively detect and segment damage in analogue media, focusing on the core challenge of distinguishing between intentional content and physical damage across diverse media types.

1.5.1 Research Questions

The central research question driving this work is: "Can machine learning approaches successfully disentangle physical damage from intended content in digitised analogue media?" This overarching question is explored through three specific research questions:

- RQ1: How can we develop statistically rigorous models of analogue media damage to enable robust synthetic data generation for supervised learning?
- RQ2: To what extent can supervised machine learning approaches generalize across different types of analogue media, damage, and contents?

[2] Clearly without realizing how time-consuming and tedious a PhD can be as well!

- RQ3: How can unsupervised methods leveraging multimodal foundation models overcome the limitations of supervised approaches in distinguishing damage from visually similar semantic content?

To address these questions, this thesis presents a systematic progression from supervised approaches trained on synthetic data to unsupervised methods leveraging the rich semantic representations in foundation models, demonstrating that multimodal understanding is essential for conceptualizing damage as a *disruption of intended meaning* rather than merely a visual pattern.

1.5.2 Scope

This thesis is limited to the analysis of digital scans of two-dimensional analogue media—paintings, photographs, manuscripts, mosaics, and similar flat artefacts. Three categories of media are deliberately excluded: (i) three-dimensional objects such as sculpture and pottery, which require volumetric or multi-view representations; (ii) moving images, including motion-picture film, whose temporal structure introduces distinct restoration challenges; and (iii) multi- and hyperspectral modalities (infrared, ultraviolet, X-ray, Terahertz), which, despite their importance in conservation practice, demand specialised acquisition pipelines not feasible for large-scale study.

The work addresses primarily the damage detection component of the restoration pipeline. Inpainting, though a natural follow-up, is omitted for two reasons: first, prior experiments (Chapter 3) confirm that accurate damage localisation is the primary technical bottleneck, while inpainting is comparatively mature; second, inpainting involves subjective aesthetic and historical decisions that vary across cultural contexts, making it less amenable to quantitative evaluation.

All methods are developed for high-resolution RGB imagery, processed with patch-based strategies where required, to accommodate the size and aspect-ratio variation typical of digitised heritage collections. These boundaries define the specific problem space addressed in the thesis: detecting and segmenting physical damage in still, two-dimensional analogue media within the constraints of standard digital imaging.

1.6 Contributions

This thesis makes the following key contributions:

1. A comprehensive, controllable statistical modeling algorithm for film damage, enabling generation of highly realistic synthetic training data validated through perceptual studies — Chapter 3.

2. FILM-R: The first-of-its-kind benchmark dataset of expertly restored high-resolution damaged film scans across multiple emulsion types —Chapter 3.

3. ARTeFACT: The first-of-its-kind comprehensive dataset of analogue media damage, comprising over 11,000 pixel-accurate annotations spanning 15 damage categories across diverse media types and image contents — Chapter 4.

4. A systematic evaluation of state-of-the-art segmentation approaches on diverse analogue media, revealing fundamental limitations in generalization across media types — Chapter 4.

5. A novel unsupervised segmentation framework leveraging self-attention mechanisms in text-to-image diffusion models for zero-shot semantic segmentation — Chapter 5.

1.7 Thesis Walkthrough

Chapter 2 reviews restoration from both cultural and computational perspectives. It traces the evolution of image restoration approaches from traditional digital signal processing to modern machine learning methods, highlighting the unique challenges that analogue media presents.

Chapter 3 introduces a statistically grounded approach for modeling and synthesizing realistic film damage. It presents a comprehensive benchmark dataset of real damaged film scans with expert restorations and demonstrates the performance improvements achieved through training on high-quality synthetic data.

Chapter 4 presents the ARTeFACT dataset spanning multiple media types and damage categories. It systematically evaluates both conventional supervised approaches and state-of-the-art vision foundation models, revealing their limitations in generalizing across media types and distinguishing damage from content.

Chapter 5 proposes a novel framework for zero-shot unsupervised segmentation leveraging the self-attention mechanisms of text-to-image diffusion models, demonstrating superior performance compared to supervised approaches without requiring specific training data.

Chapter 6 synthesizes the findings, discusses limitations of current approaches, and outlines directions for future research in analogue media damage detection and restoration.

Chapter 2

Background

This chapter establishes the theoretical foundations for analogue media restoration by examining both cultural and technical dimensions. We contrast traditional conservation approaches with computational image restoration frameworks (discussed in Section 2.1 and Section 2.2 respectively), identifying the limitations of current mathematical models when applied to physical media damage. State-of-the-art machine learning methods for digital degradations are reviewed to establish context for the novel challenges presented by analogue media 2.3. Finally, we revisit preliminary work that inspired this thesis in Section 2.4, which highlights the need for better damage simulation models and cross-media evaluation methods that form the basis for subsequent chapters.

What constitutes an "image"—and by extension, "restoration"—varies profoundly across disciplines. In conservation, an image is inseparable from its physical medium: a painting is as much the depicted scene and the utilized style as it is pigment and canvas, all with their own history and value. In technical fields, an image is abstracted as a mathematical distribution of color amplitudes, with restoration defined as recovering an ideal signal from degraded observations. These perspectives create fundamentally different approaches to restoration. *The Conservationist* grapples with subjective questions: What damage should be preserved as historically significant? What constitutes authentic restoration? *The Engineer* seeks objective mathematical models that could invert degradation processes with minimal information loss. Digital restoration offers a unique setting for a convergence of these worlds: it provides the opportunity for non-destructive intervention that preserves the physical artifact via a restored digital representation. In this overlap also lies a research question about how machine learning models represent abstract concepts like damage.

2.1 Image Restoration: Cultural Dimensions

Until the late 20th century digital revolution, humanity's visual cultural heritage existed solely in analogue media. Unlike digital media, these physical objects embody a complex duality: they simultaneously contain the artist's intended depiction and the cumulative physical changes to the medium itself over time. The restoration of damaged analogue media is not merely a technical task; it sits at the intersection of multiple disciplines, each offering perspectives that are crucial for framing the problem computationally and for guiding the development of effective machine learning solutions. These perspectives reveal the inherent complexities and nuances that must be considered when designing algorithms for damage detection and restoration.

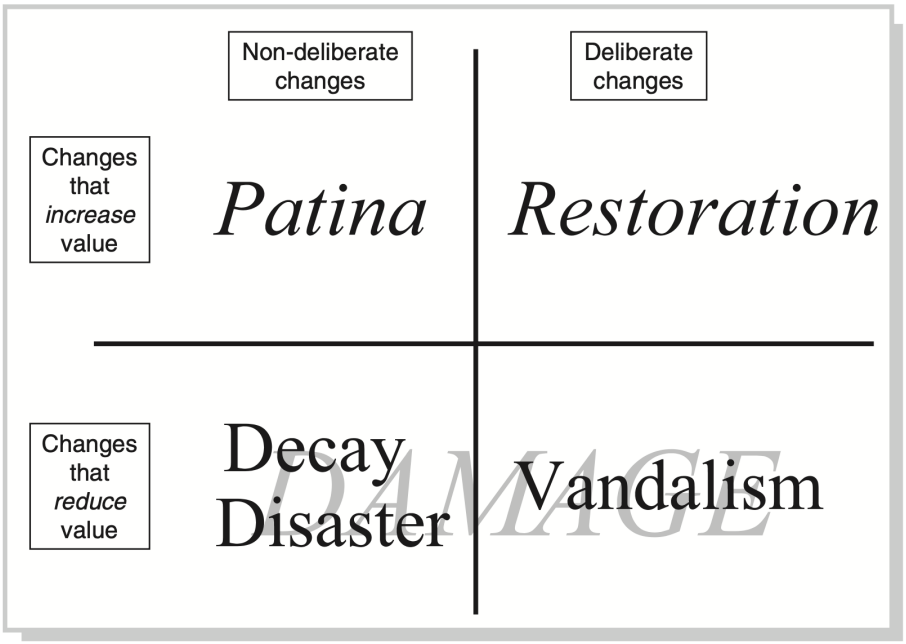


Figure 2.1: Ashley-Smith's framework for classification of a conservation object's possible alterations. Reproduced from [142].

2.1.1 The Medium is the Message

Marshall McLuhan's famous assertion from communication theory [133] works surprisingly well as a metaphor through which we can examine conservation philosophy. McLuhan distinguishes between "content" (which is easily grasped) and the "character" of the medium itself (which is often overlooked). Applied to conservation, this insight suggests that analogue media properties, including accumulated physical changes, can carry their own meanings, complementary or distinct to those of the actual depicted imagery, but equally important. While artists deliberately choose their medium as part of their artistic vision, the inevitable *changes* to that medium over time accrue value not necessarily aligned with the artist's original intention, but rather as a testimony to history and the passage of time.

Salvador Muñoz Viñas observes in his "Contemporary Theory of Conservation" that the very practice of conservation depends on the existence of damage—without actual or potential damage to address, conservation as a field would have no purpose [142]. Yet, is all material change *damage*? Muñoz Viñas extends Ashley-Smith's framework [8], which categorizes alterations to conservation objects based on two key dimensions: whether changes are deliberate or non-deliberate, and whether they increase or reduce an object's value—visualised in Figure 2.1.



Figure 2.2: From left to right: the original version of the fresco; a deteriorated version; the restored version. Reproduced from [147].

However, this neat categorization becomes problematic when intention and outcome diverge. The infamous case of the "Ecce Homo" fresco shown in Figure 2.2 illustrates this perfectly. In 2012, Cecilia Giménez, an 81-year-old amateur with no formal training, attempted to restore a deteriorating fresco of Christ. Her well-intentioned intervention transformed the painting so dramatically that it became known as "Ecce Mono" ("Behold the Monkey"). While intended as restoration (deliberate changes to increase value), the result was widely perceived as vandalism (deliberate changes that reduced value)—at least initially [1].

Indeed, the easiest way to prevent botched restorations is to not restore at all: such is the stance of Martin Heidegger in "The Origin of the Work of Art", where he discusses the "work-Being" of artworks as their natural degradation over time, arguing it should not be interfered with [86]. According to Heidegger, a Greek temple becomes a work of art [2] precisely through its ruin. This sentiment is also directly expressed by John Ruskin decades earlier, referring to the act of restoration as "a Lie from beginning to end" and equating it to destruction [172]. At

[1] Ironically, the "botched restoration" later increased the fresco's fame and tourism value, complicating its categorization further.

[2] Paradoxically, through value theory, only the subset of analogue media which is "art" is worth restoring [142].

the other extreme of conservation philosophy sits Eugène Viollet-le-Duc [3], who postulated that to restore means "... to re-establish [...] in a completed state, which may in fact never have actually existed at any given time" [142]. This approach treats all physical alterations primarily as interference to be removed.

Cesare Brandi attempted to synthesize these opposing perspectives through his 'Theory of Restoration' [24]. Drawing on Gestalt psychology [109], Brandi recognized that lacunae (missing areas) create perceptual disruption, while also acknowledging the historical value of an artifact's journey through time. His solution—using neutral, recognizable infills—aimed to balance visual unity with historical authenticity.

2.1.2 Taxonomies of Damage

The dualistic nature of content and medium degradation is further complicated by how we classify damage itself. Traditional conservation taxonomies have developed primarily to guide manual restoration practices rather than computational detection, categorizing damage based on causative factors and material impacts rather than visual characteristics.

Medium-specific taxonomies reflect the specialized nature of conservation practice. Bucklow [26] developed a classification system for craquelure patterns in paintings, identifying distinct networks associated with different historical schools. For photographic materials, Reilly [162] established a taxonomy of chemical and physical deterioration based on underlying mechanisms, distinguishing between silver mirroring, gelatin degradation, and mechanical damage. Chambah [34] categorized film damage broadly into chemical degradations (fading, color shifts) and mechanical degradations (scratches, dust, tears). Van Camp [30] proposes a "Damage Atlas", categorizing photographic material damage into mechanical (material loss), chemical (discolouration), and biological (microorganisms, rodents).

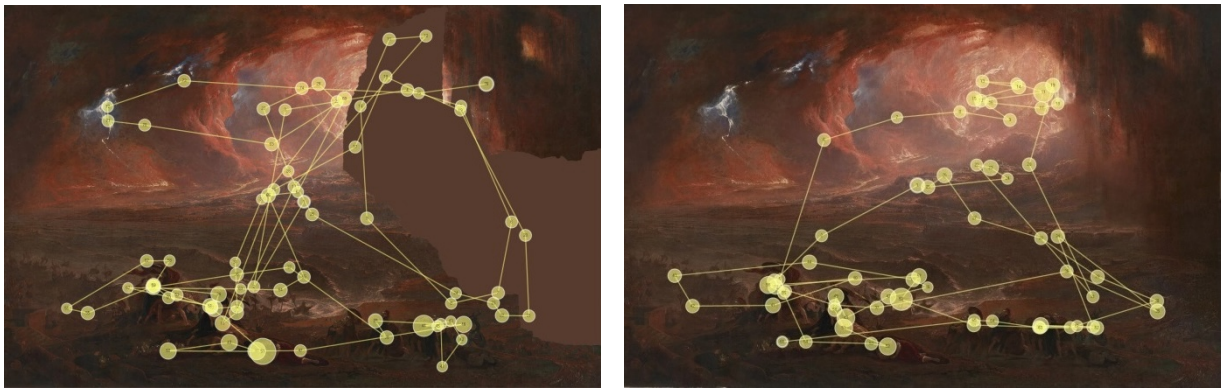
While useful conceptually, these classifications focus on the causes of degradation rather than their visual manifestations—the features that computational systems must actually detect. They tell us why damage occurs, but computational approaches need to know how damage looks. Importantly, damage with the same visual signature can exist across materials.

The ARTeFACT dataset introduced in Chapter 4 directly addresses this gap by developing a taxonomy that bridges conservation terminology with visually distinguishable characteristics detectable through computational means. By focusing on the visual manifestation of damage across different media types rather than their causal mechanisms, it creates a foundation for computational approaches that can better respect the distinctions between medium character, patina, and deterioration.

[3] Famously spent 25 years restoring the Notre-Dame Cathedral.

2.1.3 Perception of Content vs. Medium

While theoretical frameworks and philosophies, despite disagreements, have guided restoration practice for decades, recent empirical research, albeit small-scale by machine learning standards, has begun to test how humans actually perceive and process the relationship between content and medium damage. A study at Tate Britain during the restoration of John Martin's *The Destruction of Pompeii and Herculaneum* used eye-tracking technology to evaluate different restoration approaches [130]. The eye-tracking data confirmed Brandi's observation that lacunae steal away viewer attention. However, it challenged his proposed solution of neutral infills—viewers fixated on the edges of the neutral-colored restoration more frequently than on inpainted restorations with abstracted features, shown in Figure 2.3. In an additional questionnaire, despite being aware of issues of authenticity, the majority of participants rated the inpainted version as the one they would prefer to be displayed at the Tate the most.



(a) Neutral restoration as per Brandi's Philosophy (b) Inpainted restoration (preferred)

Figure 2.3: Gaze scanpaths for the Neutral Infill restoration (left) and full restoration (right) of the painting. Dots represent fixations; lines represent saccades. Reproduced from [130].

Further research by Billi et al. [20] revealed how different viewer groups process damage differently. Novices were initially more drawn to damaged areas (72% looked at lacunae in their first ten fixations) compared to restorers (64%) and art historians (35%). These results suggest that the end user's experience [4] and appreciation of artwork is largely affected by the presence of damage, especially when they are not an expert. Evidently, restoration mus strike a balance between achieving visually-pleasing results and respecting the object's provenance and the medium's other unique properties.

2.1.4 The Digital Advantage

The digital restoration approach resolves many traditional philosophical tensions. It allows the physical medium to retain its historical evidence while creating separate digital representations that can visualize original artistic intent. This can be with improved quality, and reduce

[4] Terminology intentionally borrowed from technical fields.

costs [68]. Consider again Heidegger's Greek temples and statues, long appreciated in their weathered, marble-white state. Modern research has revealed these structures were originally polychromatic [2], challenging centuries of aesthetic assumptions. Researchers are eager to use modern digital tools to restore them [183]—potentially satisfying both Ruskin's concern for material authenticity and Viollet-le-Duc's interest in revealing original appearance. However, the flexibility of digital should not be seen as a carte blanche to commit another "Ecce Mono". For instance, Wan et al.'s "Bringing Old Photos Back to Life" [204] approach (examined in Section 2.10 and further in Chapter 3) not only removes mechanical damage like scratches and dust but also eliminates film grain and modifies the distinctive color characteristics of analogue emulsions, effectively transforming analogue imagery to resemble modern digital photography. This approach implicitly categorizes medium-specific qualities of film as deterioration, despite many photographers and archivists valuing these very qualities as essential characteristics of the medium.

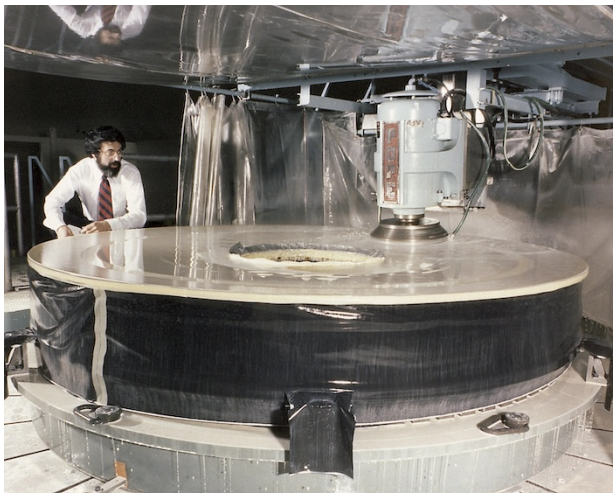
The digital advantage depends critically on accurate damage detection—the ability to computationally distinguish between the intended content, useful medium properties and medium damage. Inpainting, on the other hand, involves reconstructing the intended content, which inherently involves interpretive decisions that vary based on cultural context or expertise. By focusing on damage detection—the identification of physical alterations that might be considered deterioration—we address the most technically challenging aspect, which conservationists can then take Ruskin's, Viollet-le-Duc's, or Brandi's direction—or even in all three!

2.2 Image Restoration: A Technical Perspective

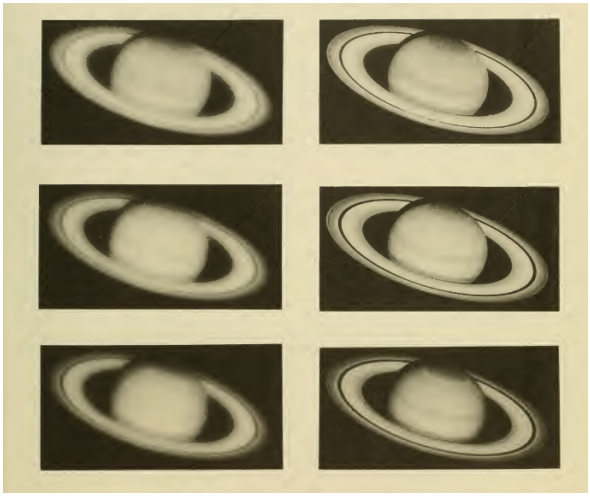
Section 2.1 investigated the multifaceted restoration of analogue media through the lens of conservation, art history, philosophy, and visual cognition. This section shifts the perspective to *image restoration* as defined and implemented within technical fields: Digital Signal Processing (DSP) and Machine Learning (ML). Here, image restoration is mathematically formulated as recovering an original, undegraded image from degraded observation. While DSP and ML have successfully addressed many digital degradations, these computational approaches differ fundamentally from the complex material degradation of analogue media. By examining the technical framework of image restoration along with tasks already solved in it, this section establishes the conceptual foundation for the thesis's core investigation: developing machine learning approaches that aim to solve the unique challenges of damaged analogue media.

2.2.1 Historical context

Digital image restoration emerged directly from pressing needs in early digital imaging, with astronomical imaging as a primary catalyst [5, 12, 143]. The Space Race era urgently demanded maximizing information extraction from degraded space-based imagery [169]. Early planetary



(a) Hubble Space Telescope's primary mirror being ground at the Perkin-Elmer Corporation's large optics fabrication facility in 1979, more than a decade before its very small but very significant flaw was discovered. Reproduced from [145].



(b) Restorations (right) via iterative methods of bias subtracted and flat-fielded HST images of Saturn. Reproduced from [122].

Figure 2.4: Knowing exactly how flat the primary mirror of the Hubble Space Telescope (HST) was allowed scientists to model the resulting spherical aberration and invert it.

missions produced valuable but compromised data: extraterrestrial imaging faced motion blur from spacecraft movement and vibration [143], and noise contamination from both faint light sources (Poisson noise) and electronic/transmission limitations (Gaussian noise).

A landmark application came with the Hubble Space Telescope in 1990, which initially produced blurred images due to a manufacturing flaw in its primary mirror [145], which was too flat. Scientists precisely modeled the spherical aberration as a well-defined point spread function [122, 214], enabling digital correction of images before physical repairs could be undertaken.

Beyond astronomy, diverse other early applications which also drove image restoration's development include biomedical imaging, industrial radiography and photoreconnaissance [5]. With the immense scientific and economic value of each image—sometimes costing millions to acquire [143]—there was a strong imperative to recover information from this degraded data. This directly fueled the development of mathematically rigorous digital image restoration framework, grounded in signal processing theory.

2.2.2 Mathematical formulation

Image restoration is formally defined as the reconstruction of an image through inverting degradation phenomena present in a degraded observation [5, 12, 73]. The classical image restoration problem is commonly formulated as:

$$y = H(f) + n, \tag{2.1}$$

where f represents the original, undegraded image (a pixel grid, $f(x,y)$ where x and y are pixel coordinates), and y is the degraded observation. The degradation operator H encompasses various deterministic distortions—point, spatial, temporal, or chromatic—or their combinations; n represents additive noise, typically modeled as a two-dimensional random process such as Gaussian or Poisson noise. Crucially, this formulation assumes that degradations occur either during the capture process or in the digital domain after capture. For instance, in early astronomical imaging, blurring from atmospheric turbulence or a flat lens (modeled via H) affects light before it reaches the sensor, while Poisson noise from faint sources (modeled via n) represents statistical uncertainty in the sensor readings themselves. In both cases, the underlying subject matter remains physically intact—only its digital representation is degraded.

The goal of image restoration thus becomes estimating f from y by inverting H and mitigating n—an objective process, defined mathematically. This is in contrast with image enhancement, which is considered a subjective attempt to improve an image's appearance for human viewing or subsequent machine processing [5], and which is beyond the scope of our investigation. This distinction underscores the differing objectives and methodologies between restoration and enhancement within DSP-based image manipulation.

This conventional formulation works reasonably well for degradations in digital photography and digital signal transmission, where the degradation affects either the light path before capture (blur, atmospheric effects) or the digital signal itself (sensor noise, compression, transmission errors). Top row of Figure 2.5 visualises some standard examples found in the literature.

As we will see in Section 2.3, this model becomes inadequate when addressing analogue media, where physical damage to the medium itself introduces a fundamentally different type of degradation that occurs between the original subject matter and its digitization and affects the medium itself. Before we discuss those in more detail, however, let's examine where traditional signal processing succeeds and fails.

2.2.3 DSP: Techniques and limitations

Image restoration, based on the $y = H(f) + n$ model, requires precise mathematical descriptions of the degradation (H), noise (n), and image priors (information about f). Historically, early digital signal processing (DSP) achieved success in specific scenarios, particularly those encountered during the Space Race. Motion blur (uniform averaging), atmospheric turbulence (Gaussian PSFs), and defocus blur (uniform circular intensity) were readily modeled due to their global uniformity and well-understood physics, primarily affecting light before capture [5, 12, 73]. Digitization/transmission noise was also statistically tractable [5, 12, 73].



original optical blur motion blur quantization additive noise

Figure 2.5: Common digital degradations shown in the top row can be modeled within the $y = H(f) + n$ framework. Example restorations via standard methods shown in bottom row. These degradations affect either the light path before capture (optical blur, motion blur) or the digital signal itself (spatial quantization, additive noise), and can be modeled with well-understood operators. They do not affect the medium itself pre-digitisation. Image from NASA Flickr Archive.

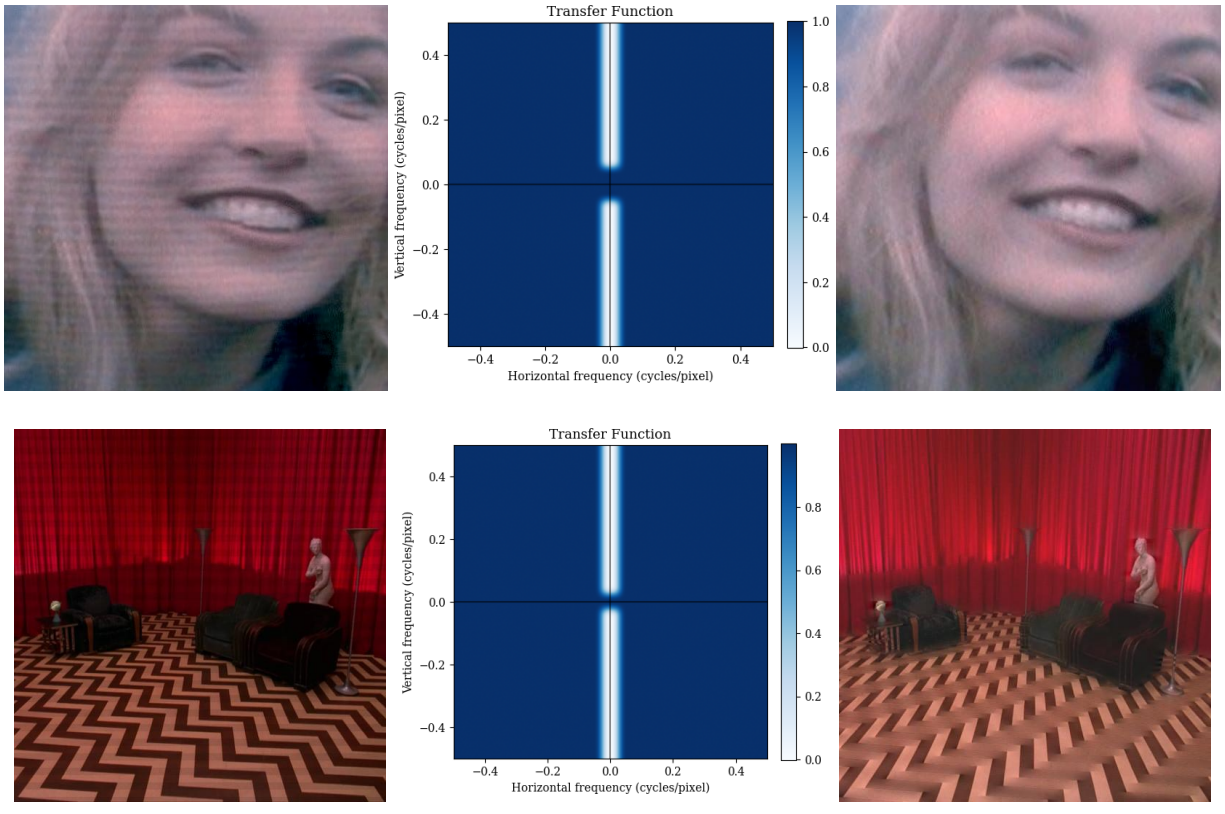
The demonstration in Figure 2.5 (bottom row) illustrates the application of various DSP restoration techniques to common digital degradations (more successful for some than others). For optical and motion blur, we employ Wiener deconvolution [73], which directly inverts the degradation operator while accounting for noise—a classic example of analytical modeling in signal processing. For spatial quantization and additive noise, we utilize more advanced techniques (bilateral filtering [199] and non-local means denoising [25], respectively) that incorporate more sophisticated image priors while still relying on handcrafted mathematical models rather than learning from data. These methods represent the evolution of DSP approaches from direct inverse filtering to algorithms that encode certain statistical properties of natural images. For a comprehensive treatment of these techniques, we direct the curious reader to standard references in image processing [12, 73].

A slightly more complex degradation example is periodic noise, often manifesting as repeating patterns in the image. It is characterized by distinct peaks in the frequency spectrum, and frequency domain filters can be specifically designed to target and attenuate these noise peaks, effectively removing periodic noise artifacts. In the top row of Figure 2.6 we show an example of CRT display line artifact removal, where frequency domain filters effectively target periodic noise peaks.

Despite their mathematical elegance and initial success, DSP techniques have limitations (as seen in Figure 2.5) especially for broader real-world degradations and analogue media. As discussed, DSP excels at global, uniform degradations, like blur. However, it struggles with

analytically intractable, complex, spatially-varying, or non-linear degradations. DSP's effectiveness, especially in more complex cases, hinges on accurate explicit models of degradation (H), noise (n), and even image priors (f). They need to be optimised for each datapoint. Even still, analytically defined DSP priors (smoothness, sparsity) are often too simplistic for complex natural images or intricate degradation artefacts. When information is fundamentally lost (as in de-quantization, see Figure 2.5, column 4), such simple priors provide limited guidance for inferring plausible details compared to ML's rich statistical models.

Bottom row of Figure 2.6 exemplifies the above issues: the analytically defined filter which works well for the example in the top row is not flexible enough when applied to another degraded image with a similar periodic noise pattern. While it removes the CRT-like periodic horizontal line pattern, it introduces new artefacts by degrading image features which have similar frequency signature, as it cannot meaningfully differentiate those from the original degradation due to a lack of prior over the image content.



Original Image Transfer Function Filtered Result

Figure 2.6: Example of frequency domain filtering in digital image restoration. The transfer function in the frequency domain (middle column) is applied to remove periodic noise patterns from the original images (left column), resulting in the restored images (right column).

The analytical modeling shortcomings, especially for complex degradations with non-linearity and spatial variation have been more recently addressed by data-driven paradigms like Machine Learning, discussed next.

2.2.4 Machine Learning Advancements

While machine learning borrows the $y = H(f) + n$ framework from traditional DSP, it fundamentally reframes how prior knowledge about H, n and even f is captured. Moving away from explicit modeling of degradation operators and prior distributions as separate components, neural networks with a much larger number of parameters can implicitly learn these relationships [88] directly from data via backpropagation [171]. As computing power finally matched the requirements of machine learning theory, a fundamental transformation occurred that has driven progress in various image restoration challenges, providing multiple benefits compared to conventional digital signal processing approaches:

- **Handles Complex Degradations (vs. DSP's Limited Scope)**: ML learns complex, nonlinear [54, 21, 74] degradations from data, overcoming DSP's struggle with analytically intractable, real-world degradations [234, 98, 117].
- **Learns Rich Priors for Context (vs. DSP's Simplified Priors)**: ML learns rich, high-order priors for degradations as well as for natural images, enabling context-aware restoration unlike DSP's basic, handcrafted priors [202, 235, 120, 103].
- **Generalizes to Unseen Examples (vs. DSP's Model Sensitivity)**: ML generalizes to unseen examples [83, 82, 21, 74], contrasting with DSP's sensitivity to specific degradation models and need for re-tuning for each data point [38, 211, 229, 119].
- **Effective Prior Application (vs. DSP's Limited Inference)**: ML applies priors learned from data to make plausible inferences in areas where information is fundamentally missing, while DSP can only apply simple priors with limited effectiveness [118, 173, 128, 167].

Through its rich learned priors, ML can synthesize plausible details consistent with available evidence, rather than merely redistributing existing information. This has enabled remarkable progress in previously intractable real-world tasks. In their comprehensive review, Zhai et al. [231] identify six main categories of image restoration as approached by modern machine learning: superresolution, denoising, deraining, degazing, deblurring, and colour correction. Of course, this categorization is by no means exhaustive; we will instead review the following four tasks which have in common the need to learn rich priors about natural images and to model complex degradation processes: superresolution, colourisation, JPEG artifact removal, and inpainting. Figure 2.7 exemplifies those degradations along with example restorations performed via state-of-the-art machine learning methods. These approaches are particularly relevant to analog media restoration, where understanding both the underlying image structure and the diverse degradation mechanisms is essential.

Super-resolution: Early ML approaches to super-resolution employed simple CNN architectures like SRCNN [60], which directly learned mappings between low and high-resolution images. Significant advances came with residual learning in VDSR [105] and EDSR [121], which enabled deeper networks and better performance. GAN-based methods like SRGAN [112] and ESRGAN [206] marked a shift by focusing on training for perceptual quality rather than pixel accuracy. Recent transformer-based models such as SwinIR [120] and diffusion models like SR3 [173] have further pushed boundaries by modeling long-range dependencies and generating highly realistic details through iterative refinement, achieving remarkable results even at high upscaling factors.

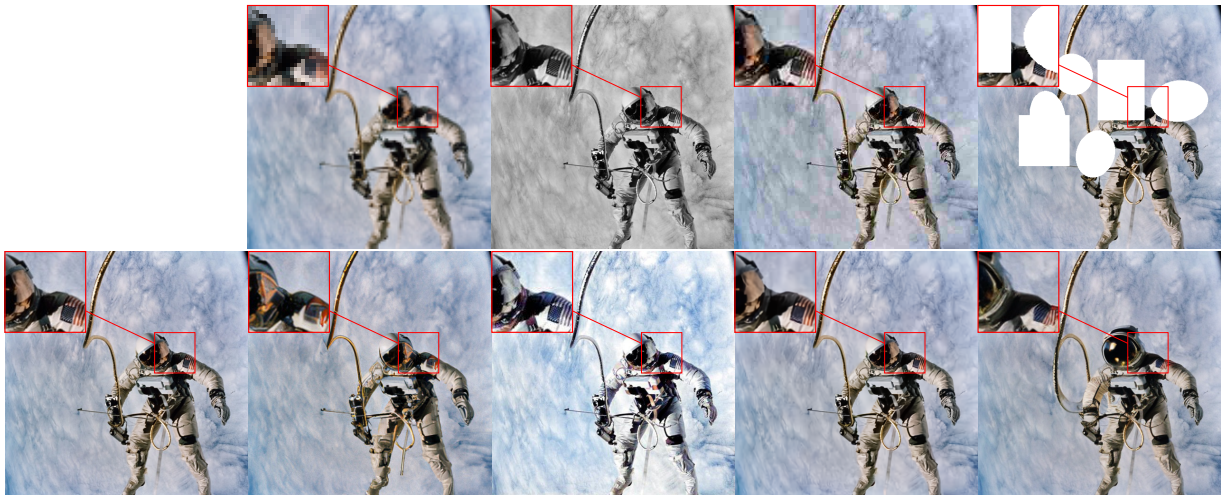
Colourisation: ML-based colourisation evolved from early user-guided approaches to increasingly automated methods. Iizuka et al. [90] pioneered end-to-end CNN colourisation with global and local features, while Zhang et al. [236] framed it as a classification problem with color probability distributions. Adversarial training dramatically improved colourisation among other image-to-image tasks in pix2pix [91] and DeOldify [6]. The challenge of inherent ambiguity [5] in colourisation led to a probabilistic VAE-based approach by Deshpande [57] that could produce multiple plausible outputs. More recently, transformer-based models like ColTran [111] have leveraged self-attention to capture long-range dependencies, while diffusion models like Palette [174] have excelled at producing natural-looking colourisation with greater consistency across complex scenes and historical imagery.

JPEG Artifact Removal: JPEG artifact removal with ML began with simple CNN approaches like ARCNN [226] that directly learned mapping functions from compressed to clean images. This task also advanced with deeper architectures incorporating residual learning, such as DnCNN [234]), which generalized well across different compression qualities. Dual-domain approaches like DMCNN [237] leveraged both pixel and Discrete Cosine Transform (DCT) domain information to better address the frequency-specific nature of JPEG artifacts. GAN-based methods [70] significantly improved restoration quality by learning to generate realistic textures with a patch-wise perceptual loss rather than just minimizing pixel differences. More recently, Chen et al. [36] proposed a transformer framework trained in a contrastive setting with multiple types of degradations, including JPEG compression, while DriftRec [213] adapts the forward stochastic diffusion process to JPEG compression and demonstrates that diffusing with small amounts of noise is enough for a diffusion model to denoise images to a clean state.

Inpainting: ML approaches to inpainting evolved from VAE-based context encoders [154], which first demonstrated CNN-based semantic hole-filling, to more sophisticated GAN architectures like DeepFill [224] that introduced coarse-to-fine prediction with attention mechanisms. Edge-guided approaches [146] improved structural coherence by first reconstructing edges before filling content. A significant breakthrough came with partial convolutions [123] and gated convolutions [223], which addressed the "blind spot" problem of regular convolutions when

[5] There are many plausible colours for an apple.

handling irregular masks. Transformer-based models like LaMa [192] leveraged fast Fourier convolutions to capture long-range dependencies for large hole inpainting. The state-of-the-art has been dramatically advanced by diffusion models such as RePaint [128] and most recently, Stable Diffusion Inpainting [167], which generates remarkably coherent and semantically meaningful content even for large missing regions via a multimodal learned prior.



original superresolution colourisation JPEG compression inpainting

Figure 2.7: Degradations solved by machine learning in the top row can also be modeled within the $y = H(f) + n$ framework. Example restorations shown in bottom row. These degradations are much more complex, but they still occur in the digital domain. Image from NASA Flickr Archive.

Overall, the advancements in these tasks follow a similar path: initially solved in an image-to-image settings, with encoder-decoder architectures, such as Variational Autoencoder (VAE) [107] and U-Net [167]; trained with sophisticated objectives, such as perceptual losses borrowed from style transfer [99], and with the advent of Generative Adversarial Networks (GANs) [75], within an adversarial training framework. More recently, transformer and diffusion models have taken over, with special interest in leveraging pre-trained generative text-to-image models such as Stable Diffusion for tasks such as inpainting. Inpainting certainly stands out among other image restoration tasks as the one most directly related and applicable to analogue media restoration, and has been recently applied to small-scale instances of the problem [219, 141]. However, most ML inpainting methods still require explicitly masked damaged areas—with "blind" inpainting approaches usually training a separate module for degradation detection [209]—assuming prior knowledge of which pixels need restoration.

2.3 Analogue Media Restoration in the Digital Domain

The success of image restoration approaches suggests potential applicability to restoring analogue damage and aligns well with conservation theory requirements of non-destructive restoration, as it is performed in the digital domain and does not affect the original artefact.

2.3.1 Understanding Analogue Damage

While digital degradations occur during image acquisition, processing, or transmission, analogue damage exists in the physical medium itself before digitization. This creates a fundamentally different restoration challenge that requires understanding several unique characteristics:

- **Physical origin:** Damage results from material interactions (mechanical abrasion, chemical reactions, environmental exposure [34, 33]) that affect the physical medium itself.
- **Heterogeneous appearance:** Different damage types (scratches, dust, tears, folds, cracks) have distinct visual signatures; multiple types of damage can exist in the same piece of analogue media.
- **Medium-specific manifestation:** How damage presents depends on the physical properties of the medium (film emulsion, canvas, parchment) and can vary significantly across media types.
- **Localized effects:** Damage typically affects specific regions rather than the entire image, requiring detection of arbitrary shapes and patterns.
- **Complex spatial patterns:** Damage exhibits non-uniform spatial distribution and irregular boundaries that defy simple mathematical modeling.

These characteristics suggest that the degradation model $y = H(f) + n$, while useful for certain digital degradations, is generally insufficient to capture the complexities of analogue damage. A more appropriate mathematical formulation for a damaged analogue image might be:

$$y = H((1 - m_d) \odot f + m_d \odot d) + n, \tag{2.2}$$

where H is the standard, non-analogue degradation operator (as before), m_d is an unknown mask indicating damaged regions, d represents the analogue damage appearance (which varies by type and medium), n represents additive noise, e.g. introduced during digitisation and $\odot$ signifies Hadamard multiplication.

This formulation highlights points of divergence with the classic signal processing formulation:

1. Both the damage mask m_d and the clean image x are unknown.

2. The damage appearance d varies significantly in terms of appearance, scale, and spatial distribution across different media and damage types.

3. The digitisation of analogue media which carries over the intended image depiction as well as any incurred damage adds an additional layer of complexity.

In this formulation, m_d represents a binary, opaque mask where each element indicates the presence (1) or absence (0) of damage at a specific spatial location (e.g. pixel or patch). This representation simplifies the model by treating damage as a discrete state rather than a continuous phenomenon. A more comprehensive model could incorporate the severity or opacity of damage through a continuous-valued mask, allowing for partial degradation. For example, discoloration might partially preserve underlying content while modifying its appearance, whereas material loss completely obliterates the original information. Such a continuous representation would model the spectrum of degradation phenomena observed in analogue media even more accurately. However, it would also significantly increase the complexity of both detection and restoration processes, as well as introduce overhead in the data collection and annotation processes. Additionally, we assume that the damage is localized rather than comprehensive, meaning m_d contains a non-trivial number of zeros, as a completely damaged image (m_d consisting entirely of ones) would render restoration impossible due to the absence of any content prior.

2.3.2 Existing Approaches to Analogue Media Restoration

To better situate analogue media restoration with the image restoration, it's helpful to review the few existing attempts to directly address it, as well as examine its relationship with better established image restoration tasks.

Hardware-Based Detection: Kodak's Digital ICE [65] technology commercially succeeded by using infrared scanning to detect dust, hairs, and other debris on certain types of film emulsion. This approach works because many common film bases are transparent to infrared light while dust and scratches are not [178], allowing for easier subsequent filtering in the Fourier domain. However, this technique is limited to specific emulsion types (primarily color negatives and slides) and requires specialized hardware.

DSP-Based Approaches: Early digital restoration attempts analytically modelled damage features such as gray levels and crack uniformity or scale in paintings [14, 72], or extreme variation in intensity levels in film [100]. Particularly interesting is the large-scale application of such approaches to restore Disney's 1937 "Snow White and the Seven Dwarfs" [67]. Cinesite and Kodak collaborated on the 1993 restoration, which used DSP to detect dirt by color density comparison and temporal frame analysis. While achieving impressive automated dust detection for its time, these approaches required extensive manual parameter tuning and artist correction

to avoid removing meaningful content details. With motion picture film, restorers also have the advantage of neighbouring frames in the temporal domain [23]; yet Disney animators had to reportedly spend hours fine-tuning parameters on individual frames to prevent the algorithm from removing "eyeballs from some of the birds" [170]—a clear example of the need to distinguish between degradation artifacts and visually similar image content, a thread which is followed throughout the rest of this thesis.

ML Approaches: Machine learning methods applied to analogue media restoration have largely emerged as narrowly-focused solutions targeting specific media types, rather than as general-purpose restoration frameworks:

- **Film Restoration**: Several approaches target film emulsion specifically. Strubel et al. [191] train a SegNet [10] encoder-decoder architecture with a cross-entropy loss to restore black and white film scans. Wan et al. [204] proposed a dual VAE [107] architecture to transform between analogue and digital photo domains, which erases not only the analogue damage but all analogue characteristics altogether. Mironica [140] employed a GAN-based [75] approach with residual learning to remove dust and scratches. These methods primarily focus on common film damage types (dust, scratches, minor tears) but struggle with more severe or unusual degradations. While Strubel et al. [191] have released their—albeit limited—training and benchmarking data, Wan et al. [204] and Mironica [140] have not.
- **Manuscript Restoration:** Calatroni et al. [29] addressed damage in illuminated manuscript miniatures via the Chan-Vese segmentation model [35], which partitions images by minimizing an energy function based on region intensity homogeneity, to separate damaged regions from content. This requires expert initialization, where a user must identify representative pixels within damaged areas to establish initial conditions for the segmentation model. The method relies primarily on intensity-based features extracted from RGB color channels and incorporates domain-specific constraints tailored to medieval manuscript pigments.
- **Painting Restoration:** Gupta et al. [78] use a two-stage approach to detect painting damage via Mask R-CNN [84] binary segmentation and U-Net [167] with partial convolutions to inpaint, evaluated on synthetically damaged (via text-overlay) paintings. Cipolina-Kun et al. [44] compares LaMa [192], CoModGans [239], and GLIDE [149] at the task of "completing" MC Escher's paintings via inpainting with pre-defined masks.
- **Fresco Restoration**: Merizzi et al. [137] successfully employ Deep Image Prior [202] with pre-defined damage masks for inpainting damage in frescoes. Xu et al. [219] train SAM-Adapter [39] for binary damage segmentation and use LaMa [192] and Stable Diffusion Inpainting [167] to fill in the identified regions. In both cases, the data is limited in diversity not only with respect to the media type, but also location, age and subject matter.
- **Mosaic Restoration:** Moral-Andres et al. [141] similarly use pre-defined masks to demonstrate that DALL-E [161] can indeed reliably inpaint the missing areas in Roman Mosaics.



Figure 2.8: Example inpaintings with LaMa [192] and Stable Diffusion Inpainting [167] across several different types of media with damage masks (binarized for inpainting) from our dataset detailed in Chapter 4.

2.3.3 Limitations

These medium-specific approaches share several limitations that hinder their broader applicability:

Evaluation Scope: All of the above approaches are evaluated exclusively on the specific medium they were designed for, with no cross-media evaluation. This leaves open questions about their generalization capabilities across different analogue media types.

Synthetic Data Reliance: Several approaches rely heavily on synthetic training data. Mironica's GAN approach [140] and Gupta et al.'s painting restoration [78] method both train on artificially damaged images with overlaid patterns. Wan et al. [204] use real photographs but employ image processing to simulate "old photo" appearances rather than authentic damage patterns.

Data Diversity Constraints: Even within single-medium evaluations, diversity is often limited. Xu et al.'s fresco dataset [219] comes predominantly from a single historical site; Merizzi et al. [137] focus on frescoes from a specific historical period and region; Moral-Andres et al. [141] test exclusively on Roman mosaics of a particular style. Taking this to the extreme, Cipolina-Kun et al. [44] evaluates on works from a single artist! This narrow diversity raises concerns about robustness when applied to the full spectrum of damaged media encountered in real-world collections.

Single-class Damage Representation: All discussed methods, regardless of whether they perform image-to-image restoration, or just inpainting, frame damage detection as a binary problem, where all types of damage are collapsed into a single *damage* class. While sufficient for the inpainting task, this simplification obscures the heterogeneous nature of analogue damage. To achieve more accurate modeling of damage, we require multiclass damage representation, which reflects the varied visual properties of each type of damage.

End-to-End vs. Two-Stage Processing: Strubel et al. [191], Wan et al. [204], Mironica [140], as well as our preliminary work [94] discussed in 2.4, implement end-to-end restoration pipelines that combine detection and inpainting. In contrast, Gupta et al. [78], Merizzi et al. [137], Xu et al. [219], Cipolina-Kun et al. [44] and Moral-Andres et al. [141] employ two-stage approaches, with separate detection and inpainting components.

2.3.4 Decomposing Analogue Restoration: Detection and Inpainting

The above review and analysis of relevant restoration tasks suggests that modern inpainting methods already achieve impressive performance when provided with accurate masks. The approaches which have empirically decoupled damage detection and inpainting and focused on the latter [78, 137, 219, 141] yield compelling results, with common state-of-the-art inpainting methods visually summarised and compared across four different types of analogue media in Figure 2.8. This leaves accurate damage detection as the critical bottleneck in achieving high-quality restoration across diverse media types.

Concurrently, the mathematical framing of analogue damage discussed at the start of this Section naturally lends itself to the same decomposition into the two distinct subtasks:

1. **Damage Detection**: Identifying and precisely delineating damaged regions, i.e. predicting m_d.

2. **Damage Inpainting**: Reconstructing the original content in the identified damaged regions, i.e. inverting $m_d \odot d$.

This formalised decomposition further clarifies the relationship between analogue restoration and other digital image processing tasks and provides a clear conceptual framework for addressing the broader restoration challenge. By treating detection and inpainting as separate problems, we can leverage the considerable progress in image inpainting while focusing research efforts on the more challenging detection task, explored in detail in Chapter 4.

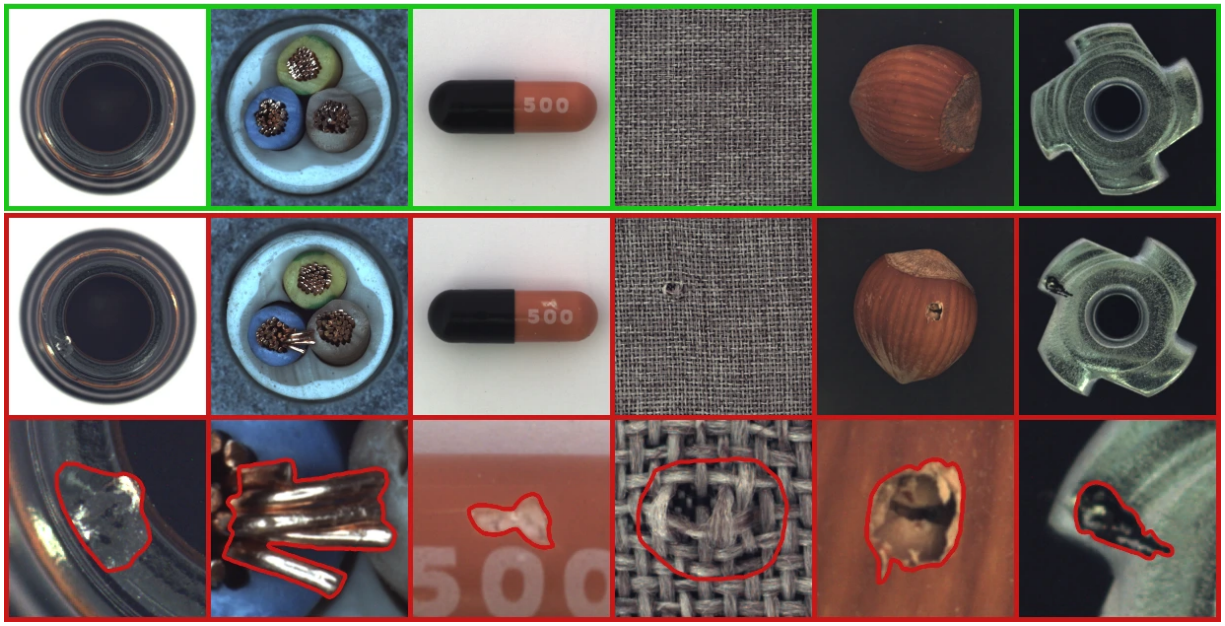
2.3.5 Related Detection Tasks

The following section discusses how several established detection tasks offer methodological parallels to analogue media damage detection, and provides insight as to why none can fully align with the diversity and ambiguity encountered with the analogue damage restoration task.

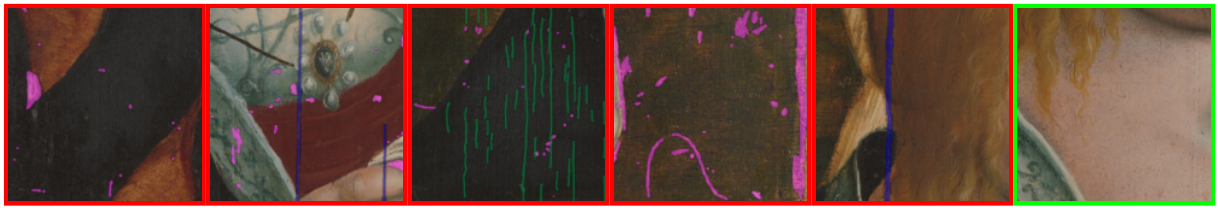
Industrial Anomaly Detection targets defects such as scratches or cracks on largely uniform manufactured surfaces. These tasks exploit the regularity of the underlying material by learning compact models of "normal" appearance and flagging deviations as anomalies, often using autoencoder [220], adversarial [185], transformer [113, 222], or diffusion-based [216, 232] reconstructions. Most recently, new trend has emerged to utilise CLIP's multi-modal feature space [47] as prior over normal samples in zero-shot anomaly detection [96, 241]. Standard datasets include MVTec AD [17] and VisA [242].

Medical Anomaly Detection similarly focuses on irregular structures within complex biological contexts, using priors built on available healthy data to detect pathologies [180, 179]. Medical anomaly detection has pioneered self-supervised and unsupervised approaches, particularly in handling class imbalance (where most pixels are non-anomalous) and in identifying rare or previously unseen damage types [40]. Most industrial anomaly detection methods are also applicable in the medical domain, including CLIP-based multimodal approaches [96, 241]. Standard benchmark datasets include CVC-ColonDB [18] and CVC-ClinicDB [19].

Semantic Segmentation partitions images into semantically meaningful regions based on content. While conventional segmentation focuses on identifying known object categories, unsupervised approaches have demonstrated the ability to discover meaningful visual entities without explicit supervision. Ji et al. [97] and Cho et al. [42] proposed clustering methods to segment without pixel-level annotations, followed by distilling feature correspondences with STEGO [81]. Foundation models have further revolutionized the segmentation landscape. Self-supervised approaches like DINO by Caron et al. [31] learn rich visual representations without explicit segmentation supervision, revealing emergent object-centric properties that enable zero-shot segmentation capabilities. Text-to-image diffusion models such as Stable Diffusion [167] have demonstrated surprising semantic understanding that can be leveraged for segmentation tasks through attention map extraction or fine-tuning approaches as shown by Baranchuk et al. [13], or even in an unsupervised setting by Tian et al. [197], where the Stable Diffusion prior allows for an open vocabulary setting. Perhaps most significantly, the Segment Anything Model (SAM) by Kirillov et al. [108] proposes a promptable segmentation framework trained on over



(a) MVTec AD Dataset. Defect-free images for example categories in top row, corresponding defects in middle row and enlarged in bottom row. Reproduced from [17].



(b) ARTeFACT Dataset. Randomly-located 512×512 crops from a *single* high-resolution (2722×3584) image in our dataset, showing how varied the distribution of damage types, frequency, severity (including no damage at all), and spatial distribution is across patches even within one image. Dataset introduced in Chapter 4.

Figure 2.9: Comparison of MVTec AD Dataset images and defects with ARTeFACT images and defects.

1 billion masks, enabling remarkably robust instance segmentation based on user input. However, while SAM is able to successfully segment visual entities of varied scales and domains, it has been shown to not encode any useful semantic information [62]. Standard benchmark datasets include Coco-Stuff [27], ADE20K [240], or more specialised ones such as CityScapes for driving scenes [49].

Each of these related tasks offers valuable methodological perspectives, but none provides a complete solution to the challenges of analogue damage detection. For instance, while conceptually similar to damage detection, industrial inspection typically benefits from controlled imaging conditions, consistent defect patterns, singular defects per object and well-defined and easily available "normal" reference states, all demonstrated in Figure 2.9a. Similarly, medical images consist of a limited set of visually and semantically consistent categories: both normal (bone, blood cells) and anomalous (lesions, tumours). In contrast, analogue media damage

presents greater complexity due to its organic, heterogeneous appearance (even within the same image, Figure 2.9b), and medium-specific nature, occurring across diverse imaging conditions and historical contexts. In this thesis, the problem is approached through the prism of semantic segmentation, where we aim to segment all pixels which constitute damage, rather than meaningful content.

2.3.6 Metrics

The related tasks of anomaly detection and semantic segmentation employ different evaluation approaches. Anomaly detection commonly uses pixel-level AUROC (Area Under the Receiver Operating Characteristic curve) and AUPR (Area Under the Precision-Recall curve), which assess detection accuracy across thresholds. As we ultimately framed damage detection as a segmentation problem, we adopted standard semantic segmentation metrics used throughout this thesis: mean Intersection over Union (mIoU), measuring overlap between predicted and ground truth segments while accounting for both false positives and false negatives; F1 score, representing the harmonic mean of precision and recall; and Pixel Accuracy, measuring the proportion of correctly classified pixels. In Chapter 3, we also use Earth mover's distance (EMD) to provide a measure of segmentation quality that is less sensitive to exact pixel overlap, particularly valuable when evaluating long, narrow features such as scratches and hairs. For unsupervised settings, we additionally employed the Hungarian matching algorithm and oracles to align predicted segments with ground truth regions before computing these metrics, following standard practice in unsupervised segmentation evaluation [97, 81, 144, 197]. This approach still suffers from an inherent "objectness" problem, which arises when a model segments at different granularities than the ground truth. For example, breaking a single "human" region into head, torso, arms and legs, or a "damage" region into multiple semantically coherent subregions, like "cracks" and "material loss". In such cases, Hungarian matching may only pair one of these sub-segments with the ground truth, treating others as false positives despite their semantic correctness. This is acknowledged in detail in the Appendix of Chapter 5.

2.3.7 The data challenge

Addressing the detection challenge through machine learning approaches faces a significant additional obstacle: the data problem.

Limited Availability of Paired Data: Authentic examples of damaged analogue media paired with their undamaged counterparts are exceedingly rare. Unlike digital degradations that can be artificially applied and reversed as in super-resolution, colourisation, inpainting, physical damage to analogue media is typically permanent, with the original state lost to time.

Complexity of Damage Simulation: Simulating representative training data for the diverse spectrum of analogue media damage is significantly more complex than for digital degradations.

Initial approaches (including the preliminary work presented in Section 2.4) naively simulated film damage by overlaying simplified patterns onto clean scanned frames [204, 140]. While computationally convenient, this approach oversimplifies real analogue damage and fails to capture its complex, medium-specific characteristics, including the interaction between damage and underlying content.

Medium-Specific Variations: Different analogue media types (film emulsion, canvas, parchment, glass plates) exhibit distinctive damage patterns governed by their physical properties. Creating accurate simulations requires modeling these material-specific behaviors—a task which requires formally defining a damage taxonomy across media types.

Limitations of Existing Analogue Media Datasets: The ML community has created numerous datasets for training and evaluating computer vision models on artworks and historical media, but these rarely focus on damage analysis. Datasets like WikiArt [175], OmniArt [190], and Met Faces [102] target tasks such as style classification, artist attribution, or aesthetic evaluation rather than damage detection. A recent trend is the emergence of multimodal art datasets, such as ArtQuest [22]. Even specialized historical document datasets like IAM-HistDB [66] and H-DIBCO [150], which include binarization challenges for degraded documents, focus primarily on text extraction rather than comprehensive damage restoration. Table 2.1 provides a summary of existing analogue media damage datasets in the literature. The majority [50, 204, 137, 141] do not utilise damage annotations at all, and only evaluate damage segmentation [50], end-to-end restoration [204] and inpainting [137, 141] qualitatively. When damage annotations do exist [140, 191, 219], they are limited to binary masks without distinguishing between damage types. All datasets are constrained to a single medium, or even to a single painting, as in Cornelis et al. [50]. Only two of the datasets [191, 219] have been made openly available, but both are small in scale: for instance, while Strubel et al. [191] state their dataset consists of over 2000 annotated damaged grayscale film scans, that is, in fact, the number of 256×256 pixel cropped *patches*, rather than entire images.

While large-scale digitized collections of analogue media exist beyond datasets compiled specifically for machine learning tasks, they are not curated with damage analysis or restoration in mind. Collections such as the Flickr Heritage Commons, Europeana Photography, the Library of Congress Prints and Photographs Collection, and various museum digitization projects offer extensive archives of historical photographs, artwork, and documents. However, these catalogs typically lack critical information for damage restoration research, as they rarely include any metadata localizing damage regions or classifying damage types in a consistent taxonomy. Furthermore, they often contain only the final, post-processed versions of digitized media, with basic restoration already applied. Such collections undoubtedly provide valuable resources for understanding the diversity of analogue media and their content, but their utility for developing and evaluating damage detection models is severely limited by the absence of ground-truth damage annotations of any kind.

Table 2.1: Analogue media damage datasets predating this thesis. Most are small, qualitative, or lack annotations. Only two are made publicly available by the respective authors.

Dataset	Medium	Damage	Images	Annotations	Public
Cornelis et al. [50]	Ghent Altarpiece (painting)	Cracks	1	None (qualitative)	✗
Strubel et al. [191]	Grayscale film patches	Dust, scratches, hairs, dirt	2,000+	Binary	✓
Wan et al. [204]	Photographs	Dust, scratches, fading	5,000+	None (qualitative)	✗
Mironica [140]	Colour film	Dust, scratches	500	Binary	✗
Xu et al. [219]	Frescoes (single site)	Cracks, flaking, fading	1,000	Binary	✓
Merizzi et al. [137]	Frescoes (regional)	Peeling, cracks	2,600	None (qualitative)	✗
Moral-Andres et al. [141]	Roman mosaics	Missing tesserae	8	None (qualitative)	✗

In Chapter 3 and Chapter 4, this limitation was addressed by manually curating and annotating images (first for a single medium, then expanding to multiple types of media), identifying examples with visible damage through careful search and visual inspection. This labor-intensive process was necessary because, even when unprocessed scans with visible damage are available, they lack the pixel-level annotations required for quantitatively evaluating damage detection algorithms.

Even with these valuable benchmark datasets, supervised ML models rely on large quantities of labeled training data face inherent limitations in this domain. The limitations in both damage detection and data generation became the primary focus of this thesis. This directly motivated exploration of alternative approaches, such as:

- Developing more physically accurate damage simulation methods to generate training data (Chapter 3).
- Investigating whether existing methods generalise across media from limited authentic examples based on large pre-trained models (Chapter 4).
- Leveraging weakly supervised, semi-supervised, or unsupervised learning techniques to detect damage (Chapter 5) .

The preliminary method discussed in Section 2.4 naively simulated film damage by overlaying simplified patterns onto clean scanned frames, creating training data analogous to super-resolution or colourisation. This established a baseline for the capabilities of ML approaches if directly applied to restore analogue media damage. While computationally convenient, this approach oversimplified real analogue damage and bypassed the core detection challenge. However, these initial explorations were a valuable first step, demonstrating ML's potential for visual restoration beyond DSP.

2.4 Preliminary work in film damage restoration

This thesis builds upon a pilot study conducted as part of my undergraduate degree, which aimed to develop a machine learning approach that could effectively restore analogue film damage. The core contribution was the modification of existing perceptual loss functions by incorporating earlier layers of pre-trained feature extractors, hypothesized to be more sensitive to the characteristics of film damage compared to losses utilising only deeper layers, as used in style transfer perceptual loss formulations. A U-Net network was trained to map from synthetically damaged analogue film scans to their clean counterparts with this novel loss formulation. Situating this approach within the literature and assessing its potential began with conducting a systematic, albeit constrained, comparative evaluation against existing methods.

2.4.1 Data challenges

An immediate challenge encountered during this initial stage was the limited availability of relevant datasets. Due to the lack of real damaged data paired with target restorations for testing, evaluation at this stage was only possible via a rudimentary approach to approximating film damage (dust/scratch textures overlaid on clean scans). This synthetic damage, while allowing for initial experiments to be conducted, was recognized as a significant simplification of the complexities inherent in genuine film degradation.

A further methodological contribution was the deliberate evaluation on clean, undamaged film images in addition to the naively "dustified" ones. Observing model responses to both damaged and undamaged inputs revealed a previously overlooked aspect of restoration quality: how restoration methods behave when no intervention is actually needed.

2.4.2 Comparative evaluation

The preliminary experiments benchmarked the proposed perceptual loss approach against "Bringing Old Photos Back to Life" [204], which was the sole directly comparable approach in the literature with publicly available model weights, alongside a simple baseline trained with a Mean Squared Error Loss.



(a) Input (b) Wan et al., 2020 (c) Ours (d) Target

Figure 2.10: Qualitative comparison between our preliminary end-to-end restoration approach [94] and "Bringing Old Photos Back to Life" [204].

SSIM was chosen to quantify changes in structural information and perceptual phenomena relative to an undamaged reference image, providing a metric for assessing restoration quality. Published in [94], the relevant quantitative results are summarized in Table 2.2.

Table 2.2: SSIM Scores obtained by a simple MSE baseline, "Bringing Old Photos Back to Life" [204], and our preliminary perceptual loss approach [94].

	SSIM on dustified images		SSIM on clean images	
	mean	std dev	mean	std dev
MSE Baseline	0.84	0.07	0.95	0.04
Wan et al.	0.68	0.06	0.72	0.07
Ours	0.88	0.05	0.98	0.02

They pointed to detection as the bottleneck: first, the perceptual loss approach achieved near-perfect preservation of clean images (0.98 SSIM), suggesting it effectively learned when not to modify the input; second, 'Bringing Old Photos Back to Life' not only missed when present, but frequently modified regions that contained no damage at all (0.72 SSIM on clean images), indicating its primary failure was in incorrectly identifying what needed restoration. These observations are supported qualitatively as well, with examples shown in Fugure 2.10.

2.4.3 Takeaways

These early experiments, despite limitations, yielded insights which directly motivate the rest of this thesis. Analysing the results from this phase clarified essential requirements for effective analogue damage restoration:

1. Models must only restore genuinely damaged areas, preserving undamaged regions.
2. Following from 1., models must detect damage correctly, not mistaking useful image features for artifacts

Furthermore, the constraints of our initial evaluation, particularly the use of naively synthesized damage limited to black artifacts on positive film scans, implicitly pointed towards a third, crucial requirement:

3. Models must generalise to different types of damage beyond their training examples.

Addressing and rigorously evaluating against these three requirements demanded significant advancements. Specifically, the limitations of using simplistic synthetic damage and the lack of comprehensive benchmarking against real-world film artifacts highlighted the need for:

- A benchmark dataset of real damaged film scans at realistic scale, paired with expert restorations, to address the lack of authentic evaluation data.
- A statistically grounded damage simulator, facilitating training of detection and end-to-end approaches despite limited real-world data.

- Decoupled evaluation of detection and detection as a precursor to inpainting, suspected as a major performance bottleneck.
- Benchmark diversity across film and damage types, enabling generalisation assessment.

These needs directly motivated some of the core contributions of this thesis, beginning with the development of a comprehensive benchmark dataset and a statistically grounded damage simulator, detailed Chapter 3. These, in turn, enabled the adaptation and evaluation of both segmentation and end-to-end restoration approaches, confirming the primary challenge of damage detection and paving the way for the systematic investigation of detection methods for diverse analogue media, which is the focus of Chapter 4.

Chapter 3

Simulating analogue film damage to analyse and improve artefact restoration on high-resolution scans

The majority of this chapter (Section 3.2 through to Section 3.8) reproduces the author's version of the accepted, peer review manuscript presented as a full paper (oral) at the The 44th Annual Conference of the European Association for Computer Graphics (Eurographics) as *Ivanova, D., Williamson, J. and Henderson, P., 2023, May. Simulating analogue film damage to analyse and improve artefact restoration on high-resolution scans. In Computer Graphics Forum (Vol. 42, No. 2, pp. 133-148).* DOI: 10.1111/cgf.14749 and does not violate the copyright of the publisher. The totality of the work presented below was undertaken by myself. I scanned and manually annotated, extracted and statistically analysed all 12,000 real film artefacts used in the damage generation pipeline. I formulated and carried out all experiments involving the benchmarking of segmentation and restoration models discussed as part of the empirical assessment of the impact of synthetic damage in downstream tasks. I created and disseminated the two perceptual surveys which evaluate the realism of the synthetic damage, and processed and interpreted the collected responses. All figures and visualisations were created by me. I drafted and rewrote the paper, with my supervisors Dr John Williamson and Dr Paul Henderson providing edits, corrections, and supervision. The 35mm film images for the "in-the-wild" test set have been shot, developed and scanned by Dmitri Tcherbadji of Analog.Cafe. The synthetic artefacts used in the damage generation pipeline have been provided by Stefan Ringelschwandtner of Mononodes.

3.1 Summary

Our preliminary work in film restoration (discussed in Section 2.4)) conceptualised the problem as an end-to-end image mapping task in pixel space, leveraging supervised learning with tailor-made perceptual losses. While this approach demonstrated promise, rigorous evaluation was infeasible due to the absence of open benchmark datasets containing authentic film damage, paired with expert restorations. Consequently, validation was limited to synthetically damaged data generated through basic overlay techniques, leaving questions about generalisation to real-world scenarios unanswered. Furthermore, the end-to-end framework obscured performance analysis, as restoration quality metrics alone cannot distinguish between different types of failure. When comparing model outputs against ground-truth un-damaged images, poor performance could arise from either detection failures, where the model misses damaged regions, identifies incorrect regions as damaged, or produces imprecise damage boundaries, or from restoration failures, where correctly identified damage is repaired poorly. These distinct failure modes are conflated in standard evaluation metrics, making it difficult to identify which aspect of the model needs improvement.

These methodological limitations inspired several key contributions. The first was a benchmark dataset of real damaged high-resolution film scans paired with expert restorations, which enabled proper quantitative evaluation. Experiments revealed two major limitations of existing end-to-end approaches. First, their restoration capabilities were inconsistent: while effective at removing small high-frequency artifacts like dust specks, they also removed desirable high-frequency details like film grain, and struggled with larger or more complex artifacts such as hairs and scratches. Second, while these methods could produce visually appealing results, they frequently introduced unwanted changes to undamaged regions—shifting colors and altering the overall character of the image.

This motivated breaking restoration into two distinct sub-tasks: damage detection (segmentation) followed by repair (inpainting). Decoupled evaluation demonstrated that state-of-the-art inpainting methods were already remarkably effective at damage repair *if* provided with precise damage masks. Hence, the primary challenge lay with *accurate damage detection*—a re-framing of the task which pertains to the remainder of this thesis.

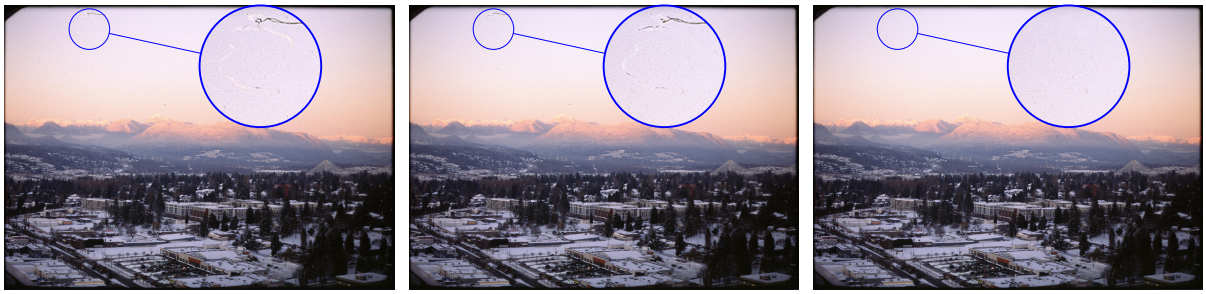
So, how do we improve damage detection? Our hypothesis is that better simulation leads to higher quality synthetic data, which in turn improves test results on real data. Hence, we focused our efforts on simulation and collected, annotated, extracted and analysed thousands of real film artifacts. Empirical evaluation on the authentic damage benchmark dataset demonstrated that training both segmentation networks and end-to-end restoration models on our synthetic data yielded substantial improvements across standard metrics, and produced superior qualitative results in comparison to models trained on alternative synthetic data.



(a) ***Input****: 4K film scan with authentic damage (hairs and dirt).*

(b) ***Artefact Segmentation****: prediction from our U-Net trained on synthetically damaged data.*

(c) ***Restoration by BOPB [204]****: using our predicted segmentation.*



(d) ***Restoration by U-Net [94]****: retrained on our synthetic damage.*

(e) ***Restoration by LaMa [192]****: best performing model, using our predicted segmentation.*

(f) ***Ground Truth****: manually restored in Photoshop by a human expert.*

Figure 3.1: Input and ground truth from our authentic artefact damage dataset, along with restorations from some of the models we evaluated, presented at full resolution. Ours is the first public dataset allowing accurate quantitative evaluation of film damage restoration.

3.2 Abstract

Digital scans of analogue photographic film typically contain artefacts such as dust and scratches. Automated removal of these is an important part of preservation and dissemination of photographs of historical and cultural importance. While state-of-the-art deep learning models have shown impressive results in general image inpainting and denoising, film artefact removal is an understudied problem. It has particularly challenging requirements, due to the complex nature of analogue damage, the high resolution of film scans, and potential ambiguities in the restoration. There are no publicly available high-quality datasets of real-world analogue film damage for training and evaluation, making quantitative studies impossible. We address the lack of ground-truth data for evaluation by collecting a dataset of 4K damaged analogue film scans paired with manually-restored versions produced by a human expert, allowing quantitative evaluation of restoration performance. We have made the dataset available at https://doi.org/10.6084/m9.figshare.21803304. We construct a larger synthetic dataset of damaged images with paired clean versions using a statistical model of artefact shape and occurrence learnt from real, heavily-damaged images. We carefully validate the realism of the simulated damage via a human perceptual study, showing that even expert users find our synthetic damage indistinguishable from real. In addition, we demonstrate that training with our synthetically damaged dataset leads to improved artefact segmentation performance when com-

pared to previously proposed synthetic analogue damage overlays. The synthetically damaged dataset can be found at https://doi.org/10.6084/m9.figshare.21815844, and the annotated authentic artefacts along with the resulting statistical damage model at https://github.com/daniela997/FilmDamageSimulator. Finally, we use these datasets to train and analyse the performance of eight state-of-the-art image restoration methods on high-resolution scans. We compare both methods which directly perform the restoration task on scans with artefacts [94, 204, 87, 89], and methods which require a damage mask to be provided for the inpainting of artefacts [192, 229, 167]. We modify the methods to process the inputs in a patch-wise fashion to operate on original high resolution film scans.

3.3 Introduction

Photographs captured on film constitute a major part of our cultural heritage and historical record, and many photographers continue to shoot on film. While it has appealing imaging qualities, film emulsion is highly susceptible to various kinds of mechanical damage: scratches, dust, hairs and dirt which mar the image when scanned and persist in darkroom and digital prints, as well as in cinematographic release prints.

Isolating dust and scratches from natural image features in film scans and inpainting the damage are not difficult tasks for humans, but can be extremely time-consuming. **Automated film restoration** aims to localise and inpaint artefacts at a quality level comparable or exceeding that of human experts.

This is challenging: mechanical artefacts can have complex shapes and are non-uniformly distributed. They must be localised accurately when performing restoration, as it is crucial that restoration is applied only on the affected areas. It is also crucial to preserve the desirable qualities of analogue film, such as grain or characteristic colour grading, avoiding over-smoothing or distorted colour distributions.

Automated film restoration can be broken down into two sub-tasks: artefact localisation (segmentation), and inpainting. Inpainting [123, 192] and segmentation [168, 10, 139] are well-studied in the literature, but relatively few models have been published that are suitable for professional-quality analogue film artefact restoration. Many existing general inpainting approaches are unsuitable as they operate at much lower resolutions than required for film scans (typically at least 4K), and rely on a mask being provided. Traditional solutions employed in commercial film scanners, such as Kodak's Digital ICE [177, 187, 71], use a separate infrared illumination process to create an inpainting mask and apply simple nearest neighbour inpainting. This requires specialist hardware, only works on a limited number of emulsions, and cannot detect every form of artefact. We focus instead on a purely image-based automated film restoration process which works for all emulsion and damage types without additional imaging.

Large datasets with realistic film damage to train and evaluate on are not easily available

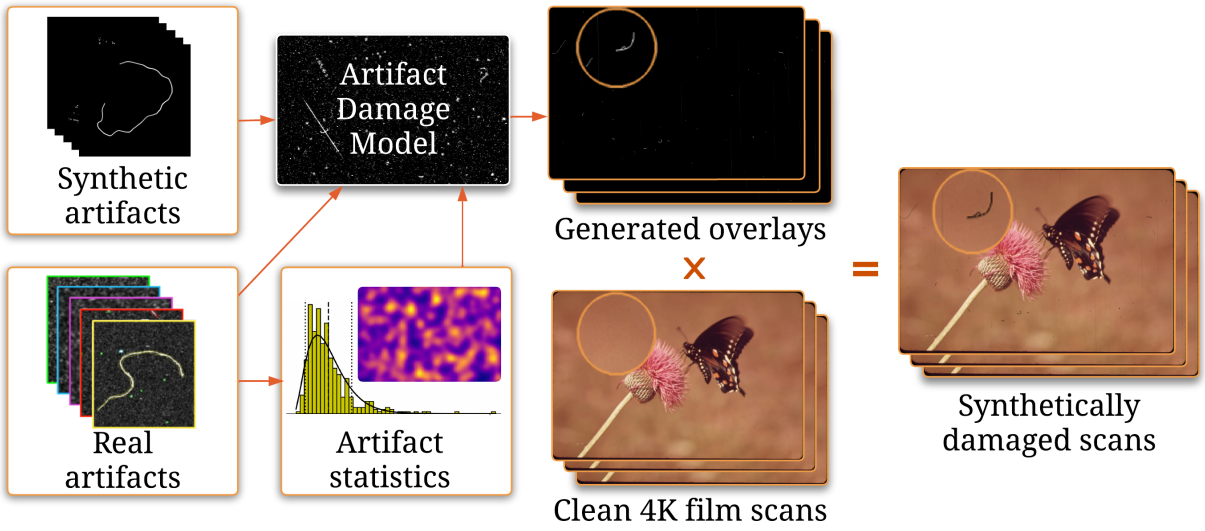


Figure 3.2: Overview of our proposed approach to generating realistic damage overlays showing analogue film artefacts (Section 3.6). We extract statistics from 12 000 manually-annotated artefacts, and use these to build a probabilistic model of artefact damage that can be applied to clean film scans. The resulting damage overlays are highly realistic, and suitable for training image restoration models.

[33] and most ML-based film restoration instead relies on naïvely generated synthetic images to approximate analogue artefacts[89, 140, 204, 94]; systems for which authentic damaged images *have* been collected have typically not made their data publicly available [140, 204]. Unlike tasks such as JPEG artefact removal, super-resolution or colourisation, there are no robust simulation models of analogue film damage to generate high-quality synthetic training samples. There is also no consensus as to what makes a "good" restoration, and therefore how a film restoration model should be evaluated, especially without human-restored ground-truth scans [34, 32].

Contributions

We make three contributions that address these issues:

1. We curate and release a dataset of real analogue film scans (Section 3.5); the dataset consists of 4K scans of a variety of positive and negative colour emulsions, with damage artefacts such as dust, scratches, dirt and hairs present, paired with corresponding professionally manually restored versions. This is the first public dataset enabling quantitative evaluation of restoration models.

2. We construct a detailed statistical model of analogue film damage, allowing us to create highly-realistic synthetic data for training (Section 3.6). To estimate the parameters of this statistical model, we collect another dataset of scanned empty film emulsion which has been scratched and left to get dusty; each damage artefact in the scans is manually annotated, resulting

in a set of over 12,000 unique artefacts. We extract statistics governing the shape, size and spatial density of different types of artefacts, which allows us to generate realistic synthetic analogue damage overlays to train restoration models. We validate the resulting statistical model through a human perceptual study. In addition, we release a dataset of analogue image scans paired with their synthetically damaged versions, which we use to fine-tune or re-train models.

3. We use these datasets to perform a detailed empirical comparison of seven state-of-the-art models on film artefact restoration (Section 3.7). We include specialised film artefact restoration models, and general models that can be adapted for this purpose. We perform evaluation on our dataset of images with real analogue damage and manually restored ground truths. We address the low-resolution limitations of existing methods that make them unsuitable for professional-grade restoration and extend processing to analogue film scans at 4K resolution by applying them patch-wise. Our synthetically damaged dataset allows us to train a segmentation network to supply damage masks to inpainting models that explicitly require them.

3.4 Background and related work

In this section we discuss definitions of film damage in the literature, and review existing machine learning approaches to film artefact damage restoration. Additionally, we consider state-of-the-art machine learning solutions to similar tasks which can be adapted to the film artefact restoration problems, and the challenges involved in doing so.

3.4.1 Film damage

Analogue film damage is a loose term covering a variety of degradation types. Chambah [32] roughly categorises analogue damage into two groups: chemical degradations and mechanical degradations. Chemical degradations, such as colour dye fading, contrast saturation and vinegar syndrome, are usually spatially homogenous [32, 204]; grain is also spatially homogenous, but while some works consider it to be another type of degradation [204], others point to it as an example of the ill-defined difference between film damage and artistically valued properties of the medium [32, 94]. On the other hand, mechanical degradations, for example abrasions to the emulsion (scratches) or dust specks or hairs stuck to the emulsion, are not spatially uniform [32, 204].

As film is the source record for printing images, artifacts which affect the photochemical emulsion will be carried over to any prints made from it [32]. Therefore, damaged film scans and the resulting damaged reproductions can be restored via the same techniques. However, further damage inflicted on the print itself, such as tears and folds, is independent of the emulsion. While some works consider film and print degradations interchangeable [204], it is notable that they have distinct properties in terms of scale and spatial distributions with respect to the medium in

which they originally occur. Furthermore, print degradations are not unique to prints of images captured on film, and as such are beyond the scope of this work.

Removal of any type of degradation can be reframed as an image-to-image translation problem, where some degradation operation has been applied to the original, non-damaged image, and needs to be inverted, similar to denoising [114, 228, 230, 120, 38], superresolution [99, 206, 120], and colorisation [236, 91, 111, 174]. For these tasks, the conventional approach is to generate training data from clean images by applying a transform which is identical to the degradation process, e.g. with colourisation, grayscale versions of the training data images can be easily derived from color images. Yu et al. [225] propose a reinforcement learning approach to modelling more complex degradations as a dynamic mix of simpler types of damage, however, it is important to note that the most common degradations in the image denoising, restoration and in-painting tasks addressed in the literature happen entirely in the digital domain. Film damage, and especially mechanical film damage, in contrast, is a product of the physical properties of the film emulsion [32], and is translated into the digital domain during the film scanning process. Due to their analogue nature, lack of spatial uniformity, variability in opacity, and randomness of shape and size, analogue artefacts such as scratches, dust specks and hairs are more challenging to model, and by extension, digitally simulate and restore [32, 204, 140, 94].

Film damage has been simulated naïvely by directly compositing a small set of full frame damage textures over clean images [89, 204, 140, 94]. Although transforms such as resizing, rotating, flipping and randomly cropping the damaged textures are employed in order to introduce more diversity in the synthetic data, to our knowledge, no approach in the literature varies the shape, location, and/or rotation of individual artefacts on the emulsion, nor their size with respect to the contents of the image and the film frame.

3.4.2 Inpainting

Inpainting is a task in which missing regions in the image, indicated by a binary mask, are filled in by estimation based on neighbouring pixel and global image context. The task lends itself to being reframed as part of the analogue artefact restoration problem: image areas obscured by the artefacts need to be suitably inpainted. However, there are some challenges: artefacts have arbitrary shape, size and location, which, in addition, are not known a priori. Recent advances in deep learning have enabled inpainting of larger image regions by generating semantically consistent content [192]. The task has been tackled by various families of models, such as VAEs [80, 156], GANs [224], and diffusion models [174, 167, 128]. While earlier approaches are constrained in the mask shape applied for inpainting, some recent models consider arbitrary masks. Another related task is single-image de-raining, which has been tackled by CNN-based approaches [163], and more recently, by transformer-based ones, such as Restormer [230], which is currently state-of-the-art. Blind image inpainting is an extension of the inpainting task, where the mask indicating areas to be inpainted is not provided [209, 28]. Hertz et al.

[87] propose a deep learning approach which attempts to address this problem in the context of watermark removal, by also predicting a mask to separate out the areas to be inpainted from the ares which are to be preserved. In line with this method, state-of-the-art inpainting approaches can be adapted to blind image inpainting via an additional segmentation network tasked with predicting masks for the inpainting network to use.

3.4.3 Processing of high-resolution data

Modern camera-equipped devices are able to capture images at very high resolution (i.e. mega- and even gigapixels). The exact definition of "high resolution" varies between applications [11]; in film scanning, it is common to scan images at 3000DPI and at least 4K resolution. Many of the image-based tasks that have drawn attention in the machine learning community in the last few years do not require for the image input to be processed at full resolution – e.g. classification models are conventionally trained on cropped and downsampled images of size 224×224 pixels to meet computational limitations. Talebi & Milanfar [193] even investigate learning the resizing operator on the input to improve network performance while maintaining the constraint for lower resolution image input. On the contrary, few state-of-the-art image restoration approaches, e.g. Restormer [230] and LaMa [192] claim the ability to process high resolution image data. When applying machine learning approaches to problems in fields such as medical imaging, processing large medical image scans at their original resolutions is crucial. The standard approach is to split the images into patches, have the model process them, and stitch them back together [158]. Processing film scans poses a very similar problem.

3.4.4 Deep learning for analogue film restoration

Due to the challenges in modelling film damage and collecting relevant data, there are few approaches in the literature which set out to solve the specific problem of film artefact restoration. Strubel et al. [191] train a SegNet model to remove dust and scratches, and provide a limited dataset of grayscale image scans for training and evaluation. Mironica [140] proposes a GAN-based approach to film artefact restoration; to generate training data, they use a set of 100 synthetic overlays applied over a set of 2500 clean film scans, which are randomly cropped to patches of 128×128 pixels. The qualitative results in the paper demonstrate some success in inpainting smaller artefacts, at the expense of overly smoothing grain. Similarly, Wan et al. [204] address the problem of limited training data by applying synthetic damage overlays onto digital images, along with synthetic grain to mimic film scans and prints. They train a segmentation U-Net to predict damage masks, and further jointly train two VAEs to translate images between two latent spaces, corresponding to the domains of damaged and restored photos, respectively. The model is trained on cropped patches of size 256×256 pixels. Since the training data is derived from digital images, the restorations produced by this approach have

the effect of overly smoothing grain and shifting color in actual film scans from the test set, i.e. the network itself introduces loss of information to the input, including novel artefacts. A follow-up work by the same authors extends the approach to analogue video sequences [203]. Similarly, DeepRemaster [89] also applies pre-rendered damage overlays to sequence frames to model analogue damage. Finally, Ivanova et al. [94] propose a U-Net restoration network trained with a perceptual loss, which tackles the artefact detection and restoration tasks simultaneously. As with previous works, film damage is again simulated by applying a set of overlays, modified by various simple transforms, over clean analogue film scans. While this restoration network is qualitatively and quantitatively shown to outperform the method of Wan et al. [204], it is limited in the resolution of inputs which it can be applied to, due to being trained on images downsampled and cropped to 256×256 pixels.

3.5 Evaluation data for damage restoration

We curate a dataset of 35mm film scans at 4K resolution, with varying degree of authentic artefact damage in the form of dust, scratches, hairs and dirt. The dataset includes 44 images of various film emulsions, both positive (slide) and negative. The content of the images is also diverse, including landscapes, architecture shots and still lifes. The images have been shot, developed and scanned by Dmitri Tcherbadji of Analog.Cafe, and are used with his kind permission. In addition, each damaged image in the set has been paired with a ground-truth restoration via manual inpainting of the artefacts in Photoshop by the same expert. Original and manually-restored versions of an example photograph from the dataset are shown in Figure 3.1. To our knowledge, this is the only public dataset of high quality damaged film scans paired with expert restoration ground truths. We have made the dataset available at www.doi.org/10.6084/m9.figshare.21803304.

3.6 Modelling analogue film damage

In this section, we describe our novel approach to generating synthetic analogue damage. Our overall approach is as follows:

1. Annotate, classify and extract over 12 000 individual real analogue film artefacts from heavily-damaged high resolution scans (Section 3.6.1).

2. Calculate statistics of the extracted artefacts, such as size, count, and spatial density (Section 3.6.2).

3. Build a probabilistic model to generate new artefact damage overlays using a combination of the extracted artefacts and synthetic ones, parameterised by the recorded statistics (Section 3.6.4).

We demonstrate the realism of damage overlays generated by our model, via a perceptual user study (Section 3.6.5). We further validate the damage by evaluating it in the context of artefact segmentation (Section 3.7.1) and artefact restoration. In both cases, the models trained with data synthesised via our approach outperform the alternatives. The full damage synthesis pipeline, including the annotated analogue damage scans, is available at www.github.com/daniela997/FilmDamageSimulator.

Artefact types. Prior works have classified damage as dust, scratches or hairs [32, 204]. We extend this taxonomy by splitting hairs into two classes, short and long, due to their large variation in size; we also define an additional class, dirt, to capture artefacts of irregular shape which are larger than dust specks, but are neither hairs nor scratches.

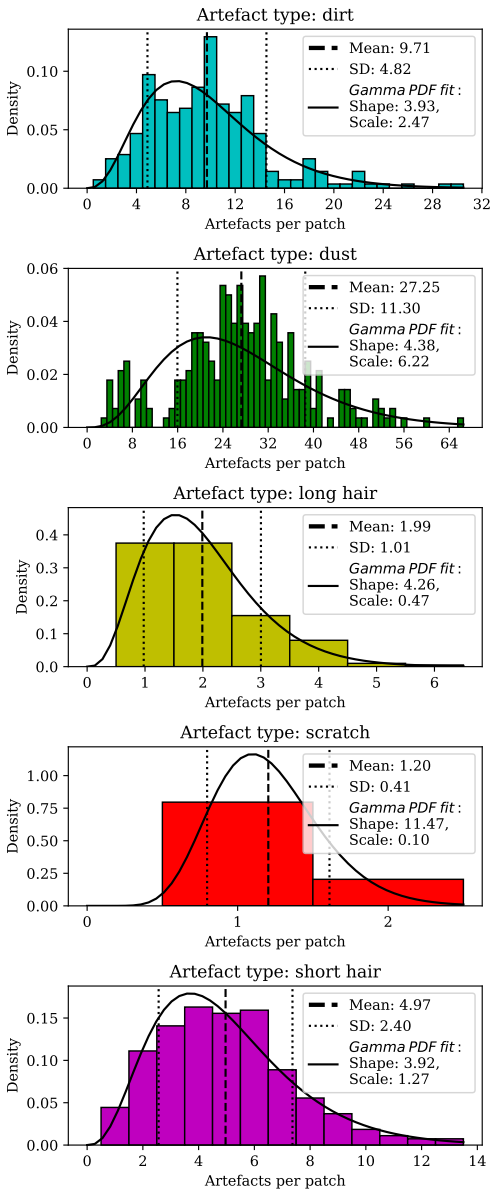
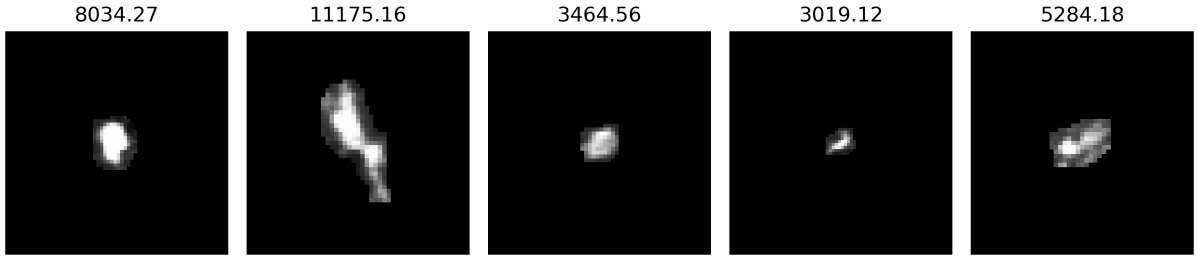


Figure 3.3: Number of artefacts per 256 × 256 pixel patch for each artefact type across all scans, and fitted Gamma distributions.

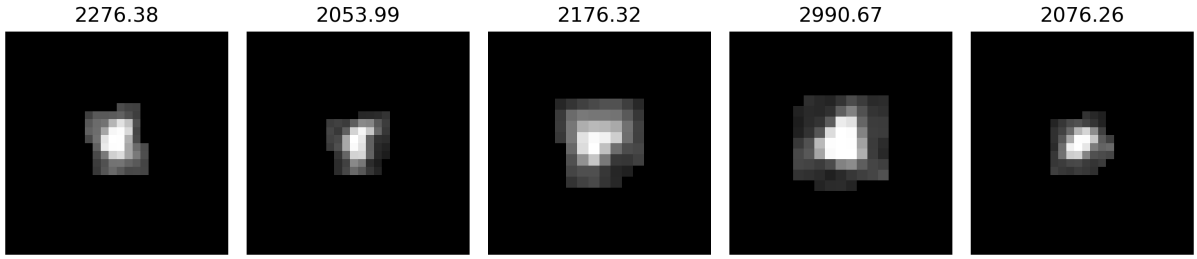
3.6.1 Capturing real analogue film artefacts

To gather samples of real-life dust specks, scratches, hairs and dirt, we scan 10 heavily damaged, empty 35mm film frames of Lomochrome Color Negative 400 ISO film using a Plustek OpticFilm 8100 dedicated film scanner and SilverFast 9. The resulting scans are saved at 4K resolution.

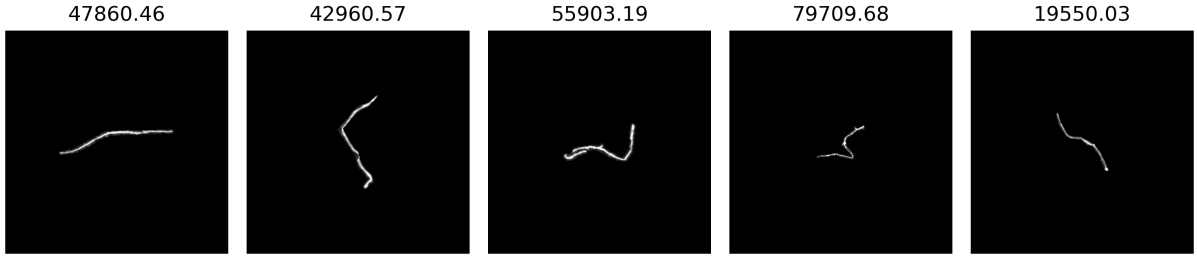
We manually annotate individual artefacts in the scans with bounding polygons, and classify each as dirt, dust, long hair, short hair or scratch (see Figure 3.5 for examples). We calculate the area of each polygon, and convert these to physical units of square microns based on the ratio between the scanned frame's size in pixels, and its size in millimeters – 35mm on the long edge and 24mm on the short edge. Finally, we extract each artefact, zero padded to square, to create a bank of isolated artefacts to sample from when generating new overlays; examples for each class are visualised in Figure 3.4. In total we have annotated 12135 artefacts across the 10 scanned frames.



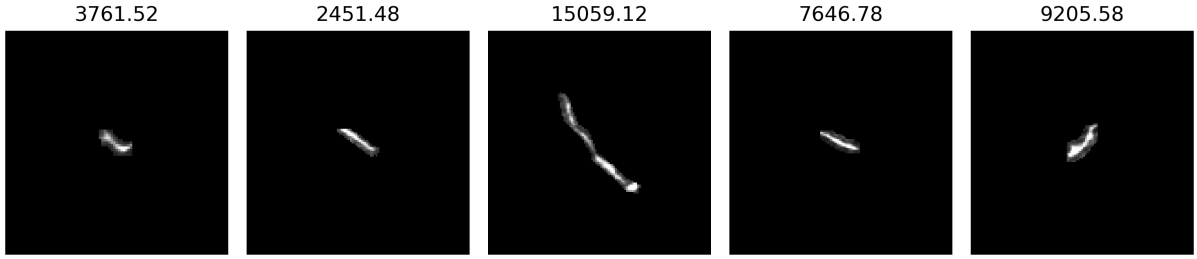
(a) ***Dirt*** *artefacts are irregularly shaped and of varying size.*



(b) ***Dust*** *specks are round and relatively small in size.*



(c) ***Long hairs*** *are thin, overall large in size and of varying shape and extent.*



(d) ***Short hairs*** *are similar to long hairs but smaller in length (and size).*

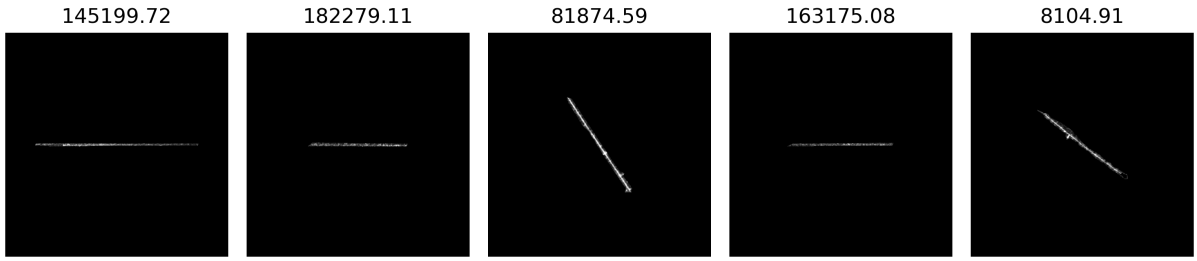


Figure 3.4: Examples of extracted and padded artefacts. The number above each is its area in μm^2 (square microns).

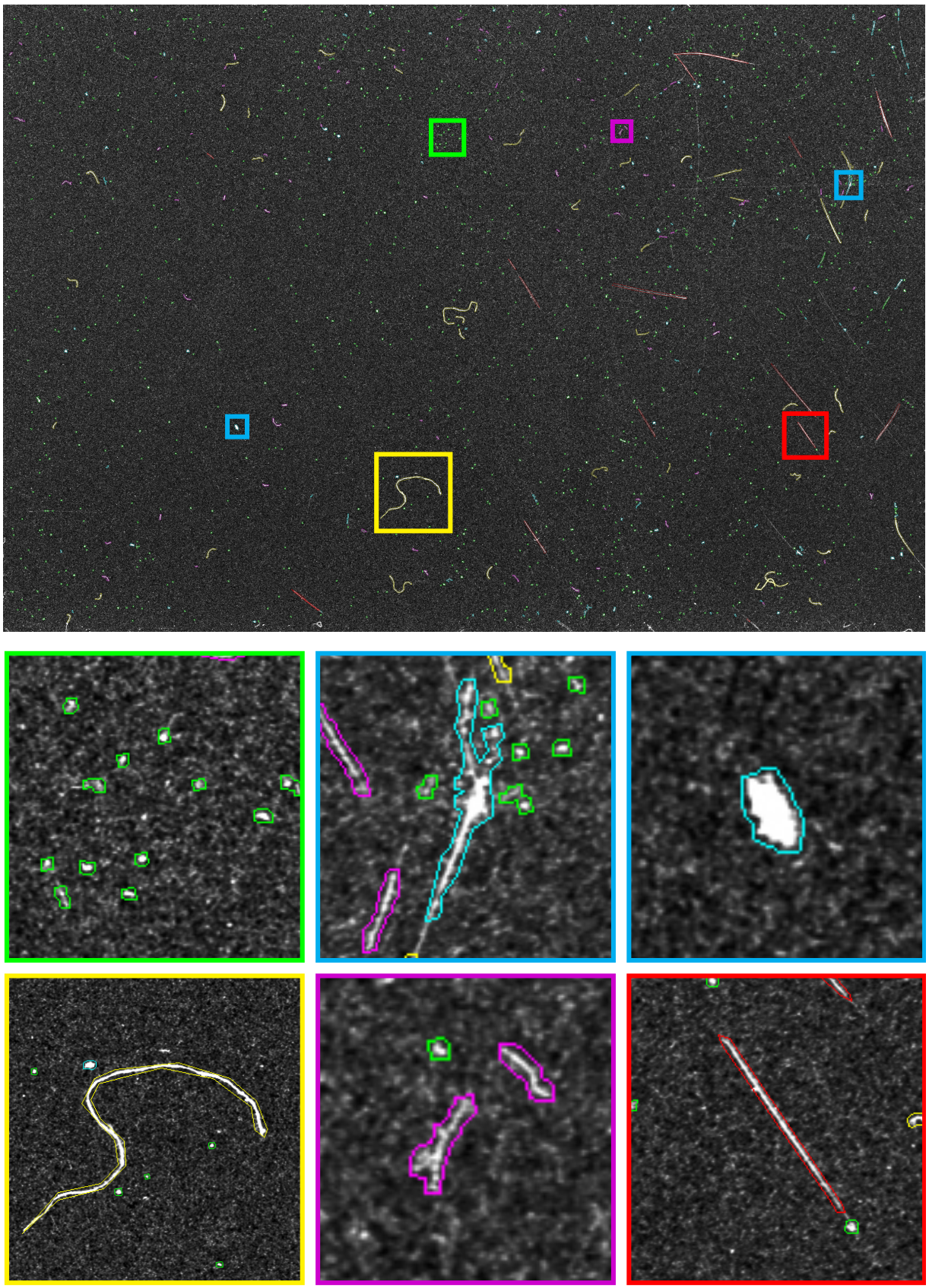


Figure 3.5: An example of an annotated scan. Full resolution (4944 by 3396 pixels) is shown at the top, along with zoomed-in regions around examples of each type of artefact below. Overlaid annotations for dust (green), dirt (blue), long hairs (yellow), short hairs (magenta), and scratches (red).

3.6.2 Analysing artefact statistics

For each artefact class, we collect several statistics to ensure our generated overlays match the distribution of real film artefact damage. We measure individual artefacts (area in μm^2, square microns), as well as their distribution over the entire frame (counts, spatial frequency). Artefact counts and sizes are summarised in Table 3.1; Figure 3.6 displays the full distributions of sizes for each class.

Spatial frequencies. We visualise the spatial distribution of artefacts via kernel density plots, for an example class in Figure 3.7a and for all classes in the supplementary material (Figure S5). While there is no obvious pattern to the distributions of artefact occurrences, we can observe that four artefact types are more frequent near the frame's upper and lower left corners; this could be related to the direction in which the film strip is advanced inside the camera and/or the scanner.

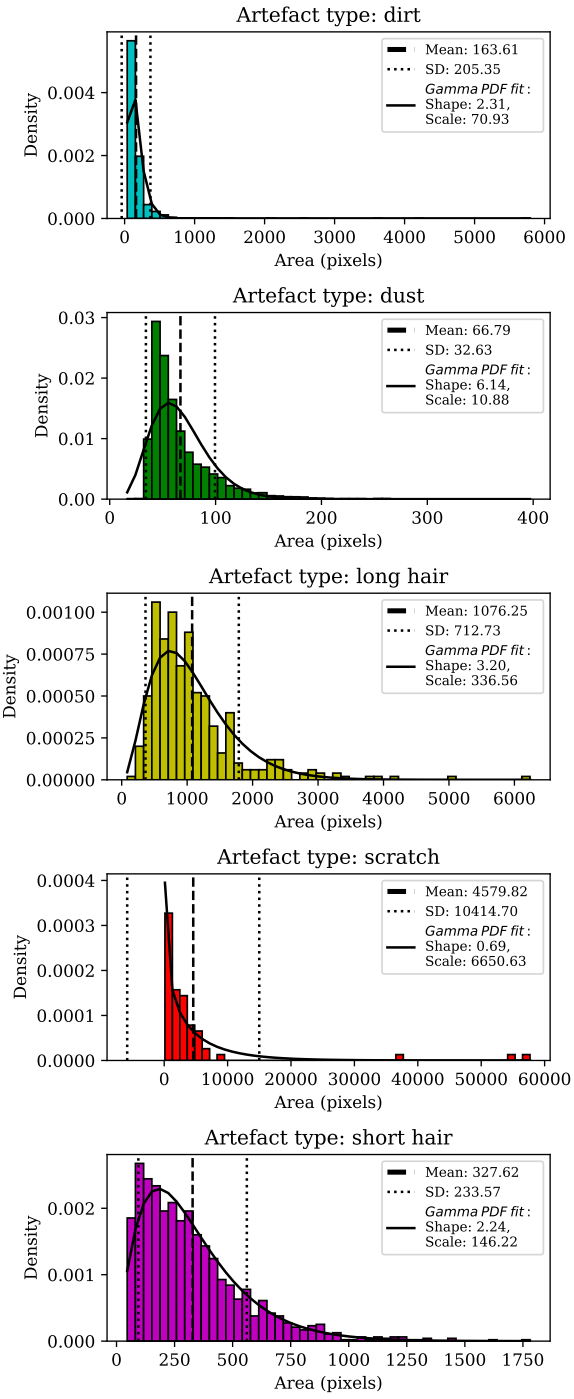
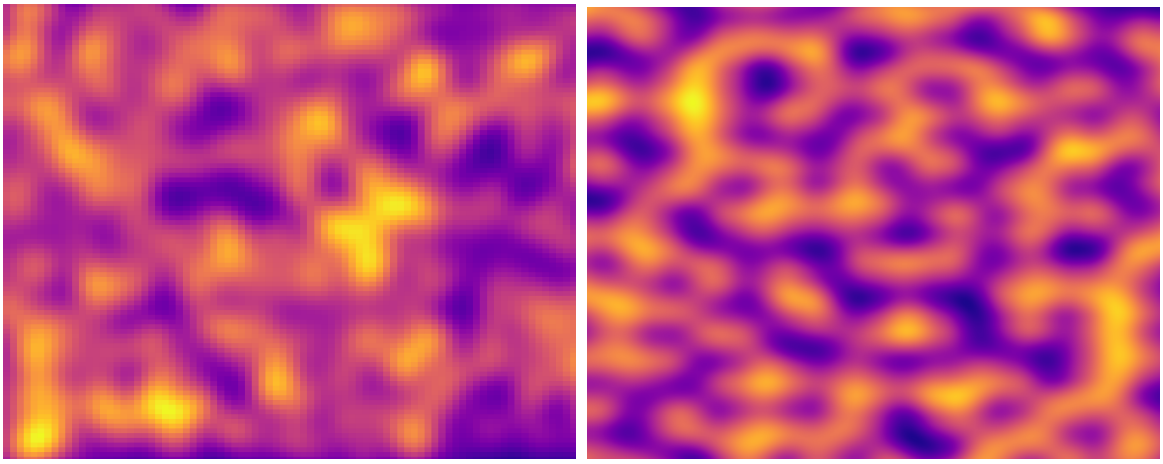


Figure 3.6: Artefact size histograms in pixels, and fitted Gamma distributions. While average sizes for each type of artefact vary, the sizes of artefacts are distributed in a similar way.

3.6.3 Synthetic artefacts.

In addition to the artefacts extracted from the 10 scans, we also use a set of 6100 synthetic dust, scratch, hair, lint, and dirt artefacts, which were manually drawn in Photoshop and kindly shared by Stefan Ringelschwandtner of Mononodes. We do this to further increase the diversity of rarer artefact classes. Every artefact in this set is of size 400×400 pixels, and therefore must be rescaled to match the observed distribution of areas for real artefacts of the corresponding class. As the scratch class is particularly underrepresented, we also programatically generate additional scratches matching the appearance of those found in real scans.



(a) (b)

Figure 3.7: Visual comparison between the true spatial distribution of short hairs (by kernel density estimation on annotation centroids) **(a)**, and Perlin noise **(b)**, which we use as an approximation when sampling new overlays. Each plot shows the distribution over a full 35mm frame, with brighter colors corresponding to higher density.

Artefact type	**Count**	**Avg. area (μm^2)**	**Std. dev. area (μm^2)**
dirt	2700	8194	10304
dust	7631	3344	1637
long hair	398	53501	35524
short hair	1341	16365	11735
scratch	65	229660	521345

Table 3.1: Statistics of annotated artefacts – we report the total count and average area for each artefact class. See Figure 3.6 for the full distributions of areas.

3.6.4 Generating damage overlays

We develop a probabilistic model based on the measured artefact properties, which allows us to generate new synthetic full-frame damage overlays with a realistic distribution of artefacts. Our generative process is as follows:

1. **Sample the numbers of artefacts for each class:** For each artefact class, we sample a target count from a Gamma distribution fit by maximum likelihood estimation to the empirical counts in Figure 3.3 and round to the nearest integer. As the recorded counts are per 256×256 pixel patch, we scale the sampled counts depending on the target overlay resolution.
2. **Sample the artefact sizes:** Similarly, we sample artefact sizes from Gamma distributions fitted to the observed artefact sizes in Figure 3.6. The sampled sizes are further rescaled according to the target overlay size, using the relevant pixel-to-micron conversion ratio.
3. **Sample the artefact appearances:** For each artefact, we randomly choose its appearance from among the real and synthetic artefacts of the relevant class. We ensure that the size of each (in pixels) is not divergent from the target sizes sampled in the previous step, in

order to avoid excessive upsampling of small artefacts in classes with high variance in area (e.g. dirt).

4. **Sample locations and rotations:** We approximate the spatial densities discussed earlier with Perlin noise [157], as it produces smooth, non-repetitive random fields whose scale, frequency, and octaves can be tuned to match the clustering patterns seen in real scans. Perlin noise allows these variations to be expressed coherently, while avoiding the artificial regularity introduced by grid-, sinusoid-, or Gaussian mixture-based sampling. We empirically justify this choice by a visual comparison with the true artefact density aggregated over all scans (Figure 3.7). We sample the location of each artefact independently from this noise distribution; we also sample a rotation from the uniform distribution on $[-\pi, \pi]$.
5. **Compose the final overlay:** Given the artefact appearances, sizes, locations and rotations, we alpha-composite them into the frame at the required output resolution. The overlay can be used to simulate damage to film negatives (artefacts are white), or to developed slides (artefacts are black).

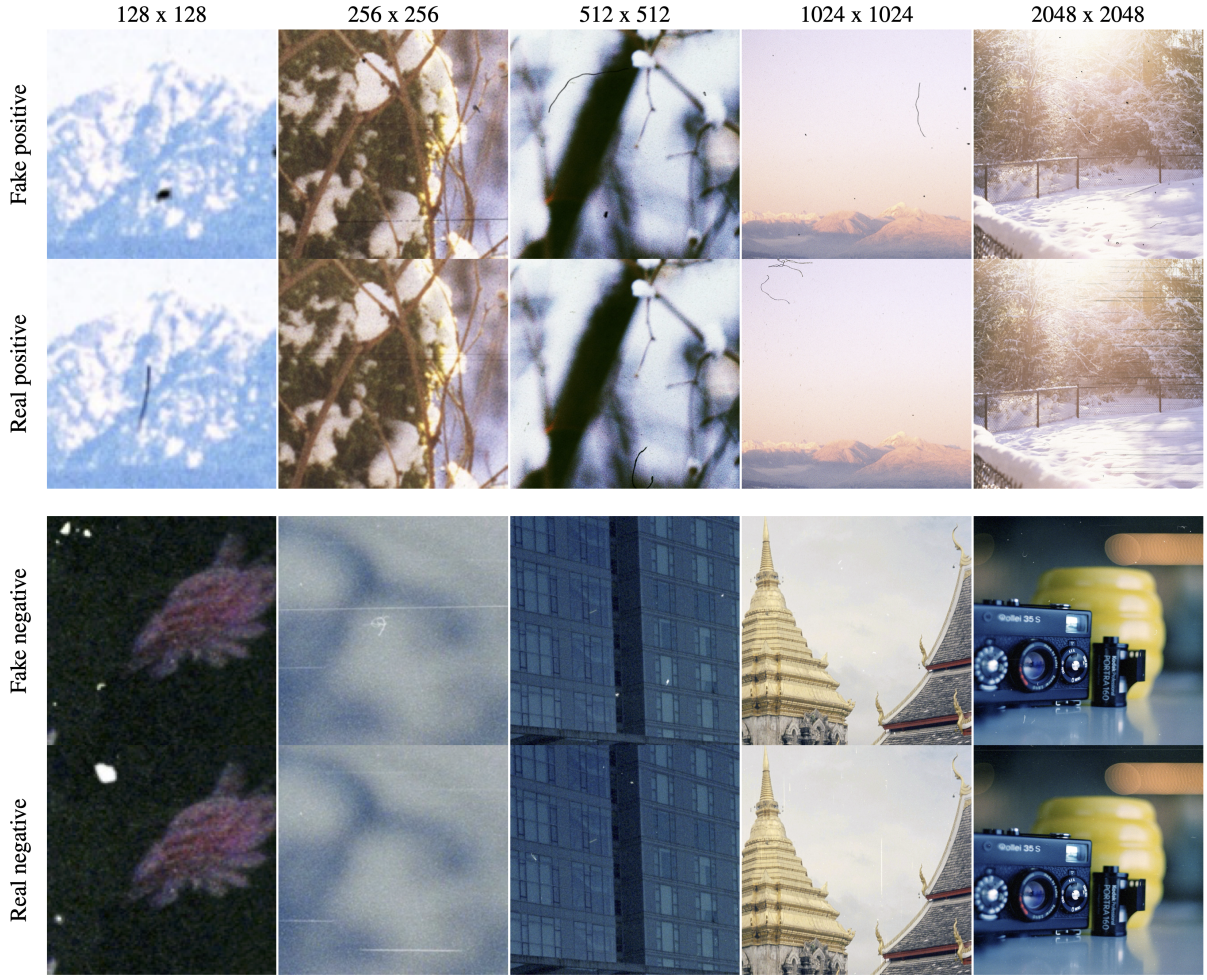


Figure 3.8: Example image pairs (positives and negatives) shown to perceptual study participants at five different crop resolutions (in pixels).

3.6.5 Human validation of generated artefact damage

We conduct two user perceptual surveys to verify that our damage overlay generation process yields realistic results. We discuss the study design and summarise the results below. In both cases, we also ask each participant to self-report their familiarity level with analogue film and its associated artefacts, as 'Not familiar at all', 'Somewhat familiar' or 'Very familiar'.

Real vs. synthetic damage. In the first survey, we compare real analogue damage to the synthetic damage produced by our model. The generated damage is applied to the ground truth restored images from our test set. We pair the synthetically damaged image scans with their real damaged versions. Each participant was shown 100 such pairs (see Figure 3.8 for examples): 50 simulating negative film damage (i.e., white artefacts), and 50 simulating positive (slide) film damage (i.e. black artefacts), and is asked to choose one image per pair which they believe shows real analogue film damage. The images are cropped to five different target resolutions to reflect the relationship between artefacts and image features of varying size.

We collected 251 responses to this first survey. We calculate the percentage of pairs for which each participant chose the example with real artefacts. We define the 'score' of a user as being the overall fraction of pairs for which they correctly selected the real damage. A score of 50% is ideal, as it would mean that the user is unable to distinguish real from synthetic damage better than chance. Participant score breakdown is summarsed in Table 3.2. We find that unfamiliar participants scored on average 51.51% with standard deviation of 6.88%, moderately familiar participants scored on average 51.93% with standard deviation of 7.05%, and very familiar participants scored on average 51.43% with standard deviation of 7.0%. Thus, all groups of users found our synthetically damaged images to be indistinguishable from original damaged scans, and there was no statistically-significant variation among the groups.

Ours vs. Ivanova et al. [94]. In the second survey, participants are shown 30 pairs of film scans from the Documerica photographic collection [196]. In each pair, one version of the image is damaged using the earlier approach of Ivanova et al. [94], whereas the other image is damaged using our proposed approach. Participants were asked which image in the pair showed more realistic damage.



Figure 3.9: High resolution image, generated damage overlay and resulting synthetically damaged image.

We collected 78 responses for the second survey. In this survey, unfamiliar participants preferred our damage on average in 63.88% of the examples with standard deviation of 23.17%, moderately familiar participants preferred our damage in 73.33% of the examples with standard deviation of

Familiarity	Participants	Resolutions											
		128×128		256×256		512×512		1024×1024		2048×2048		Overall	
		Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Not Familiar	49	52.4	12.7	52.6	11.1	50.6	9.6	52.5	12.4	50.3	13.2	51.5	6.9
Somewhat Familiar	81	52.9	10.1	54.1	12.0	51.7	10.0	54.4	13.6	51.1	15.9	52.9	7.0
Very Familiar	121	54.9	11.1	53.5	11.3	50.4	12.2	49.9	12.1	48.1	16.1	51.4	7.0
Overall	251	53.8	11.2	53.5	11.5	50.8	11.0	51.9	12.7	49.6	15.5	51.9	7.0

Table 3.2: Summary of participant scores (in %) based on responses to the Real vs Synthetic Damage survey. A score of 50% is optimal since it indicates that our synthetic damage is on average indistinguishable from real damage.

21.13%, and very familiar participants preferred our damage in 75.58% of the examples with standard deviation of 21.71%. These results indicate that all groups found our synthetic damage to be more realistic than that proposed in the prior work [94]; expert participants favored our damage the most out of the three groups.

3.7 Experiments

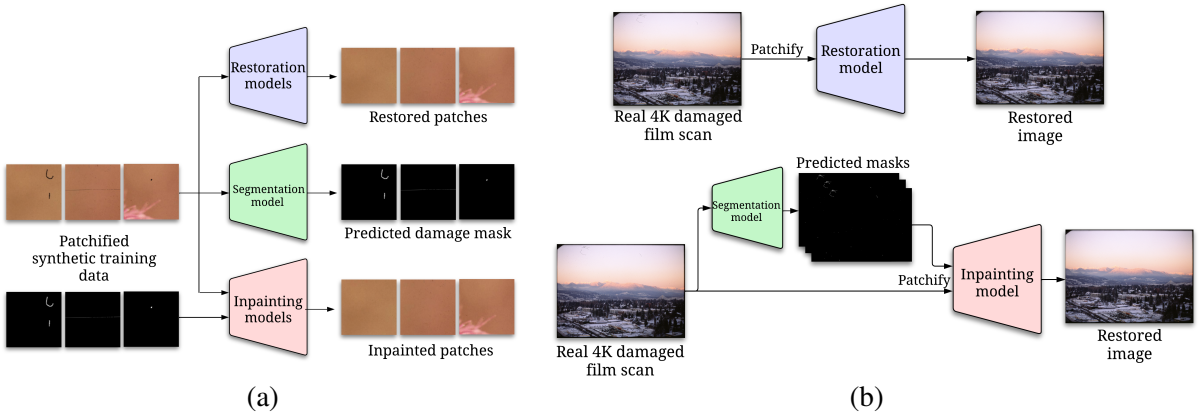


Figure 3.10: Overview of (a) training and (b) evaluation workflows for restoration, segmentation and inpainting models. We retrain existing restoration models on our synthetic data where relevant, and train an additional segmentation model to adapt state-of-the-art inpainting methods to the film artefact restoration task.

Equipped with our damage simulator (Section 3.6) and our expertly curated test set of real analogue damage with hand-restored ground-truths (Section 3.5), we perform two sets of experiments. In Section 3.7.1, we directly compare our damage synthesis pipeline with those from two prior works [89, 204]. In Section 3.7.2, we evaluate several approaches on our final goal of film restoration. The following sections discuss the chosen approaches (and how we adapt them to high resolution data), describe our experimental setup, and report the results of our experiments. The training and evaluation workflows for restoration, inpainting and end-to-end restoration are visualised in Figure 3.10.

Evaluation metrics for restoration. Since our evaluation set has ground-truth restored images (Section 3.5), we can evaluate restoration quality directly using standard image similarity metrics. Specifically, we use peak signal-to-noise ratio (PSNR), structural similarity (SSIM) [212], and the learnt perceptual metric LPIPS [237]. These are calculated over full frames (not individual patches), and we report the average over the evaluation set.

Evaluation metrics for segmentation. We obtain approximate ground-truth segmentations of the authentically damaged dataset by subtracting the damaged images from the restored ones and binarising. We use the standard image segmentation metrics intersection-over-union (IoU) and F1 score. Since these metrics are sensitive to exact pixel-value overlap, which in turn is influenced by the choice of binarisation threshold, we test several thresholds and choose the one which maximises the scores obtained by the baselines. We also borrow a popular point-cloud comparison metric, the earth-mover's distance (EMD) [1, 63], to provide an additional measure which is less sensitive to exact overlap, but more accurately compares proximity of long, narrow features such as hairs and scratches.

Synthetically damaged training data. We generate synthetic damage overlays using our proposed approach (Section 3.6) for 6228 clean 4K image scans from the Documerica collection [196] (see Figure 3.9 for an example). We render the synthetic damage either as negative or positive artefacts (corresponding to damage on film negatives or developed slides respectively), with equal probability for each image. The overlays are also binarised to produce segmentation maps indicating where damage was added. We split the set of 6228 images into training and validation sets in the ratio of 9:1. For the segmentation experiment, we use a variant of this dataset produced by damaging the same set of Documerica images with damage overlays as provided by DeepRemaster [89]. These datasets are used to retrain restoration models where relevant, as well as to train a segmentation model to detect artefact damage in our real artefact damage evaluation set. The data is available at www.doi.org/10.6084/m9.figshare.21815844.

3.7.1 Comparison against existing damage simulations

We train two U-Net [168] segmentation networks to detect artefact damage in real damaged film scans – one using damage generated by our model, and one using the damage provided by DeepRemaster [89]. We train both for 20 epochs on 256×256 crops of the corresponding damaged training set with learning rate 10^{-3} and a batch size of 16. We compare the predictions from the two segmentation networks directly to those from the artefact segmentation module of BOPB [204] as our second baseline.

Results. Segmentations obtained from the U-Net trained on our synthetically damaged data are of much higher quality than those obtained from the same model trained on the damage from DeepRemaster, as well as those obtained from the segmentation module of BOPB [204], as shown in Figure **??**. Quantitatively, the model trained on our artefacts outperforms BOPB and DeepRemaster, achieving lower EMD and higher IoU and F1 scores, see Table 3.3.

Segmentation model training	IoU ↑	F1 ↑	EMD ↓
Our damage	0.256	0.408	0.007
DeepRemaster [89] damage	0.003	0.006	0.016
BOPB [204] pre-trained module	0.010	0.019	0.016

Table 3.3: Comparison between the predicted segmentations for the evaluated models trained on different simulated damage.

3.7.2 Comparison of restoration models

We compare a diverse set of approaches to artefact restoration. First, we select three methods that directly perform damage restoration:

- *Bringing Old Photos Back to Life* (BOPB) [204], an approach specifically targeting analogue damage, for which the authors have provided pre-trained weights. We compare variants using our segmentation model vs. theirs, and with fully-convolutional vs. patch-wise processing.
- A restoration U-Net trained with perceptual loss [94], another approach which specialises in film artefact removal, for which the pre-trained weights are available. We also test a variant re-trained on crops of our synthetic 4K dataset, instead of the 256×256 downsampled images used in the original. Lastly, we test a variant re-trained on the same crops of our synthetic dataset, damaged with overlays from DeepRemaster [89].
- Adobe Photoshop's *Dust & Scratch filter*, which is a commonly-used off-the-shelf solution for artefact restoration.

Second, we select three state-of-the-art methods for image inpainting. In order to apply these to the restoration task, we also train a model for segmenting artefacts (see below), which is used to determine which pixels must be inpainted:

- *LaMa* [192], a state-of-the-art approach for high resolution irregular hole inpainting, with a pre-trained model available.
- *Stable Diffusion* [167], popular for its state-of-the-art text-to-image generation abilities, also supports image inpainting.
- *RePaint* [128], another diffusion approach focusing on inpainting, proposing a resampling technique which conditions the inpainting on a pre-trained diffusion model.

Finally, we consider two blind inpainting approaches:

- *Blind Visual Motif Removal From a Single Image* (BVMR) [87], which simultaneously detects the pixels representing the visual motif to be inpainted, and synthesises new content for the affected pixels. The provided pre-trained model is for semi-transparent emoji watermark removal, which we re-train for the artefact restoration task.

- *Restormer* [230], an approach which achieves state-of-the-art results in the tasks of blind denoising and deraining, and is applicable to large resolution inputs. Out of several models provided, we find that the closest to the task of film artefact removal is the one trained for deraining.

Pre-segmentation of artefacts for inpainting methods. LaMa, RePaint and Stable Diffusion rely on masks to be supplied for the inpainting task. As they cannot identify artefacts automatically, they cannot be applied to the film restoration task out-of-the-box. To remedy this, we employ the segmentation U-Net trained on our synthetically damaged data (described in Section 3.7.1) to predict damage masks on the test set of authentically damaged data. When evaluating LaMa, Stable Diffusion, and RePaint on damage restoration, we first pass the images through the trained segmentation U-Net, then inpaint the regions it indicates are damaged (i.e. we use the output of the segmentation model as the mask input for the inpainting model). In addition, while Wan et al.'s [204] method predicts its own damage masks, we also evaluate a variant using the masks from our segmentation model (shown in Section 3.7.1 to produce artefact segmentations of superior quality).

Processing high resolution scans. Most of the models are designed to operate on 256×256 pixel inputs. We adapt these to process the 4K image scans in our test set without downsampling them first. During inference, we process the images in our dataset in patches of size 256×256, and we stitch back the predictions using 50% overlap and a smooth blending function, as proposed by Pielawski & Wählby [158]. For the inpainting models which require masks, we increase efficiency by only processing the patches for which the corresponding mask indicates the presence of artefacts; if no artefacts have been detected in the patch, we simply copy the input patch to the output tensor to be stitched with the rest of the restored image. This does not apply to the Photoshop Dust & Scratch filter, nor to LaMa, both of which can natively process 4K images.

Method	**Notes**	**Patch-wise**	**PSNR (dB)** ↑		**SSIM** ↑		**LPIPS** ↓	
			Mean	Std	Mean	Std	Mean	Std
BOPB [204]		Yes	14.8	3.14	0.756	0.063	0.561	0.060
BOPB [204]		No	24.0	2.80	0.904	0.048	0.432	0.062
BOPB [204]	w/ our masks	Yes	14.9	3.69	0.769	0.068	0.556	0.067
BOPB [204]	w/ our masks	No	24.4	2.34	0.917	0.043	0.425	0.425
LaMa [192]	**w/ our masks**	**No**	**39.6**	**5.24**	**0.991**	**0.012**	**0.042**	**0.038**
Photoshop Dust & Scratch Filter		No	28.9	3.10	0.965	0.022	0.277	0.088
U-Net + perceptual loss [94]	trained on downsampled frames	Yes	31.8	2.62	0.979	0.015	0.224	0.047
U-Net + perceptual loss [94]	re-trained on DeepRemaster damage [89]	Yes	27.48	1.36	0.945	0.015	0.249	0.043
U-Net + perceptual loss [94]	**re-trained on our synthetic damage**	**Yes**	**33.0**	**2.46**	**0.984**	**0.011**	**0.215**	**0.044**
Stable Diffusion [167]	w/ our masks	Yes	28.6	2.43	0.959	0.020	0.305	0.040
BVMR [87]	re-trained on artefact damage	Yes	17.8	4.08	0.798	0.108	0.392	0.153
Restormer [230]	de-raining model	Yes	32.0	2.52	0.978	0.015	0.238	0.044

Table 3.4: Quantitative evaluation of restoration results on our high resolution dataset of authentic film damage.



(a) ***Input****: 4K film scan with authentic damage.*

(b) ***Artefact Segmentation****: prediction from U-Net trained on synthetic damage.*

(c) ***Segmentation by BOPB [204]****.*

(d) ***Restoration by U-Net [94]****: using original weights.*

(e) ***Restoration by BOPB [204]****: using our segmentation.*

(f) ***Restoration by BOPB [204]****: using their segmentation.*

(g) ***Restoration by U-Net [94]****: retrained on synthetic damage.*

(h) ***Restoration by LaMa [192]****: best model with our segmentation.*

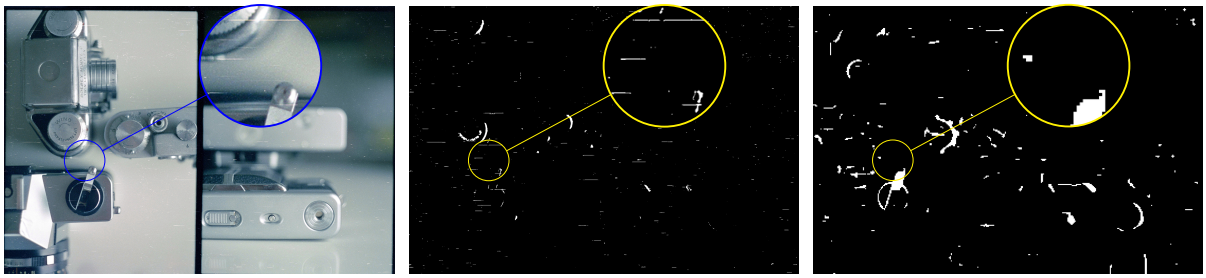
(i) ***Restoration by Stable Diffusion [167]****: using our segmentation.*

(j) ***Restoration by BVMR [89]****: retrained on synthetic damage.*

(k) ***Restoration by RePaint [128]****: using our segmentation.*

(l) ***Ground Truth****: manually restored by a human expert.*

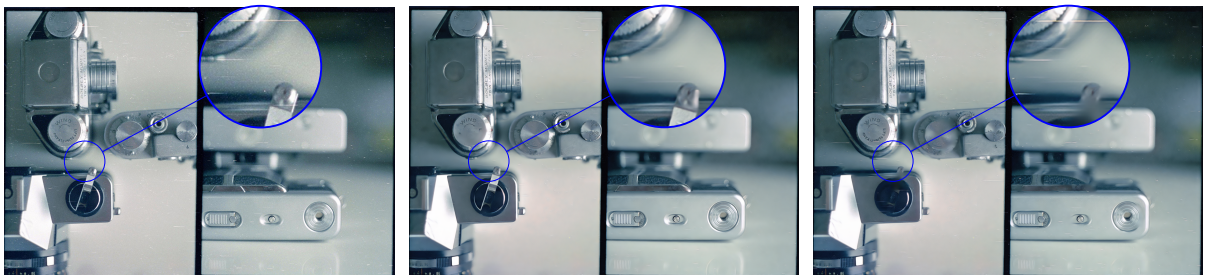
Figure 3.12: Input and ground truth from our authentic artefact damage dataset, along with chosen restorations.



(a) ***Input****: 4K film scan with authentic damage.*

(b) ***Artefact Segmentation****: prediction from U-Net trained on synthetic damage.*

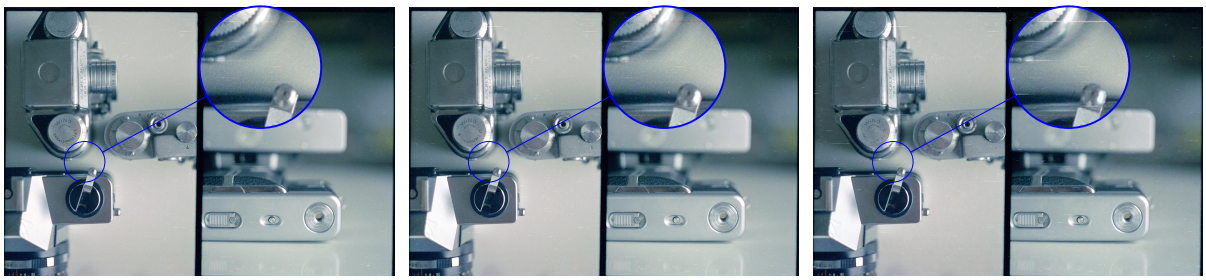
(c) ***Segmentation by BOPB [204]****.*



(d) ***Restoration by U-Net [94]****: using original weights.*

(e) ***Restoration by BOPB [204]****: using our segmentation.*

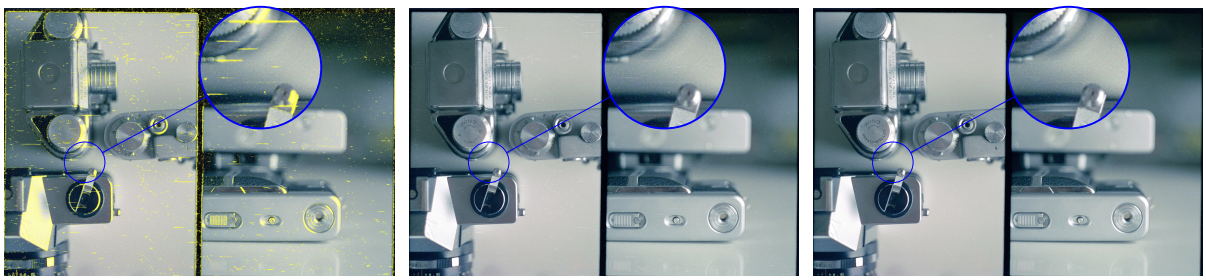
(f) ***Restoration by BOPB [204]****: using their segmentation.*



(g) ***Restoration by U-Net [94]****: retrained on synthetic damage.*

(h) ***Restoration by LaMa [192]****: best model with our segmentation.*

(i) ***Restoration by Stable Diffusion [167]****: using our segmentation.*



(j) ***Restoration by BVMR [94]****: retrained on synthetic damage.*

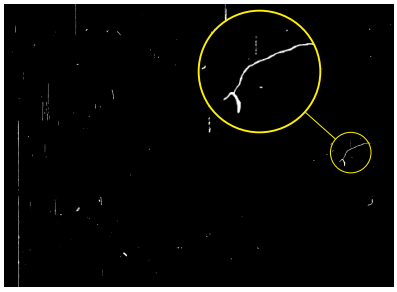
(k) ***Restoration by RePaint [128]****: using our segmentation.*

(l) ***Ground Truth****: manually restored by a human expert.*

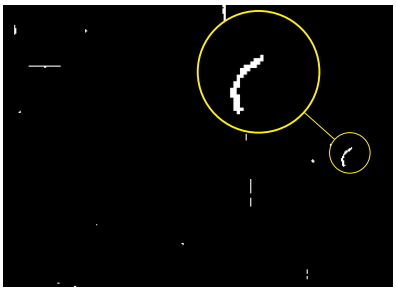
Figure 3.12: Input and ground truth from our authentic artefact damage dataset, along with chosen restorations.



(a) ***Input****: 4K film scan with authentic damage.*



(b) ***Artefact Segmentation****: prediction from U-Net trained on synthetic damage.*



(c) ***Segmentation by BOPB [204]****.*



(d) ***Restoration by U-Net [94]****: using original weights.*



(e) ***Restoration by BOPB [204]****: using our segmentation.*



(f) ***Restoration by BOPB [204]****: using their segmentation.*



(g) ***Restoration by U-Net [94]****: retrained on synthetic damage.*



(h) ***Restoration by LaMa [192]****: best model with our segmentation.*



(i) ***Restoration by Stable Diffusion [167]****: using our segmentation.*



(j) ***Restoration by BVMR [94]****: retrained on synthetic damage.*



(k) ***Restoration by RePaint [128]****: using our segmentation.*



(l) ***Ground Truth****: manually restored by a human expert.*

Figure 3.12: Input and ground truth from our authentic artefact damage dataset, along with chosen restorations.

3.7.3 Restoration results and discussion

Table 3.4 summarises our quantitative results, showing PSNR, SSIM and LPIPS metrics with respect to the ground-truth professionally-restored image set. All three metrics largely agree, with LaMa and the U-Net retrained with our damage as the best and second best performing models respectively. We can directly observe the substantial improvement that re-training the simple U-Net restoration model [94] with our simulated damage brings to its restoration performance. On the other hand, re-training the same model with the damage from DeepRemaster [89] results in deteriorated restoration performance.

BOPB works especially poorly in patch-wise mode as it introduces color distribution shift in each patch, leading to a very distorted result and correspondingly low scores. We attribute this to the fact that BOPB aims to remove film grain and modify color to match modern digital photographs, in addition to removing artefacts. This means it performs poorly at our specific task of damage removal (and is thus unsuitable for applications where the qualities of analogue photographs are desirable). Interestingly, the standard "Dust and Scratches filter" in Photoshop exceeds the performance of BOPB (patchwise and fully-convolutional modes), Stable Diffusion and BVMR. Restormer and the best U-Net model have comparable performance, but LaMa exceeds the performance of all of the competitive models on all three metrics, particularly as measured by LPIPS. This is in spite of LaMa being trained with inpainting masks that are significantly different in size and shape to typical analogue damage artefacts. Note that we do not include results for RePaint in Table 3.4. This model requires over 12 minutes for the restoration of a single 256×256 pixel patch on a Tesla P100 GPU. Given that one 4K image from our test set is split into around 200 overlapping patches, about 40 hours are needed to process just one image. While this is clearly impractical in a real-life restoration scenario, we process a subset of 10 images from our test set for qualitative evaluation.

We present qualitative results in Figure 3.12, with additional examples in the supplementary material, Figure 3.13. BOPB performs significantly better when we substitute its predicted artefact segmentation masks (which are strongly under-segmented) with ours. However, it still modifies colors and smoothes the film grain, significantly changing the character of images. Even when guided by our segmentation masks, Stable Diffusion fails to inpaint the damage. Similarly, BMVR, even when re-trained using our synthetic damage, struggles with the same issue as BOPB and injects additional damage. In general, we find that methods based on inpainting which do not modify regions away from the artefacts perform much better than BOPB, as they are less likely to introduce new damage or distortions. In line with our observations from the quantitative evaluation, retraining the U-Net model [94] on our improved synthetic damage results in higher quality restorations. Visually, LaMa performs slightly better than [94]; however both methods fail to inpaint part of the upper hair in Figure 3.1. RePaint's performance, guided by our segmentation masks, is comparable to that of LaMa and [94]. Despite producing restorations of relatively high visual quality, the model's runtime limits its practical usefulness.

3.8 Conclusions

Automated restoration of mechanical damage in scanned film is a challenging task. Fully automated restoration would be transformative in improving the quality of imaging in the enormous archives of extant film held around the world. However, conventional inpainting and restoration processes are unsuitable for professional-quality restoration. All of the machine learning approaches we tested for film restoration perform well below the level required to be competitive with professional hand restoration. Models trained on low-resolution patches performed very poorly when applied patch-wise, but even state-of-the-art models like LaMa applied at native resolution were unable to adequately inpaint artefacts without introducing undue distortion. Many of the algorithms would also be computationally impractical for processing large image collections at high resolution.

Progress towards film restoration that can operate truly automatically at realistic scan resolutions requires better models of the film damage process. To that end, we have presented both a sophisticated statistical model for synthesising large quantities of realistic mechanical artefacts. Our extensive human validation of the synthetic artefact model suggests that even experts cannot reliably distinguish between our synthetic damage models and real damage at any level of zoom. Furthermore, we have demonstrated that training with damage generated by our model leads to a significant improvement in both artefact detection and end-to-end restoration tasks when tested on authentic damaged images. We have also published a high-quality baseline dataset to drive progress in restoration research.

Overall, we conclude that there remains significant work to achieve acceptable automatic restoration quality. We see the use of high-quality damage simulators, multi-scale approaches that can incorporate wide image context while operating at high resolution, and evaluation with challenging full-resolution image benchmarks as important directions to achieve these goals.

3.9 Afterword

This chapter demonstrated that our statistically-grounded approach to synthetic damage generation produces training data that closely aligns with real film artifacts. Through two perceptual studies, the simulator was shown to produce synthetic damage which was not only superior to the previously used naive damage overlays, but was also indistinguishable from authentic damage even by experts. This enabled significant improvements in supervised damage detection, as verified through extensive empirical evaluation on real film scans with expert-annotated ground truth. These experiments established the performance bounds achievable through supervised learning when synthetic training data is optimally aligned with the target domain.

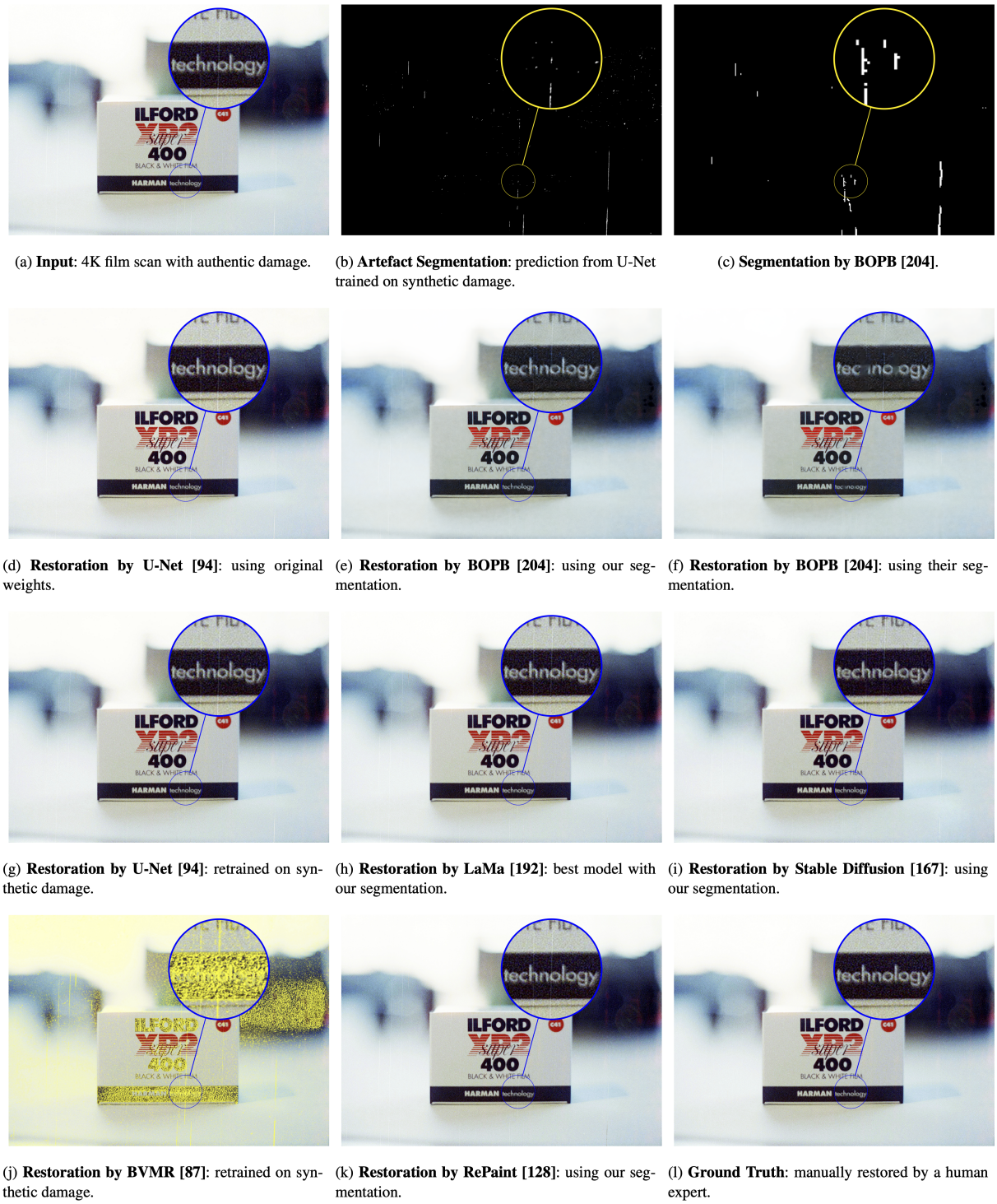
However, fundamental questions about model understanding and generalisation remain. Even with strong performance on film damage detection, would these models generalise across different media types? While scratches on film negatives share conceptual characteristics with cracks in paintings, folds in photographs, or tears in ancient textile—patterns easily recognised by humans despite their visual differences—the question of cross-domain transfer remained unexplored.

Furthermore, error analysis revealed another critical aspect of generalisation: the evaluated models' ability to disambiguate between actual damage and visually similar semantic content. For instance, distinguishing between film scratches and visually similar structures like cat whiskers or technical drawing lines presents a challenging test of whether models have learned meaningful damage representations beyond simple low-level pattern matching.

Both the cross-media generalisation and semantic disambiguation challenges point to a fundamental question in representation learning: even with optimal training data, are models learning robust damage representations or merely memorising media domain-specific visual patterns?

These open questions motivated expanding the investigation beyond film artefacts. The subsequent work, presented in the next chapter, systematically examines both material-based generalisation, i.e. testing damage recognition across diverse media types, and content-based generalisation, i.e. evaluating semantic disambiguation across varied image contents.

3.10 Supplementary



(a) **Input**: 4K film scan with authentic damage.

(b) **Artefact Segmentation**: prediction from U-Net trained on synthetic damage.

(c) **Segmentation by BOPB [204]**.

(d) **Restoration by U-Net [94]**: using original weights.

(e) **Restoration by BOPB [204]**: using our segmentation.

(f) **Restoration by BOPB [204]**: using their segmentation.

(g) **Restoration by U-Net [94]**: retrained on synthetic damage.

(h) **Restoration by LaMa [192]**: best model with our segmentation.

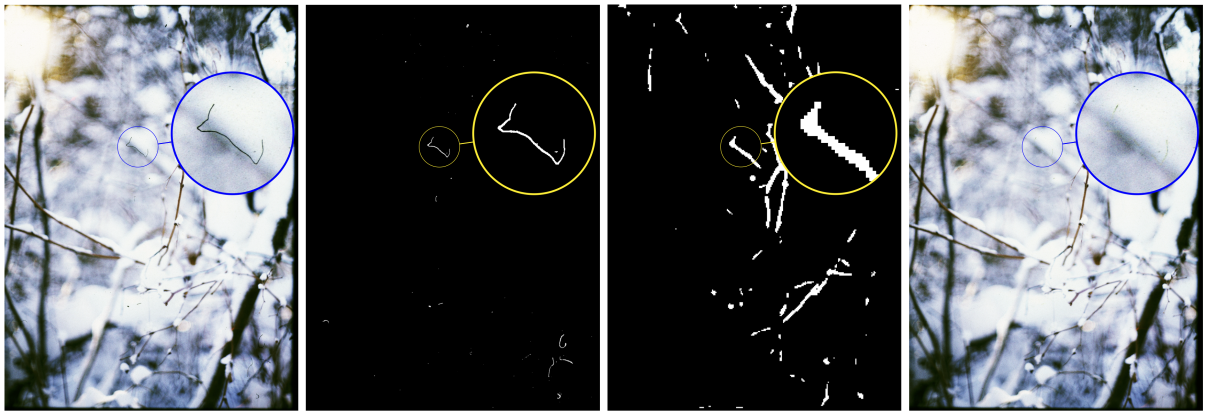
(i) **Restoration by Stable Diffusion [167]**: using our segmentation.

(j) **Restoration by BVMR [87]**: retrained on synthetic damage.

(k) **Restoration by RePaint [128]**: using our segmentation.

(l) **Ground Truth**: manually restored by a human expert.

Figure 3.13: Input and ground truth from our authentic artefact damage dataset, along with chosen restorations.



(a) **Input**: 4K film scan with authentic damage.

(b) **Artefact Segmentation**: prediction from U-Net trained on synthetic damage.

(c) **Segmentation by BOPB [204]**.

(d) **Restoration by LaMa [192]**: best performing model, using our segmentation.

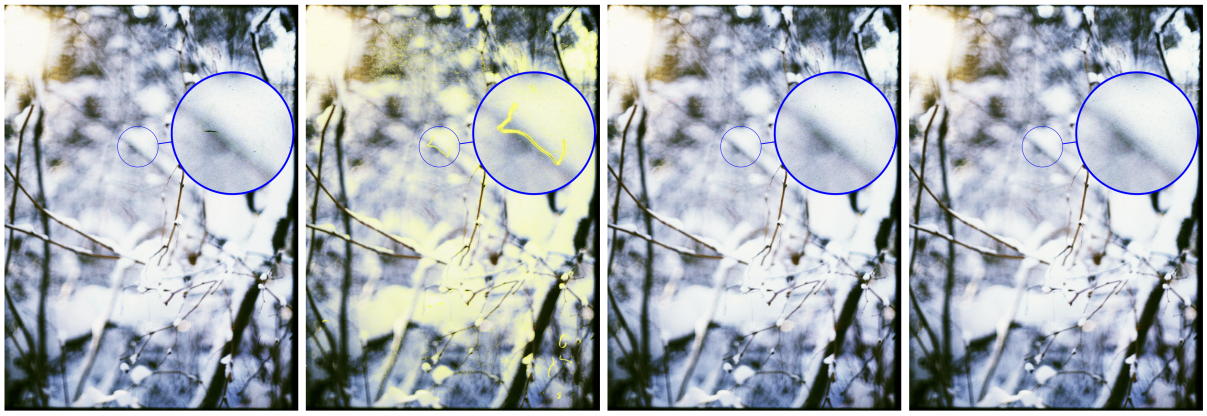


(e) **Restoration by U-Net [94]**: using original weights.

(f) **Restoration by BOPB [204]**: using our segmentation.

(g) **Restoration by BOPB [204]**: using their segmentation.

(h) **Restoration by Stable Diffusion [167]**: using our segmentation.



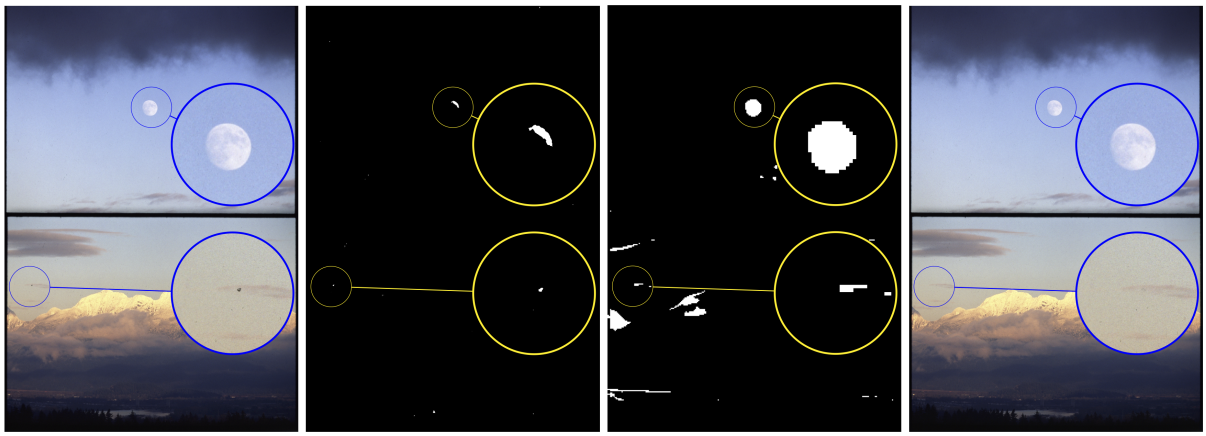
(i) **Restoration by U-Net [94]**: retrained on synthetic damage.

(j) **Restoration by BVMR [87]**: retrained on synthetic damage.

(k) **Restoration by RePaint [128]**: using our segmentation.

(l) **Ground Truth**: manually restored by a human expert.

Figure 3.13: Input and ground truth from our authentic artefact damage dataset, along with chosen restorations.

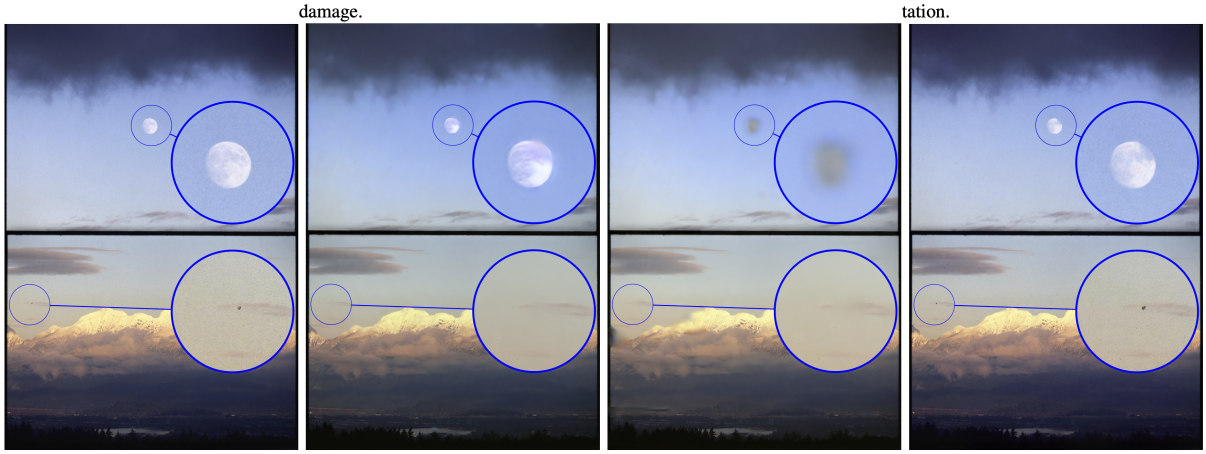


(a) **Input**: 4K film scan with authentic damage.

(b) **Artefact Segmentation**: prediction from U-Net trained on synthetic damage.

(c) **Segmentation by BOPB [204]**.

(d) **Restoration by LaMa [192]**: best performing model, using our segmentation.

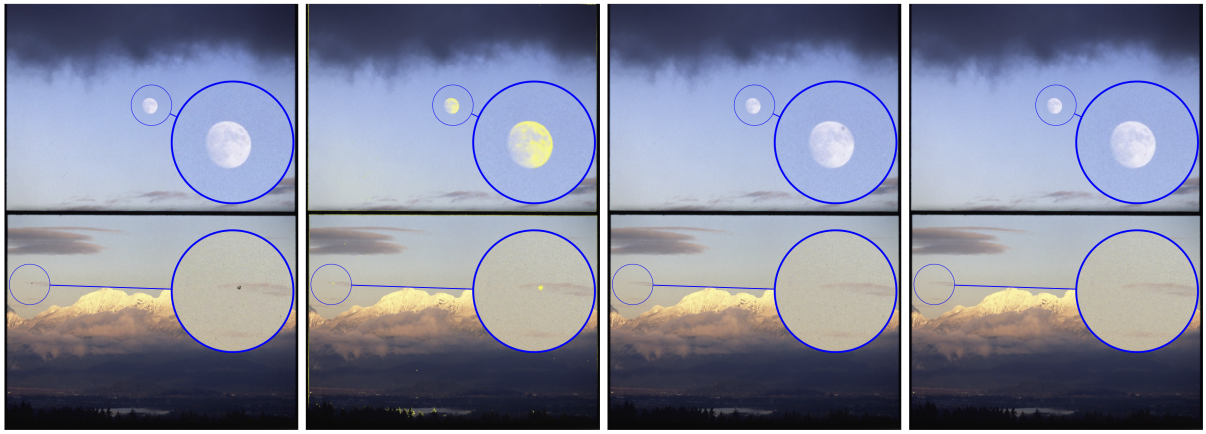


(e) **Restoration by U-Net [94]**: using original weights.

(f) **Restoration by BOPB [204]**: using our segmentation.

(g) **Restoration by BOPB [204]**: using their segmentation.

(h) **Restoration by Stable Diffusion [167]**: using our segmentation.



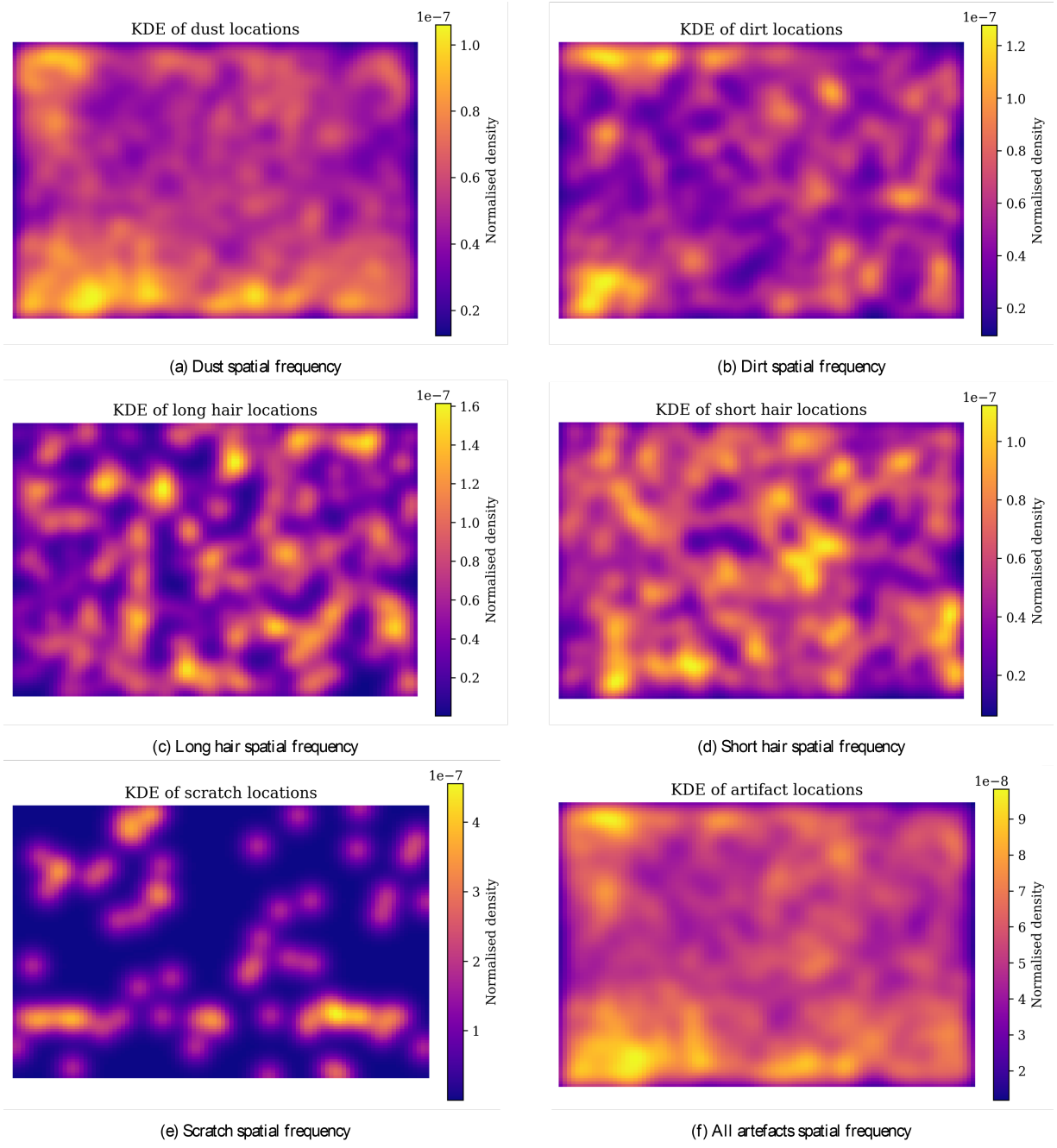
(i) **Restoration by U-Net [94]**: retrained on synthetic damage.

(j) **Restoration by BVMR [87]**: retrained on synthetic damage.

(k) **Restoration by RePaint [128]**: using our segmentation.

(l) **Ground Truth**: manually restored by a human expert.

Figure 3.13: Input and ground truth from our authentic artefact damage dataset, along with chosen restorations.



(a) Dust spatial frequency

(b) Dirt spatial frequency

(c) Long hair spatial frequency

(d) Short hair spatial frequency

(e) Scratch spatial frequency

(f) All artefacts spatial frequency

Figure 3.13: Spatial frequencies aggregated across all 10 scans, broken down by artefact type ((a), (b), (c), (d), (e)) and for all artefact types ((f)).

Chapter 4

State-of-the-Art Fails in the Art of Damage Detection

The majority of this chapter (Section 4.2 through to Section 4.7) reproduces the author's version of the accepted, peer review manuscript presented as a full paper at the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) as *Ivanova, D., Aversa, M., Henderson, P. and Williamson, J. (2024) ARTeFACT: Benchmarking Segmentation Models on Diverse Analogue Media Damage. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2024), Tucson, AZ, USA, 28 Feb - 4 Mar 2025,* and does not violate the copyright of the publisher. A short paper version of this chapter was published at the ECCV VISART workshop as *Ivanova, D., Aversa, M., Henderson, P. and Williamson, J. (2024) State-of-the-Art Fails in the Art of Damage Detection. In: Vision for Art VISART VII Workshop at the European Conference of Computer Vision (ECCV) 2024, Milan, Italy, 30 Sep 2024.* The majority of the work presented below was undertaken by myself. I collected and manually annotated all images and performed statistical analysis of the data. Manual verification of the annotations was performed by Sebastían Aírton. I formulated and carried out all experiments involving the fine-tuning and benchmarking of segmentation models, with the exception of evaluating GroundingSAM, which was carried out by Marco Aversa. The idea behing the LOOCV experimental setting was concieved by Dr John Williamson and Dr Paul Henderson and executed by me. All figures and visualisations were created by me. I drafted and rewrote the paper, with my supervisors Dr John Williamson and Dr Paul Henderson providing edits and corrections.

4.1 Summary

The success of our synthetic data-driven approach in film damage detection demonstrated in the previous chapter raised fundamental questions about how well supervised methods could generalise across different types of analogue media and damage, and how well they could simultaneously distinguish damage from visually similar image content, especially in the context of

artwork, where depictions are highly stylised. These questions also highlighted a practical gap in the field—while some damage datasets existed, they were typically limited to a single type of media or sourced from a single collection. No comprehensive dataset existed for studying how damage manifests across different materials and content types.

This chapter introduces ARTeFACT, a comprehensive dataset of varied damage across diverse analogue media. Comprising over 11,000 pixel-accurate annotations spanning 15 damage categories across a number of materials (paintings, photographs, textiles, mosaics, frescoes), along with semantic descriptions of the contents of each image, ARTeFACT provides unprecedented scope for analyzing damage characteristics and benchmarking detection approaches. The dataset's categorisation by material properties and content types facilitates investigation of both physical damage's manifestation across various media, and the impact of visual complexity on detection performance.

Extensive evaluation of state-of-the-art supervised segmentation approaches on this dataset in a leave-one-out cross-validation setting confirmed the shortcomings of supervised learning in the analogue damage detection task.

Experiments with Segment Anything (SAM)—a powerful foundation model for segmentation—proved particularly illustrative: when presented with an artwork containing multiple cracks, SAM treats each crack as a separate visual entity requiring its own prompt, rather than understanding "cracks" as a semantic category that can encompass multiple (sometimes hundreds of) disjoint instances within the same image. On one hand, this leads to the need of an impractically large number of prompts to get visually satisfactory segmentations, and on the other, it demonstrated a lack of semantic discriminability by the model, which is unable to produce interpretable labels to group together segmented damaged regions and distinguish them from non-damaged ones. Fine-tuning and evaluating SAM in the same fashion as the rest of the benchmarked supervised segmentation models to address this, caused SAM to fail in similar ways to the rest: poor cross-domain generalisation and inability to disambiguate damage from visually similar semantic content.

The above observations motivated our exploration of text-to-image (T2I) diffusion models like Stable Diffusion. These models, trained on vast datasets of image-text pairs, have demonstrated a remarkable understanding of diverse visual content. Beyond generating natural images with incredibly realistic appearance, they can create images in specific artistic styles—from oil paintings to ancient frescoes—while accurately reproducing the characteristic artifacts and material properties of each medium. This suggests their internal representations capture both semantic content and physical characteristics across different media types. Most intriguingly, these models appear to have developed an implicit understanding of how damage manifests in different contexts. To what extent damage could be visually segmented and conceptually disentangled from overall style representations, however, remained to be answered. To this end, we evaluated two complementary approaches to leveraging Stable Diffusion: with DiffSeg, we

explore whether the model's self-attention features inherently encode meaningful visual segmentation of damage without supervision, and through DiffEdit, we test whether text guidance can leverage multimodal features to identify damage. While both approaches revealed limitations in controlling granularity and disambiguating damage from artistic elements, they highlighted the rich semantic information in diffusion model representations and pointed toward the potential for a more principled unsupervised approach.



Figure 4.1: Examples from our dataset of damaged artwork, categorised by Material (rows 1 and 2) and Content (row 3). Annotation colours correspond to different types of damage. Note the diversity of media and content, and pixel-accurate damage masks.

4.2 Abstract

Accurately detecting and classifying damage in analogue media such as paintings, photographs, textiles, mosaics, and frescoes is essential for cultural heritage preservation. While machine learning models excel in correcting degradation if the damage operator is known a priori, we show that they fail to robustly predict *where* the damage is even after supervised training; thus, reliable damage detection remains a challenge. Motivated by this, we introduce ARTeFACT, a dataset for damage detection in diverse types analogue media, with over 11,000 annotations covering 15 kinds of damage across various subjects, media, and historical provenance. Furthermore, we contribute human-verified text prompts describing the semantic contents of the images, and derive additional textual descriptions of the annotated damage. We evaluate CNN, Transformer, diffusion-based segmentation models, and foundation vision models in zero-shot, supervised, unsupervised and text-guided settings, revealing their limitations in generalising across media types. Our dataset is available at https://daniela997.github.io/ARTeFACT/ as the first-of-its-kind benchmark for analogue media damage detection and restoration.

4.3 Introduction

Artworks on analogue media are at risk of deterioration over time due to environmental factors, human intervention, or simple aging. Accurately distinguishing and analysing such damage is a crucial step in the preservation process, with many practical applications—including improved archiving and curation, provenance research, and downstream restoration via digital tools. Manual identification of damage is labour-intensive, requiring specialized software, significant financial investment, and extensive time commitment from skilled specialists [34, 29, 92]. These constraints limit the scale and frequency of restoration efforts.

Whether machine learning can successfully be used to automate the damage detection has not yet been established. This can be attributed to the constraints posed by data availability: damage in analogue media is rarely detailed in metadata, making data collection challenging. Selecting and annotating examples which cover the varied manifestations of damage across different materials and contents in analogue media is difficult but essential for assessing a model's ability to generalise at such a complex task [32]. Previous work has only targeted one type of analogue media at a time and sourced all of their data from the same place [29, 125, 227, 137]. Consequently, their results do not reflect the evaluated approaches' performance on unseen data, and whether the models are able to truly learn what damage *is*. Effective evaluation requires a diverse dataset and a protocol that accounts for the manifold of analogue media types and their differences.

We propose the ARTeFACT dataset, which is designed to enable a comprehensive evaluation of current and future approaches. Our contributions are as follows:

- We produce an extensive dataset for the real-life task of damage detection in analogue media. For each image, we provide a pixel-accurate damage mask, with over 11,000 annotations in total, across 418 high-resolution images spanning various cultures and historical periods. This dataset is the first of its kind.
- We propose and justify a comprehensive taxonomy of analogue damage, categorizing deterioration into 15 distinct classes, shown in Figure 4.2. We classify the images into 10 material categoriesv (see Figure 4.1, rows 1 and 2) to represent how damage manifests across different media based on material properties. Additionally, we use 4 content categories (see Figure 4.1, row 3) to capture differences in content complexity and structure.
- We perform a thorough evaluation of state-of-the-art segmentation models in various settings - including a vision foundation model trained specifically for segmentation - and demonstrate that all models fall short of generalising across different analogue media domains and damage types.
- We further benchmark two state-of-the-art diffusion-based segmentation methods. Our results suggest that current text-to-image models have insufficient specificity in conditioning for damage, resulting in the segmentation of irrelevant areas.

4.4 Related work

Damage in analogue media. Advances in digital technology have revolutionized access to cultural artifacts by enabling digitization. Digitization creates durable replicas, but also captures *damage* that was present on the original media. The field of cultural heritage preservation has faced ongoing challenges in defining what *is* and *isn't* damage, and hence - what should and should not be restored [189], complicating conservators' decisions on restoration scope and methods[152]. By detecting and restoring damage digitally, we can address and mitigate the risk of causing further harm to the original analogue medium; this aligns with the crucial conservation principle of reversibility [15, 152]. However, this also necessitates skilled use of specialised software by restoration professionals [34, 33, 32]. Damage in analogue media, which includes a range of materials such as film, paintings, mosaics, murals, drawings, textile and others, can manifest in diverse forms. Chambah [34] categorizes such damage broadly into two groups: chemical degradations and mechanical degradations. While this classification found in Chambah's work pertains to film emulsion [33, 32], similar ontologies are commonly applied in the field of mathematical modelling of painting damage [56, 155, 29, 55].

Image restoration and segmentation. The task of image restoration typically involves reversing a degradation process H applied to the undamaged image $I = f(x,y)$ to produce the observed damaged image I'. Common restoration tasks include denoising [114, 228, 230, 120, 38], superresolution [99, 206, 120], and colourisation [236, 91, 111, 174]. Training data is often generated by applying transformations to undamaged images, like creating grayscale versions for colourisation. These tasks are primarily digital and easy to simulate, unlike analogue media damage, which stems from the physical properties of the medium and is only digitised after photography or scanning. The irregular nature of such damage—material loss, cracks, scratches—presents significant challenges for digital simulation [32, 204, 140, 94, 92, 29, 176, 50, 55].

In most restoration tasks, such as superresolution and colourisation, the degradation operator is global; when damage is localized (e.g., inpainting), the damaged areas are typically assumed to be known. Inpainting has been addressed using VAEs [80, 156], GANs [224], and diffusion models [174, 167, 128, 221, 210]. Stable Diffusion [167] has demonstrated state-of-the-art inpainting capabilities and is now the go-to tool for conditional image inpainting. However, these models rely on masks indicating the damaged areas. This suggests that research should shift focus from inpainting to *detecting* damage, the most challenging and understudied part of the machine learning pipeline.

Blind inpainting, where the mask indicating the damaged area is not provided, typically requires a segmentation module to detect these areas, using models like autoencoders [209, 28, 87] or U-Nets [140, 92]. State-of-the-art architectures like UPerNet [217] and SegFormer [218] could be trained for damage segmentation if annotations are available. Self-supervised models such as DINO [31], DINOv2 [153], SAM [108], and CLIP [160] have also shown promise in learning semantically rich representations useful for segmentation tasks. However, their ability

to generalise to the complex task of damage segmentation, with its variability in shape and distribution, has remained an open question, which we address in this work.

Recent text-to-image diffusion models like Stable Diffusion [167] have inspired methods such as DiffEdit [51], which uses pre-trained diffusion models to detect damaged areas based on text prompts, followed by inpainting. DiffSeg [197], the state-of-the-art for unsupervised segmentation, leverages diffusion models for semantic segmentation without additional supervision.

Datasets. Several art datasets have been developed for classification [53, 175] and semantic segmentation [45] in cultural heritage. More recent datasets pair artwork images with textual descriptions [47, 22], and even leverage text conditioning to synthesise artwork variants via diffusion models [43]. However, there is a notable scarcity of datasets of damaged analogue media, and the few that do exist tend to focus exclusively on one type of media: paintings [50, 176], illuminated manuscript miniatures [29], film emulsion scans [92], frescoes [137].

4.5 Damaged Analogue Media Dataset

We present a new dataset for damage detection in analogue media. Our dataset consists of 418 high-resolution images of various kinds of analogue media, including examples of: manuscript miniatures, photographs, maps, ephemera (such as posters and leaflets), mosaics, drawings and sketches, paintings, frescoes, carpets, tiles, lithographs, book covers, technical drawings and blueprints, wallpapers, letters, tapestries and stained glass (indicated as a `Type` attribute in the dataset's metadata). We provide over 11,000 pixel-level annotations covering 15 types of damage, and categorise them based on material and content. In addition to these annotations, we also make available natural language annotations, disentangled into content and damage descriptions. Images were manually selected from digitized collections of museums and galleries, with additional sources from WikiArt and Flickr groups, all collected under a CC license at the highest possible resolution.

4.5.1 Annotation methodology

All images were annotated with 15 physical damage types by one expert annotator, followed by two rounds of manual review to ensure quality. A second expert conducted a final review, resulting in over 11,000 annotated instances. Our expert annotators specialise in digitally restoring analogue media and have experience with all media types in the dataset. Each image is also classified by material and content based on its metadata, yielding 10 material classes and 4 content classes. For a detailed look at our images and image-level annotations through the prism of CLIP, refer to Section 4.9.2 in the Supplementary.

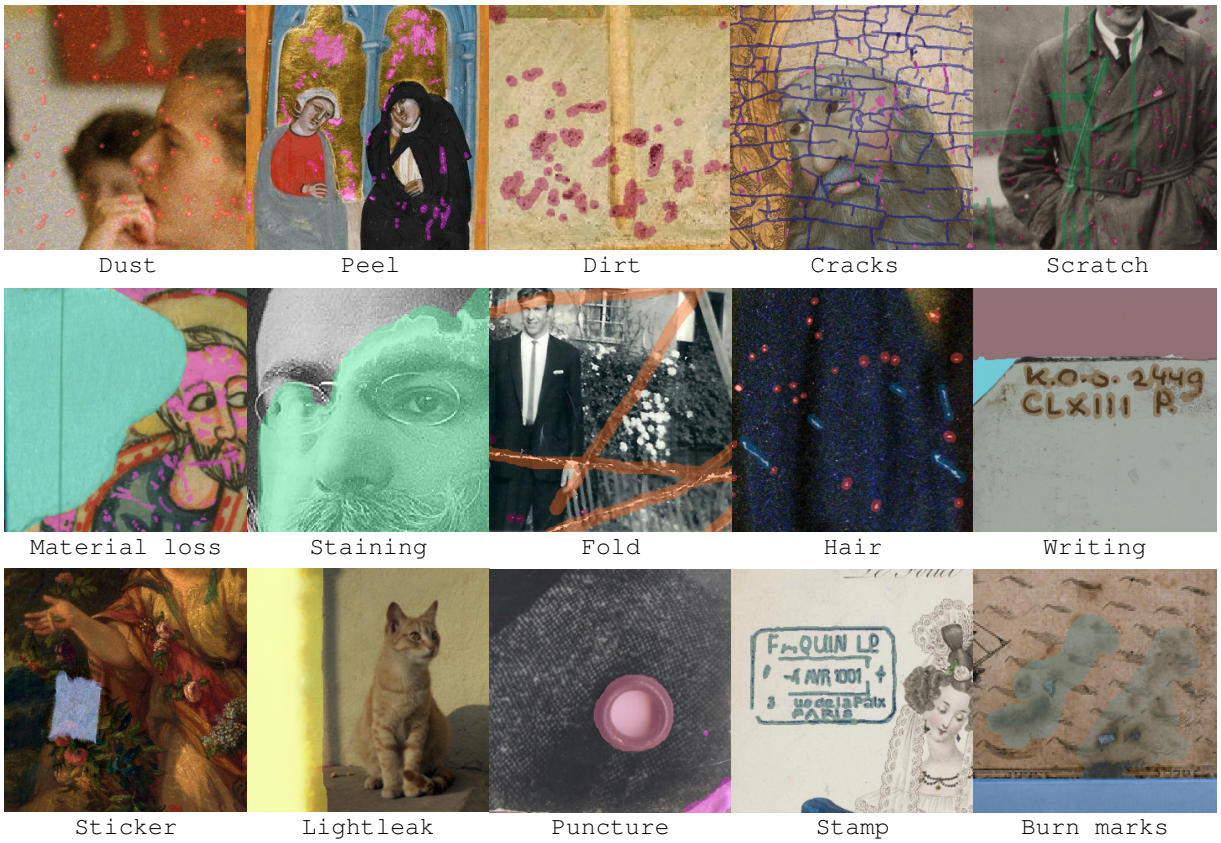


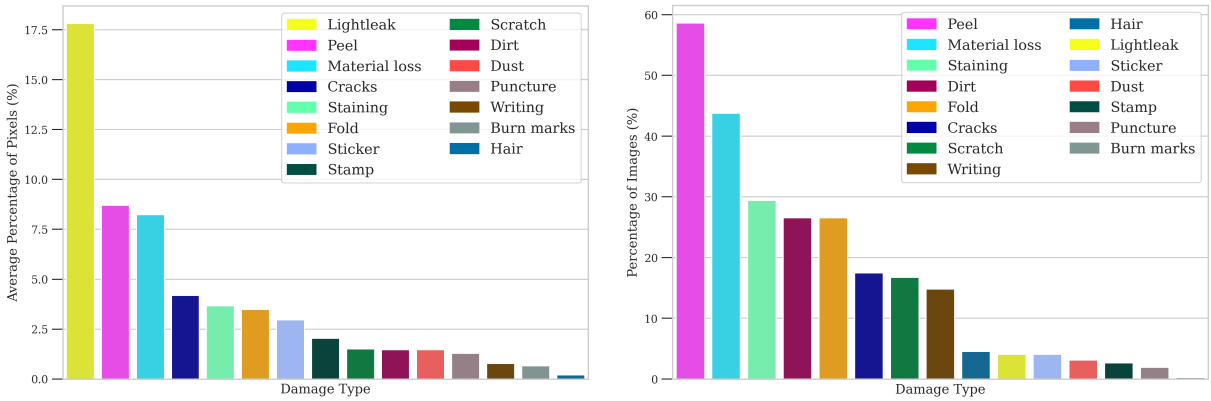
Figure 4.2: Damage types found in our dataset, demonstrating the variety of shape, scale and severity of damage.

4.5.2 Damage Types

We define damage types by their properties (scale, opacity), causes (mechanical, chemical), and effects (occluding information, absence of information, structural deformity). Each pixel-level annotation is assigned a damage type from the proposed taxonomy, with `Clean` denoting parts of the artwork which are undamaged, and remaining pixels classified as `Background`, covering the framing surface of digitisation, if visible (usually in the cases of non-rectangular artwork). This taxonomy was developed in consultation with multiple heritage preservation experts.

The types of damage include `Material loss`, which refers to missing or eroded sections of the artwork, and `Peel`, indicating areas where layers (e.g. of pigment) have separated or are flaking off. `Dust`, `Hair`, and `Dirt` denote different kinds of surface contaminants. `Scratch`, `Puncture`, `Fold` and `Cracks` capture mechanical damage and deformities, potentially caused by mishandling or accidents, or due to age. `Stamp` and `Sticker` highlight intentional or accidental markings that could be seen as defacements. `Staining` and `Burn marks` are indicative of chemical or thermal damage. `Lightleak` is a specific type of damage often found in photographic materials, caused by unintended exposure to light.

Damage types vary widely in terms of shape, size and spatial distribution; Figure 4.3 shows the distribution and prevalence of each type of damage within the dataset. Figure 4.3a considers



(a) Severity of Damage Types in Affected Images (b) Prevalence of Damage Types in the Dataset

Figure 4.3: Overview of the prevalence and severity of different damage types in the Dataset.

the extent of damage as a percentage of pixels covered. In Figure 4.3b, each distinct type of damage present in an image is instead counted as a single instance, regardless of its extent or interconnectedness. We can see that Peel and Material Loss are the most common types of damage found in the dataset, whereas Lightleaks are less common. At the same time, Lightleaks are very severe (i.e. cover many pixels) when they occur. Dirt, Fold, and Staining are common but vary in severity. Further insight can be gained by categorising the images in the dataset by material and content.

4.5.3 Image-level Annotations

In addition to the pixel-level damage annotations, we also provide three types of image-level annotations.

Material categorisation. How damage manifests itself is correlated to the materials used in the imaging process; each material type has unique properties and susceptibilities to different forms of damage. We propose the following categories, based on information sourced from the images' metadata during the collection process: *Parchment*, *Film emulsion*, *Glass*, *Paper*, *Tesserae*, *Canvas*, *Lime plaster*, *Textile*, *Ceramic*, *Wood*; examples for each category are shown in Figure 4.1, rows 1 and 2.

Content categorisation. Besides classifying based on material, we also provide content categorisation to capture the differences of overall image attributes across analogue media types, which can vary in complexity and style—for instance, photographs capture natural scenes, while paintings are stylised depictions, tile and carpet designs usually consist of repetitive geometric or abstract patterns. The following categories are proposed based on the observed differences in types content in each type of imagery: *Artistic depiction*, *Photographic depiction*, *Line art*, *Geometric patterns*; examples for each category in Figure 4.1, row 3.

Textual descriptions. We provide textual descriptions for each image, detailing both content and damage types. The expert descriptions are derived by correcting draft captions produced by

LLaVA [124]. Additional damage descriptions were compiled from our pixel-level annotations. For detailed examples and discussion on how much the expert captions improve those of LLaVA refer to Section 4.9.2 in the Supplementary.

4.6 Benchmark

We first benchmark Segment Anything, as it is state-of-the-art for many segmentation tasks. We evaluate the model in six different prompting settings at the task of segmenting damage. Next, we evaluate several state-of-the-art semantic segmentation methods on our dataset, and compare them alongside linear probes of SAM and another state-of-the-art vision foundation model - DINOv2. For rigorous evaluation in this supervised setting, we define a LOOCV protocol that measures how well models generalise to unseen media materials and contents, where not only the damaged, but the undamaged regions exhibit variance as a result of material and content properties. We further evaluate two diffusion segmentation methods based on Stable Diffusion, demonstrating the dataset's versatility and underscoring the task's complexity. We analyze each method's effectiveness across Material and Content types within the dataset, finding that although some are capable of detecting damage within specific categories or through tailored prompting, no single method excels across the whole dataset, and all methods have different shortcomings in addressing our task.

4.6.1 Zero-shot Segmentation with SAM

In its intended zero-shot setting, SAM requires prompting which can be done in several ways (points, bounding boxes, masks, or text, though the latter is not publicly available), and which would not be available in a real-life damage restoration scenario. Furthermore, SAM is inherently not a semantic segmentation model; it is designed to segment individual visual entities without assigning semantic labels to them or grouping them, meaning that each damage instance must first be segmented individually and then manually assigned the correct label.

Evaluation protocol We address these complexities and extensively benchmark SAM using an oracle to assign the correct label (Damaged or Clean) post-segmentation. We evaluate the model's ability to segment damage by deriving the required prompts from our ground-truth labels in several settings: point per segmentation label, point per segmentation instance (connected component), bounding box per segmentation label, and bounding box per segmentation instance. Additionally, we also evaluate prompting via a grid of points, which is the authors' original solution for cases where a prompt is unavailable, and prompting via text by employing Grounding SAM [164], where the model is prompted with textual description of the damage types present in each image. We report macro-averaged metrics across the entire dataset as well as for each Material and Content split in Table 4.1.

Table 4.1: Results from benchmarking SAM at zero-shot damage segmentation across various prompt modes. N corresponds to average number of prompts (i.e. points, bounding boxes, text) required as input. Point grid was evaluated with 64 points, sampled uniformly [108].

Test Class	Point per class				BBox per class				Point per instance				BBox per instance				Point Grid				Grounding SAM			
	Acc	F1	mIoU	N	Acc	F1	mIoU	N	Acc	F1	mIoU	N	Acc	F1	mIoU	N	Acc	F1	mIoU	N	Acc	F1	mIoU	N
Wood	0.85	0.15	0.10	2	0.86	0.15	0.11	2	0.88	0.41	0.31	206	0.90	**0.60**	0.49	206	0.87	0.41	0.32	-	0.83	0.05	0.03	2
Ceramic	0.85	0.39	0.27	4	0.79	0.09	0.06	4	0.85	0.50	0.35	185	0.89	**0.57**	0.44	185	0.86	0.33	0.25	-	0.67	0.03	0.01	4
Textile	0.89	0.39	0.32	2	0.85	0.16	0.15	2	0.93	0.68	0.56	76	0.91	**0.68**	0.57	76	0.87	0.37	0.28	-	0.87	0.03	0.01	2
Lime Plaster	0.87	0.41	0.32	3	0.85	0.31	0.25	3	0.90	0.57	0.46	90	0.90	**0.67**	0.56	90	0.90	0.60	0.49	-	0.60	0.02	0.01	3
Canvas	0.88	0.19	0.15	2	0.88	0.14	0.10	2	0.90	0.33	0.26	95	0.90	**0.57**	0.44	95	0.89	0.32	0.24	-	0.93	0.21	0.12	2
Tesserae	0.89	0.43	0.36	2	0.86	0.26	0.23	2	0.90	0.59	0.49	43	0.93	**0.72**	0.62	43	0.91	0.60	0.49	-	0.77	0.03	0.01	2
Paper	0.92	0.36	0.28	3	0.91	0.23	0.19	3	0.93	0.44	0.34	37	0.95	**0.60**	0.49	37	0.93	0.45	0.35	-	0.86	0.05	0.02	3
Glass	0.91	0.35	0.26	5	0.90	0.23	0.18	5	0.92	0.49	0.38	335	0.93	**0.63**	0.50	335	0.93	0.47	0.37	-	0.89	0.10	0.05	5
Film emulsion	0.94	0.38	0.35	2	0.95	0.62	0.54	2	0.95	0.52	0.45	261	0.97	**0.79**	0.69	261	0.93	0.55	0.47	-	0.94	0.08	0.04	2
Parchment	0.92	0.18	0.13	3	0.92	0.13	0.10	3	0.93	0.31	0.24	170	0.94	**0.58**	0.45	170	0.93	0.31	0.23	-	0.88	0.06	0.03	3
Geom Patterns	0.90	0.43	0.35	3	0.86	0.12	0.10	3	0.91	0.58	0.47	100	0.92	**0.65**	0.54	100	0.89	0.40	0.31	-	0.82	0.04	0.02	3
Line Art	0.93	0.33	0.23	3	0.93	0.20	0.16	3	0.95	0.46	0.36	38	0.95	**0.56**	0.45	38	0.95	0.44	0.33	-	0.92	0.05	0.03	3
Photo Depiction	0.92	0.36	0.27	3	0.91	0.30	0.24	3	0.93	0.45	0.35	163	0.94	**0.63**	0.51	163	0.93	0.45	0.35	-	0.91	0.11	0.06	3
Art Depiction	0.90	0.31	0.23	3	0.89	0.21	0.17	3	0.91	0.44	0.35	93	0.93	**0.63**	0.51	93	0.91	0.45	0.36	-	0.80	0.04	0.02	3
Overall	0.91	0.33	0.25	3	0.90	0.23	0.19	3	0.92	0.46	0.36	112	0.93	**0.63**	0.52	112	0.92	0.45	0.32	-	0.85	0.05	0.02	3



Figure 4.4: Qualitative results for SAM at zero-shot damage segmentation. Top row shows initial segments from N prompts as predicted by SAM (unique colour per prompt), bottom row shows the segments after being assigned binary class (Clean or Damaged) via an oracle.

4.6.2 Supervised Semantic Segmentation

We benchmark SOTA supervised semantic segmentation methods alongside linear probes trained over the features of vision foundation models on two tasks which require predicting pixel-wise segmentation maps: either for detecting damage (binary) or for detecting and classifying it (multiclass).

Evaluation protocol We split the data by Material and Content categories, performing a leave-one-out evaluation with one category left out for testing, and the rest split 8:2 into training and validation sets. This results in 14 splits: 10 for Material and 4 for Content. This approach assesses model performance on unseen media properties. For comparison, we train each model using a stratified split, ensuring all categories are in training, validation, and test sets. In total, each model is trained and evaluated 15 times across these settings.

Evaluated methods

Due to the range of scales exhibited by different damage types with respect to the image plane, we chose to evaluate segmentation models which have been shown to excel at recognition at various scales:

- **UPerNet** [217] is a framework which can leverage different vision backbones to learn from heterogeneous segmentation annotations. The flexibility of this framework allows us to evaluate both a convolutional variant, using a **ConvNeXt**[127] backbone, and a transformer variant, using **Swin Transformer** [126].
- **SegFormer** [218] utilises a hierarchically structured Transformer encoder which outputs multiscale features, and combines those with a lightweight Multilayer Perceptron (MLP).
- **DINOv2** [153] is a foundation vision model trained in a self-supervised setting, without any labels. The resulting features have been demonstrated to have emergent semantically meaningful properties. The authors utilise this by attaching a linear classifier, which is trained for semantic segmentation as a downstream task, achieving competitive results. We adopt this setup for our evaluation as well.
- **SAM** [108] is a foundation vision model that can perform zero-shot image segmentation given prompts. To assess how useful its feature space is for the task of damage detection and segmentation, we train a linear probe over the largest available SAM vision encoder (ViT-H) in the same setting as for DINOv2.

For all except SAM and DINOv2, we fine-tune from ADE20K [240] weights. Due to data imbalance towards Clean, we use macro-averaged Dice loss. We measure macro-averaged F1 Score and Mean Intersection over Union (mIoU), fine-tuning models until the validation F1 score has not improved for 10 epochs. For completeness, we also report macro-averaged Accuracy. Models are optimized with Adam [106], using the best learning rates and weight decay from the respective papers.

Data Augmentation

Our dataset consists of images of varying high resolutions and aspect ratios. During training, we apply standard augmentations: random scaling, cropping to the model's input size, and flipping. For validation and testing, we split our images into 512×512 pixel overlapping patches, make predictions on each patch, and stitch them back together using Hann windows [158] before calculating the metrics.

4.6.3 Diffusion-based segmentation

In addition to the supervised methods, we also benchmark two recent diffusion-based segmentation approaches. Following DiffEdit [51], we use contrasting predictions from a diffusion model conditioned on pairs of different single-word text prompts. Our positive prompts are

"Flawless", "Unblemished", "Undamaged", paired with corresponding negative prompts "Damaged", "Deteriorated", "Defaced". Each image is evaluated against all nine resulting prompt pairs; we report metrics for the pair with the best F1 score. This approach requires no training, relying on the learned prior of Stable Diffusion [167]. For our second approach, we benchmark DiffSeg [197], which is an unsupervised segmentation model relying on self-attention features extracted from Stable Diffusion. We set the KL-divergence threshold parameter to 0.5 in order to produce segments of higher granularity and successfully segment damage artefacts, which are often missed with the author-recommended value of 0.3 due to their relatively small scale. We use an oracle to assign each predicted segment to the Damaged or Clean class, same as with SAM in 4.6.1. Both methods are evaluated on inputs center-cropped and resized to 512×512; for DiffEdit we also provide results in full resolution in the Supplementary, Figure 4.14.

4.6.4 Results

Evaluation results are provided for all modes of prompting SAM in a zero-shot setting, all methods for the supervised semantic segmentation task (binary and multiclass setting) as well as for diffusion-based segmentation (text-guided and unsupervised). None of the benchmarked approaches manage to consistently perform well across different types of materials and contents.

SAM is missing the point. We found that SAM can produce good segmentation results only *if* prompted correctly (Figure 4.4), but doing so requires an impractical number of prompts. As shown in Table 4.1, using points-whether one per label or per artefact instance-yields poor results. Bounding boxes per class also perform poorly, especially when artefacts are spread across the image, as the boxes are too imprecise. Text-based prompts perform even worse, indicating that damage is not easily captured semantically. Point grid sampling also fails, as it does not guarantee all artefacts, especially fine ones, are sampled. The only method that provides adequate segmentation - achieving the highest average F1 score of 0.63 - is using a bounding box for each artefact instance, but this requires an average of 112 boxes per image (up to 3407 for some), making it impractical for real-world use in terms of both human effort and computational resources, and essentially solving the detection task for SAM a priori. Furthermore, an oracle was needed to assign labels to the segmented regions.

Supervised semantic segmentation. In the supervised setting, models can distinguish some damaged areas in the binary task, as seen in Figure 4.5 top row, but struggle with segmenting and classifying damage types in the multiclass setting for unseen contents or materials, regardless of pre-training or architecture, visualised in Figure 4.5 bottom row. Table 4.2 summarizes the binary segmentation results, where no model achieves an F1 score over 0.7. Notably, DINOv2 outperforms SAM despite being a smaller model. In the multiclass setting (Table 4.3), all models perform poorly, with no F1 score exceeding 0.3. Damage missclassifications are detailed in Section 4.9.3 of the Supplementary. DINOv2 performs best, likely due to its semantically meaningful features [153], aiding in differentiating damage types. SAM, however, struggles to

distinguish between damage types and produces imprecise segmentations when trained to segment all artefacts simultaneously, as it is unable to semantically group artefacts of the same type together within its feature space. This underscores the fact that the SAM is wholly inappropriate for this task, where multiple instances of damage may be present in the same image, and the model needs to meaningfully distinguish them.

Table 4.2: Results across both Material and Content splits, for all baselines, trained on the task of binary damage segmentation. Best F1 scores for each test class in bold. Last row is a stratified split over Material x Content.

Test Class	**Segformer**			**UPerNet + Swin**			**UPerNet + ConvNeXt**			**DINOv2 + MLP**			**SAM ViT-H + MLP**		
	Acc	**F1**	**mIoU**	**Acc**	**F1**	**mIoU**	**Acc**	**F1**	**mIoU**	**Acc**	**F1**	**mIoU**	**Acc**	**F1**	**mIoU**
Wood	0.85	0.48	0.32	0.84	0.36	0.22	0.85	**0.59**	0.23	0.85	0.47	0.31	0.81	0.40	0.25
Ceramic	0.87	0.46	0.46	0.82	0.43	0.27	0.83	0.46	0.30	0.82	**0.61**	0.44	0.80	0.37	0.23
Textile	0.83	0.49	0.33	0.86	**0.60**	0.43	0.88	**0.60**	0.43	0.89	0.59	0.42	0.84	0.48	0.30
Lime Plaster	0.85	**0.66**	0.49	0.84	0.58	0.40	0.83	0.64	0.48	0.82	0.61	0.44	0.79	0.52	0.35
Canvas	0.93	**0.70**	0.53	0.92	0.69	0.52	0.92	0.67	0.50	0.90	0.62	0.45	0.82	0.50	0.33
Tesserae	0.85	**0.54**	0.37	0.81	0.47	0.31	0.84	0.47	0.31	0.85	0.52	0.36	0.79	0.47	0.30
Paper	0.92	0.56	0.39	0.90	0.53	0.35	0.92	**0.59**	0.42	0.89	0.50	0.34	0.84	0.44	0.28
Glass	0.89	0.30	0.17	0.89	0.37	0.22	0.90	0.44	0.28	0.88	**0.46**	0.30	0.86	0.36	0.22
Film emulsion	0.85	0.34	0.20	0.85	0.50	0.33	0.93	**0.71**	0.56	0.86	0.46	0.30	0.89	0.63	0.46
Parchment	0.89	**0.44**	0.28	0.90	**0.44**	0.28	0.89	0.35	0.21	0.88	0.35	0.21	0.84	0.32	0.19
Geom Patterns	0.90	**0.64**	0.48	0.89	0.61	0.44	0.84	0.57	0.39	0.89	0.62	0.45	0.86	0.52	0.35
Line Art	0.93	**0.62**	0.45	0.93	0.55	0.38	0.92	0.56	0.39	0.89	0.41	0.26	0.90	0.46	0.30
Photo Depiction	0.87	0.41	0.26	0.89	**0.58**	0.40	0.89	0.54	0.37	0.85	0.45	0.29	0.83	0.46	0.30
Art Depiction	0.88	**0.49**	0.33	0.86	0.41	0.26	0.87	0.43	0.27	0.87	0.45	0.29	0.84	0.36	0.22
Stratified (all)	0.91	0.63	0.46	0.91	0.60	0.43	0.91	0.65	0.48	0.89	0.59	0.42	0.87	0.53	0.36

Table 4.3: Results across both Material and Content splits, for all baselines, trained on the task of multiclass damage segmentation. Best F1 scores for each test class in bold. Last row is a stratified split over Material x Content.

Test Class	**Segformer**			**UPerNet + Swin**			**UPerNet + ConvNeXt**			**DINOv2 + MLP**			**SAM ViT-H + MLP**		
	Acc	**F1**	**mIoU**	**Acc**	**F1**	**mIoU**	**Acc**	**F1**	**mIoU**	**Acc**	**F1**	**mIoU**	**Acc**	**F1**	**mIoU**
Wood	0.14	0.11	0.08	0.14	0.10	0.08	0.14	0.11	0.08	0.23	**0.20**	0.15	0.05	0.06	0.05
Ceramic	0.18	0.13	0.10	0.17	0.13	0.10	0.17	0.12	0.09	0.28	**0.20**	0.16	0.10	0.10	0.08
Textile	0.10	0.10	0.08	0.10	0.10	0.09	0.09	0.09	0.08	0.20	**0.18**	0.15	0.07	0.07	0.06
Lime Plaster	0.13	0.12	0.09	0.14	0.14	0.10	0.18	0.17	0.12	0.38	**0.30**	0.22	0.07	0.08	0.06
Canvas	0.17	**0.17**	0.13	0.17	0.16	0.13	0.18	**0.17**	0.13	0.18	**0.17**	0.13	0.06	0.07	0.05
Tesserae	0.07	0.08	0.06	0.09	0.08	0.07	0.08	0.09	0.07	0.16	**0.15**	0.12	0.06	0.06	0.05
Paper	0.14	0.11	0.09	0.17	**0.13**	0.10	0.16	0.12	0.10	0.18	0.11	0.09	0.07	0.06	0.05
Glass	0.15	0.15	0.12	0.18	**0.19**	0.14	0.16	0.17	0.13	0.15	0.14	0.11	0.07	0.07	0.05
Film emulsion	0.12	0.07	0.06	0.10	0.08	0.07	0.11	**0.09**	0.07	0.13	0.06	0.05	0.05	0.06	0.05
Parchment	0.14	**0.14**	0.10	0.12	0.12	0.09	0.15	0.13	0.10	0.18	**0.14**	0.11	0.07	0.07	0.06
Geom Patterns	0.15	0.08	0.05	0.20	0.17	0.13	0.17	0.16	0.12	0.23	**0.19**	0.15	0.09	0.09	0.07
Line Art	0.15	**0.15**	0.11	0.18	0.14	0.12	0.16	0.13	0.10	0.20	0.13	0.11	0.09	0.06	0.05
Photo Depiction	0.13	0.12	0.09	0.13	0.12	0.10	0.16	**0.13**	0.10	0.14	0.10	0.08	0.07	0.06	0.04
Art Depiction	0.12	0.11	0.09	0.13	0.11	0.09	0.15	**0.14**	0.10	0.17	0.12	0.09	0.08	0.07	0.05
Stratified (all)	0.21	0.20	0.15	0.23	0.26	0.21	0.17	0.18	0.14	0.30	0.25	0.18	0.12	0.12	0.09



Figure 4.5: Qualitative comparison for supervised binary (top row) and multiclass (bottom row) damage segmentation.

Diffusion-based segmentation. In diffusion-segmentation approaches, we found that while models may localize or cluster damaged areas—such as in DiffSeg—the predictions are too imprecise for acceptable segmentation quality. As shown in Table 4.4, the results are comparable to those from supervised models, with DiffEdit performing worse than DiffSeg. Qualitatively, DiffEdit sometimes identifies damaged areas but often highlights unrelated features, such as faces and high-frequency details, due to diffusion models' tendency to generate fine details at later steps [129]. DiffSeg performs slightly better, aided by an oracle for labeling, but while it groups damage instances effectively due to the emergent semantic meaning found in Stable Diffusion's features—its segmentation remains imprecise since it relies on low-dimensional features.

Table 4.4: Results across both Material and Content splits diffusion based text-guided (DiffEdit) and unsupervised (DiffSeg) binary damage segmentation.

Test class	DiffEdit			DiffSeg		
	Acc	**F1**	**mIoU**	**Acc**	**F1**	**mIoU**
Wood	0.77	0.13	0.07	0.90	**0.18**	0.15
Ceramic	0.70	0.13	0.07	0.90	**0.39**	0.31
Textile	0.79	0.19	0.11	0.93	**0.57**	0.47
Lime Plaster	0.72	0.18	0.10	0.90	**0.42**	0.35
Canvas	0.79	0.21	0.12	0.91	**0.28**	0.21
Tesserae	0.76	0.11	0.06	0.92	**0.53**	0.45
Paper	0.81	0.16	0.09	0.94	**0.29**	0.23
Glass	0.78	0.19	0.11	0.94	**0.38**	0.30
Film emulsion	0.78	0.08	0.04	0.98	**0.41**	0.43
Parchment	0.83	0.13	0.07	0.93	**0.18**	0.15
Geom Patterns	0.80	0.20	0.12	0.94	**0.45**	0.39
Line Art	0.84	0.14	0.08	0.95	**0.16**	0.13
Photo Depiction	0.79	0.18	0.10	0.95	**0.36**	0.30
Art Depiction	0.79	0.13	0.07	0.93	**0.30**	0.24

4.7 Conclusion

Damage is ubiquitous in analogue media, particularly in heritage artifacts. While inpainting for restoration is relatively well-solved, reliably identifying damaged areas from digital images remains a significant challenge. We have introduced a damage taxonomy spanning diverse materials and contents to guide detection systems, and our extensive, permissively-licensed dataset sets a rigorous baseline for future work. Our benchmark results show that no state-of-the-art method performs acceptably in detecting damage. Supervised models fail to generalize to unseen materials and contents, while foundation models require excessive prompt engineering and still struggle to label damage accurately. There is substantial room to develop machine learning pipelines that can detect damage at a human-equivalent level, particularly for the pixel-perfect precision required in conservation. Our experiments highlight the complexity of the task, and we hope our dataset will inspire advancements in this area.

4.8 Afterword

The unique challenges of damage detection distinguish it from conventional segmentation tasks. While typical semantic segmentation deals with objects having consistent visual and semantic properties, the ARTeFACT dataset demonstrates that damage manifests with visual diversity and dramatic variance in scale, shape, opacity, density, and distribution patterns across different media types. Systematic evaluation of state-of-the-art architectures established that damage detection cannot be treated as a simple extension of existing segmentation approaches.

The empirical limitations revealed through the ARTeFACT benchmark point to a fundamental challenge: models need to learn damage as an abstract semantic concept rather than merely a collection of low-level visual patterns. Even with comprehensive training data spanning multiple media types, supervised architectures failed to develop robust representations, exhibiting poor generalisation to unseen materials and contents.

Multimodal foundation models, particularly generative text-to-image diffusion approaches such as Stable Diffusion, suggested a promising alternative direction. Concurrent work demonstrated these models' feature spaces encode rich semantic information despite no explicit supervision, which can be successfully utilised in downstream tasks such as keypoint matching [129], depth estimation [41], and relevantly—semantic segmentation [144, 197]. Indeed, initial experiments with DiffEdit and DiffSeg revealed useful emergent properties for damage detection, while highlighting two key challenges: controlling segmentation granularity and disambiguating damage from intentional artistic elements or semantic content through text conditioning alone.

This observation motivated rethinking the segmentation problem itself. Rather than developing another domain-specific approach, we investigated whether a general unsupervised framework could leverage the representations in text-to-image (T2I) diffusion models while providing principled control over segmentation granularity—a limitation identified in DiffSeg. The next chapter explores this idea, proposing a novel framework for unsupervised segmentation based on T2I diffusion model features. While not specific to damage, this approach demonstrates strong performance in semantic segmentation without requiring any training data, and importantly, offers a mechanism for controlling hierarchical grouping through matrix exponentiation. This suggests that T2I diffusion models not only learn to distinguish meaningful visual entities and group them semantically, but that their self-attention mechanisms can be systematically leveraged through spectral clustering methods.

4.9 Supplementary

4.9.1 Data collection & curation

We sourced most of our dataset from the digitised collections of the following cultural heritage institutions:

- The Metropolitan Museum of Art
- The US National Archives
- The Library of Congress
- The Philadelphia Museum of Art
- The J. Paul Getty Museum
- The Louvre
- The Rijksmuseum
- The Victoria & Albert Museum
- The British Museum
- The British Library
- The National Gallery of Scotland
- The Burrell Collection
- The British National Portrait Gallery
- The National Library of Ireland
- The Fitzwilliam Museum
- The Bavarian National Museum
- The Albert-Kahn Museum

The remainder of the images were collected from Wikimedia, and from Flickr groups dedicated to vintage photographs, such as Wallet Worn and Vintage Scans. All images were made available in the Public Domain by their authors, and collected under CC0, CC-BY and CC-BY-NC licenses. For each image we acquired the highest resolution available, with the average image height and width in our dataset being 2496 and 2306 pixels respectively, and the maximum height and width reaching 11233 and 12797 pixels, allowing for flexibility in preprocessing depending on the task. All text annotations (Content and Materual classifications, Type classification and descriptions) are provided in a tabular format along with source attribution. Ground-truth segmentation masks are provided in a standard format as single-channel PNG files, along colour-coded RGB segmentations for visualisation purposes. Examples shown in Figure 4.6.



Type: Painting
Material: Wood
Content: Artistic depiction
LLaVA description: A woman wearing a dress and a crown is depicted in a painting on wood.
Our description: Portrait painting of woman wearing a dress and a crown.
Damage description: The painting has cracks, dirt spots, scratches and peels.



Type: Photo
Material: Paper
Content: Photographic depiction
LLaVA description: A man wearing a uniform and a hat is smoking a cigarette.
Our description: Photo portrait of a man wearing a uniform.
Damage description: The photo has staining, folds, scratches and peels.



Type: Book cover
Material: Paper
Content: Artistic depiction
LLaVA description: A book cover featuring a woman and a bird, with the woman wearing a flowing dress and holding a bird in her hand.
Our description: Book cover illustration of a woman in a dress in a mirror facing a bird with long feathers sitting on a branch.
Damage description: The book cover has stickers, writings, folds, scratches, peels and tears.

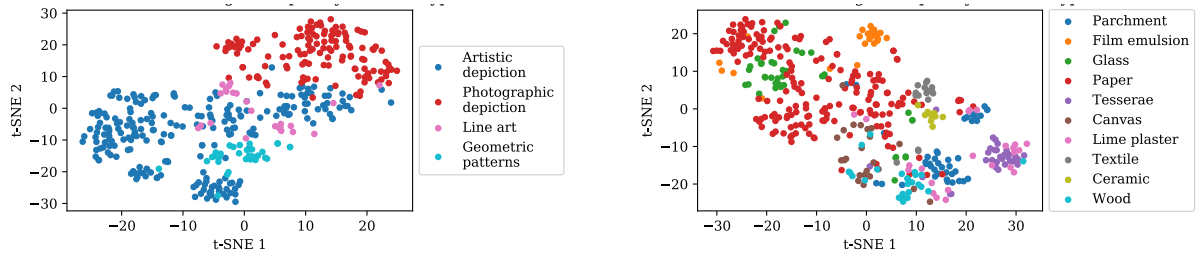


Type: Mosaic
Material: Tesserae
Content: Artistic depiction
LLaVA description: A mosaic on tesserae depicting a cat and a duck, with the cat standing over the duck, and a bird in the foreground.
Our description: Mosaic of a cat catching a duck above other ducks eating from a pile of fish and clams in the foreground.
Damage description: The mosaic has cracks and peels.



Type: Tile
Material: Ceramic
Content: Geometric pattern
LLaVA description: A tile on ceramic depicting a green and red design, with a flower pattern and a cross in the center.
Our description: Ceramic tile depicting a green and red design, with a flower pattern and a cross in the center.
Damage description: The tile has discolourations, scratches and peels.

Figure 4.6: Example segmentation, class and text annotations from our dataset.



(a) Image embeddings coloured by Content **(b)** Image embeddings coloured by Material

Figure 4.7: CLIP embeddings of the images in the dataset, dimensionality reduced via t-SNE. Each colour corresponds to a material (a) or content (b) category as provided in our dataset.

4.9.2 Image-level caterogorisation and textual descriptions

Content and Material categories.

Figure 4.7 a) shows that our Content categories assigned by our experts are well separeted in the CLIP image feature space, since CLIP focuses on semantic image content. Material labels form “fuzzier" regions in Figure 4.7 b), demonstrating the diversity of objects depicted in the images across Material classes, while also indicating that certain subject matters are more commonly depicted on certain materials, e.g. Textile will most commonly have stylised Geometric pattern depictions.

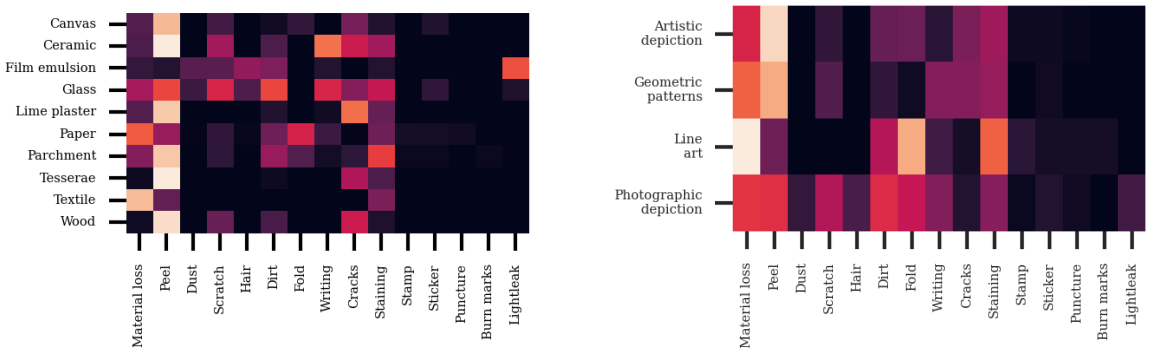


Figure 4.8: Relative damage type frequency per material (a) and content (b) class. Lighter colours indicate higher frequency.

It can also be observed that certain types of damage occur more frequently in images of certain content, as shown in Figure 4.8 b), however, this is less pronounced than for different material types, as shown in Figure 4.8 a).

Textual descriptions.

We provide textual descriptions for each image, detailing both content and damage types. The expert descriptions are derived by correcting draft captions produced by LLaVA [124], by

Table 4.5: Cosine distances between the CLIP embeddings of image descriptions generated by LLaVA and our descriptions, averaged across Material and Content classes.

LLaVA vs	Material										Content			
	Canvas	Ceramic	Film emulsion	Glass	Lime plaster	Paper	Parchment	Tesserae	Textile	Wood	Artistic depiction	Geometric patterns	Line art	Photographic depiction
Ours	0.801	0.861	0.731	0.772	0.744	0.73	0.700	0.792	0.752	0.747	0.742	0.789	0.682	0.751
LLaVA + Damage description	0.917	0.924	0.818	0.895	0.928	0.909	0.884	0.971	0.951	0.935	0.920	0.935	0.923	0.883
Ours + "Damaged"	0.740	0.812	0.649	0.696	0.713	0.659	0.669	0.756	0.722	0.705	0.699	0.746	0.646	0.656
Ours + Damage description	0.761	0.820	0.613	0.716	0.743	0.689	0.665	0.787	0.736	0.725	0.715	0.766	0.660	0.685

prompting it with the following: *This is a <type> on <material>. Describe what this image depicts concisely, in a single sentence, but be as detailed as possible, also noting the initial information about the type of image and the media it is depicted on.* The resulting descriptions were then manually verified and corrected by our annotation experts. We derived separate damage descriptions from our pixel-level annotations. To illustrate how much human experts improve on LLM-generated captions, we computed cosine similarity between LLaVA and expert descriptions using CLIP text embeddings (Table 4.5), quantifying LLaVA's "incorrectness" regarding semantic image content. We also report distances between original LLaVA descriptions and: extended LLaVA descriptions, extended expert descriptions, and expert descriptions prefixed with "Damaged." Specifying damage presence decreases cosine similarity, showing LLaVA's failure to recognize damage, often ignoring or misinterpreting it. For instance, in the second example in Figure 4.6, LLaVA's caption is `A man wearing a uniform and a hat is smoking a cigarette`, entirely omitting the presence of damage, and misinterpreting a peeled area pf the photograph as a cigarette. In contrast, our expert annotation is `Photo portrait of a man wearing a uniform`, while the damage description reads `The photo has staining, folds, scratches and peels`.

4.9.3 Qualitative assessment

Supervised segmentation

We qualitatively compared the segmentation masks produced by various models across distinct materials. We show the outcomes for both binary and multi-class segmentation tasks. The selected images cover a wide range of damage type where state-of-the-art models encounter difficulties in achieving optimal performance.

Binary segmentation Binary segmentation is the easier setting of the task, where the models learn to differentiate between damaged and non-damaged areas of the images, without needing to classify the type of damage. The qualitative results shown in Figure 4.9 for Material categories and in Figure 4.10 for Content categories support our quantitative evaluation results, showing that SegFormer marginally outperforms the other models, but overall model performance is poor. All models are prone to making false positive predictions, and struggle with both small-scale artefacts such as Dust and Hairs in photos. The DinoV2 + MLP model, which is the worst performing according to the quantitative evaluation, fails at predicting large missing areas, and

instead only focuses on the edges of such artefacts (e.g in Glass, Tesserae, Artistic depiction categories).

Table 4.6: Weighted average metrics per damage types across Material categories for each model at the task of multiclass semantic segmentation of damage.

Damage Type	Segformer			UPerNet + Swin			UPerNet + ConNeXt			DinoV2 + MLP			SAM ViT-H + MLP		
	Acc	F1	IoU	Acc	F1	IoU	Acc	F1	IoU	Acc	F1	IoU	Acc	F1	IoU
Material loss	0.458	0.543	0.38	0.767	**0.71**	0.572	0.713	0.685	0.548	0.517	0.427	0.277	0.294	0.095	0.051
Peel	0.236	0.294	0.176	0.31	0.371	0.233	0.327	**0.392**	0.249	0.395	0.359	0.226	0.101	0.108	0.059
Dust	0	0	0	0	0	0	0	0	0	0	0	0	0.005	0.004	0.002
Scratch	0.008	0.007	0.004	0.047	0.049	0.026	0.087	**0.101**	0.056	0	0	0	0.017	0.009	0.004
Hair	0	0	0	0	0	0	0	0	0	0	0	0	0	0.001	0
Dirt	0.04	0.047	0.025	0.141	0.132	0.071	0.149	**0.14**	0.076	0.007	0.003	0.002	0.008	0.007	0.004
Fold	0.021	0.018	0.01	0.493	0.489	0.33	0.583	**0.517**	0.356	0.012	0.009	0.005	0.009	0.013	0.007
Writing	0.292	0.245	0.167	0.474	**0.405**	0.263	0.448	0.36	0.224	0.592	0.302	0.197	0.016	0.024	0.013
Cracks	0.051	0.062	0.033	0.099	0.131	0.073	0.124	0.14	0.078	0.245	**0.216**	0.126	0.02	0.029	0.015
Staining	0.038	0.036	0.019	0.245	0.221	0.141	0.287	**0.257**	0.168	0.023	0.024	0.013	0.002	0.003	0.001
Stamp	0	0	0	0	0	0	0	0	0	0	0	0	0.003	0	0
Sticker	0	0	0	0	0	0	0	0	0	0	0	0	0.017	0.002	0.001
Puncture	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Burn marks	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Lightleak	0	0	0	0	0	0	0	0	0	0.008	**0.007**	0.004	0.003	0.003	0.001

Table 4.7: Weighted average metrics per damage types across Content categories for each model at the task of multiclass semantic segmentation of damage.

Damage Type	Segformer			UPerNet + Swin			UPerNet + ConNeXt			DinoV2 + MLP			SAM ViT-H + MLP		
	Acc	F1	IoU	Acc	F1	IoU	Acc	F1	IoU	Acc	F1	IoU	Acc	F1	IoU
Material loss	0.572	**0.522**	0.357	0.576	0.517	0.356	0.621	0.517	0.353	0.539	0.409	0.262	0.306	0.115	0.061
Peel	0.181	0.231	0.133	0.2	0.218	0.124	0.207	0.24	0.138	0.345	**0.333**	0.206	0.069	0.078	0.042
Dust	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Scratch	0.001	0.001	0.001	0.012	**0.02**	0.01	0.011	0.012	0.006	0	0	0	0.04	0.008	0.004
Hair	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Dirt	0.02	0.012	0.007	0.046	0.017	0.009	0.06	**0.075**	0.039	0	0	0	0.002	0.002	0.001
Fold	0.197	0.155	0.086	0.166	0.128	0.07	0.188	**0.186**	0.104	0.194	0.131	0.07	0.021	0.021	0.011
Writing	0.106	0.021	0.012	0.172	0.095	0.063	0.216	0.112	0.068	0.433	**0.14**	0.079	0.033	0.023	0.012
Cracks	0.021	0.006	0.003	0.026	0.011	0.006	0.026	**0.012**	0.006	0.016	0.009	0.005	0.003	0.004	0.002
Staining	0.04	0.034	0.018	0.131	0.077	0.041	0.083	**0.079**	0.041	0.004	0.005	0.002	0.019	0.014	0.007
Stamp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Sticker	0	0	0	0	0	0	0	0	0	0	0	0	0.057	0.001	0
Puncture	0	0	0	0	0	0	0.051	**0.053**	0.029	0	0	0	0	0	0
Burn marks	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Lightleak	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Multi-class segmentation In the multi-class segmentation setting, models also need to differentiate between the different types of damage. We provide qualitative results in Figure 4.11 across Material test categories, and in Figure 4.12 across Content test categories. The predictions are similarly poor as in the binary setting across all four models. We can observe that even when the damaged area is correctly detected, models often assign it an incorrect class. We further examine this behavior by computing metrics for each specific damage type, as summarised in Table 4.6 and Table 4.7 for Material and Content categories respectively. The values in the tables are averaged across all LOOCV splits and weighted based on the frequency of Damage type per each Material or Content category. In some cases (LOOCV Content splits), all instances of a certain type of damage belong in the same test class - e.g., all Lightleaks would be in the Photographic depiction category, hence they are either all in the training or the test data.

Qualitatively, performance is exhibited across all damage types in the Content splits, including ones such as Material loss and Peel, which are common in multiple categories. We see that the models cannot successfully segment large-scale damage; they also fail at detecting very small-scale damage types, such as Dust, Scratches, and Hairs, contributing to its poor performance in the Photographic depiction class.

When Lightleaks are not unique to a specific category, as in the Material split, Table 4.6 shows that the linear probes of DinoV2 and SAM are the only models which score above 0 in all three metrics for this damage type, but still achieve an extremely low score. Dust and Scratches are again where the other three models fail as well; furthermore, all models struggle with Stamps, Stickers, Burn marks.

Where models do make somewhat successful predictions, they show similar performance to the binary setting, highlighting edges and failing at more "blobby" damage instances. Still, false positives are incredibly common, with models detecting arbitrary image features as damage.

4.9.4 Zero-shot generative segmentation

In addition to quantitatively comparing DiffSeg and DiffEdit in Section 4.6.3, we visualise qualitative results in Figure 4.13. Additionally, we show a qualitative comparison between DiffEdit used on center-cropped images and in full resolution following the patch-process methodology described in Section 4.6.4, Figure 4.14. Quantitative results are shown in Table 4.8. Results are presented over all categories of Material and Content, and cover various types, scales, and degree of damage.

DiffEdit

We can observe that while the approach may sometimes predict damaged areas correctly, it tends to also highlight generally salient features, discriminative features such as faces, as well as high frequency features, none of which are related to the prompt. The detection of high-frequency features is particularly common when working in full resolution. Since the denoising process is performed over the later portion of steps in the denoising chain, this behaviour can be explained via the property of diffusion models to generate more fine-detailed features at later steps [129]. We find that the approach is extremely sensitive to scale, making inconsistent predictions over center-cropped areas compared to over the entire images. The results support our quantitative evaluation, demonstrating further that language guidance is too imprecise to describe damage. Furthermore, we show that the averaged noise prediction approach used by DiffEdit does not provide disentanglement between predicted noise variance from language conditioning and predicted noise variance over high-frequency features due to the choice of denoising step range. These results illustrate that while modern multi-modal generative approaches may show promising results in general purpose image segmentation via text conditioning, they

fail at the task of damage detection - this highlights the complexity of our task.

DiffSeg

Although DiffSeg achieves better quantitative scores (partly due to the use of an oracle), we can see in Figire 4.13 that it struggles with higher frequency damage. This is due to the fact that with this approach the segmentation process is performed at a very low-resolutions at which the self-attentions is extracted, the highest being 64×64; the even lower latent resolutions of 16×16 and 32×32 might be helpful when forming general semanic clusters, but here they also fail to capture finer artefacts.

Table 4.8: Results across both Material and Content splits for DiffEdit binary mask generation in both center cropped and full resolution settings.

Test class	**DiffEdit (Crop)**			**DiffEdit (Full Res)**		
	Acc	**F1**	**mIoU**	**Acc**	**F1**	**mIoU**
Wood	0.77	0.13	0.07	0.74	0.15	0.08
Ceramic	0.70	0.13	0.07	0.67	0.19	0.10
Textile	0.79	0.19	0.11	0.74	0.18	0.10
Lime Plaster	0.72	0.18	0.10	0.70	0.18	0.10
Canvas	0.79	0.21	0.12	0.78	0.26	0.15
Tesserae	0.76	0.11	0.06	0.71	0.90	0.05
Paper	0.81	0.16	0.09	0.78	0.19	0.11
Glass	0.78	0.19	0.11	0.76	0.21	0.12
Film emulsion	0.78	0.08	0.04	0.73	0.11	0.06
Parchment	0.83	0.13	0.07	0.78	0.11	0.06
Geom Patterns	0.80	0.20	0.12	0.76	0.19	0.11
Line Art	0.84	0.14	0.08	0.79	0.14	0.08
Photo Depiction	0.79	0.18	0.10	0.77	0.22	0.13
Art Depiction	0.79	0.13	0.07	0.75	0.14	0.08



Figure 4.9: Qualitative comparison for binary damage segmentation across Material categories.



Figure 4.9: Qualitative comparison for binary damage segmentation across Material categories.

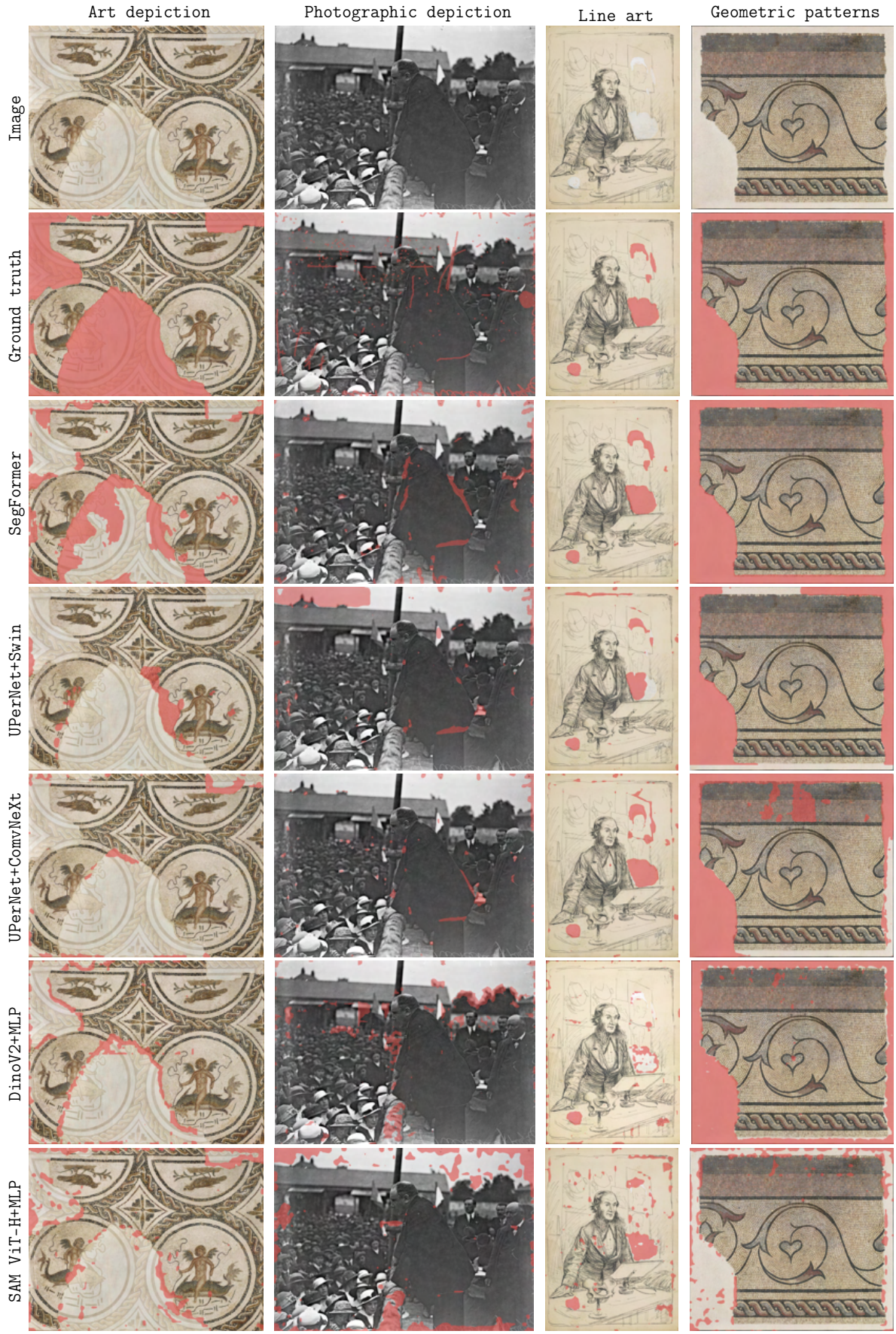


Figure 4.10: Qualitative comparison for binary damage segmentation across Content categories.

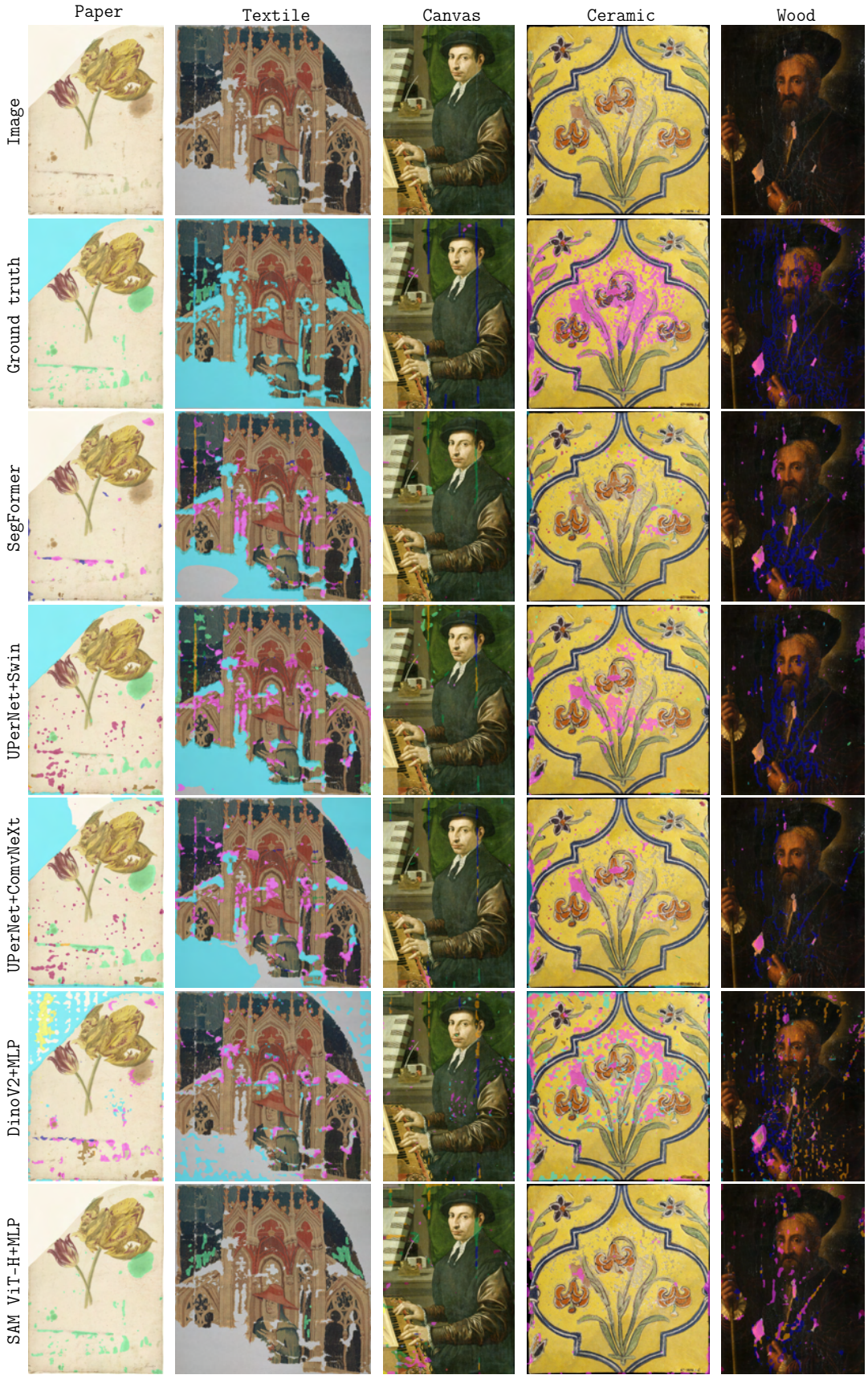


Figure 4.11: Qualitative comparison for binary damage segmentation across Material categories.



Figure 4.11: Qualitative comparison for multiclass damage segmentation across Material categories.



Figure 4.12: Qualitative comparison for binary damage segmentation across Content categories.

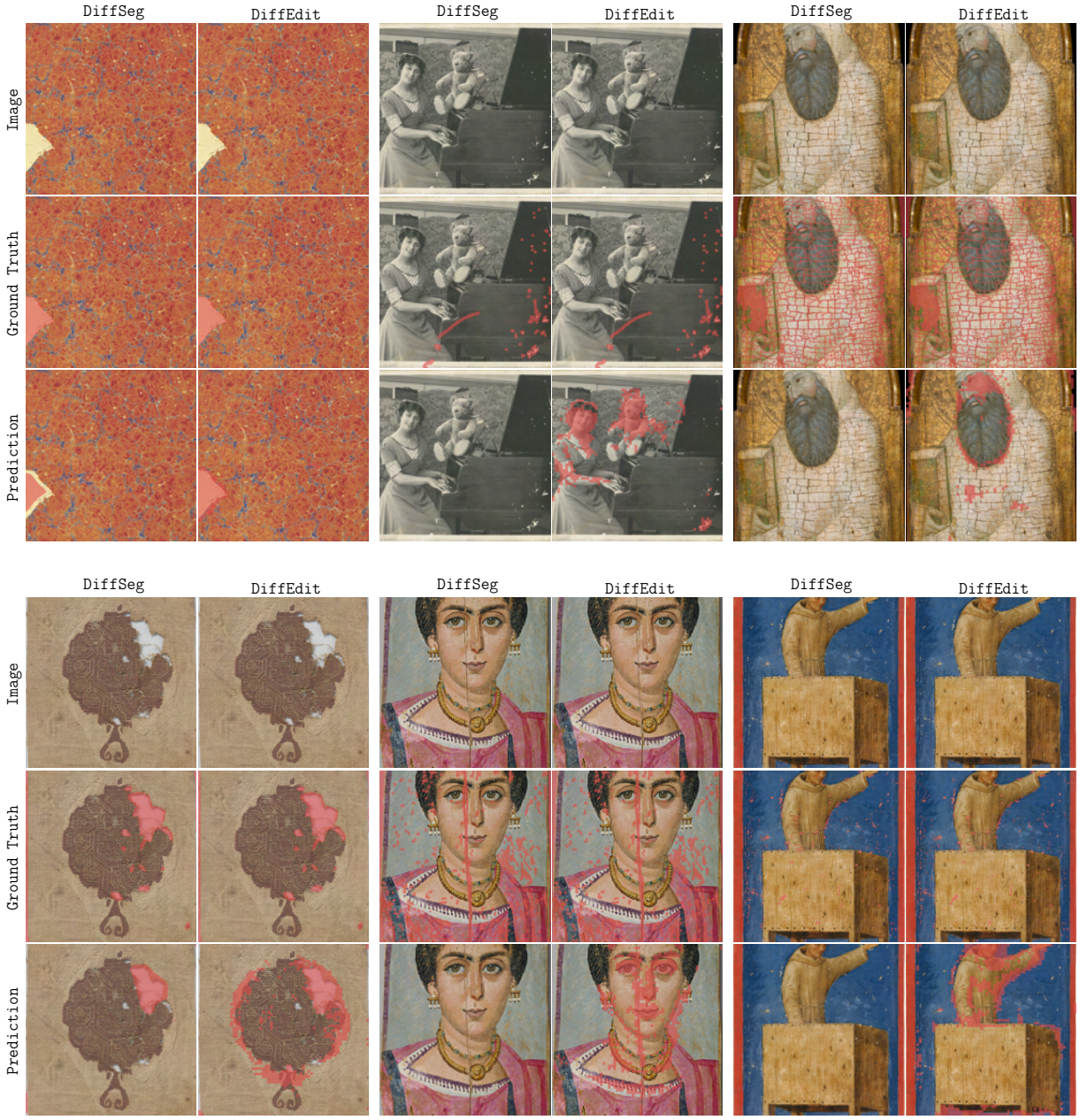


Figure 4.13: Qualitative comparison for diffusion-based damage segmentation between DiffSeg and DiffEdit.

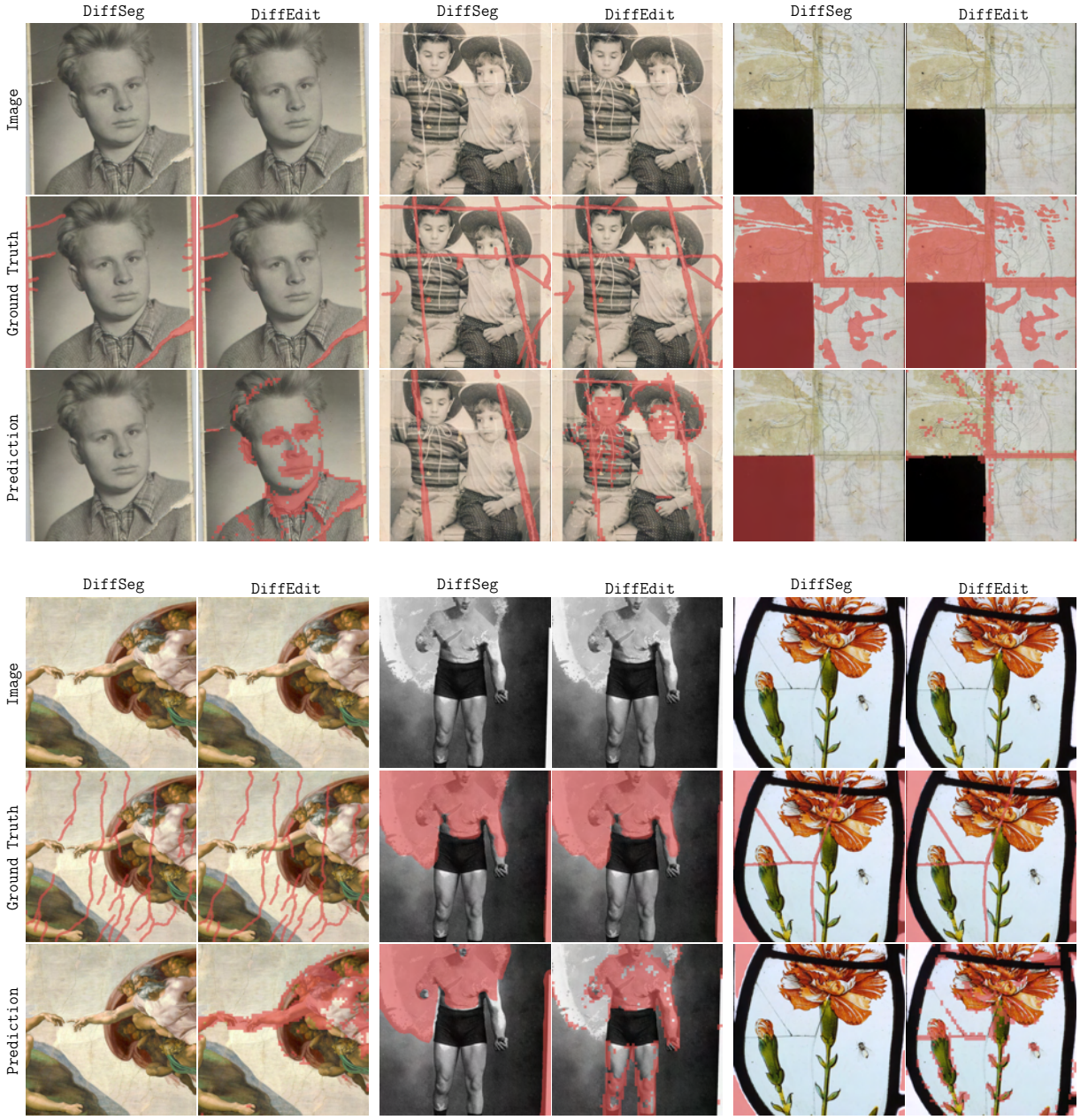


Figure 4.13: Qualitative comparison for diffusion-based damage segmentation between DiffSeg and DiffEdit.

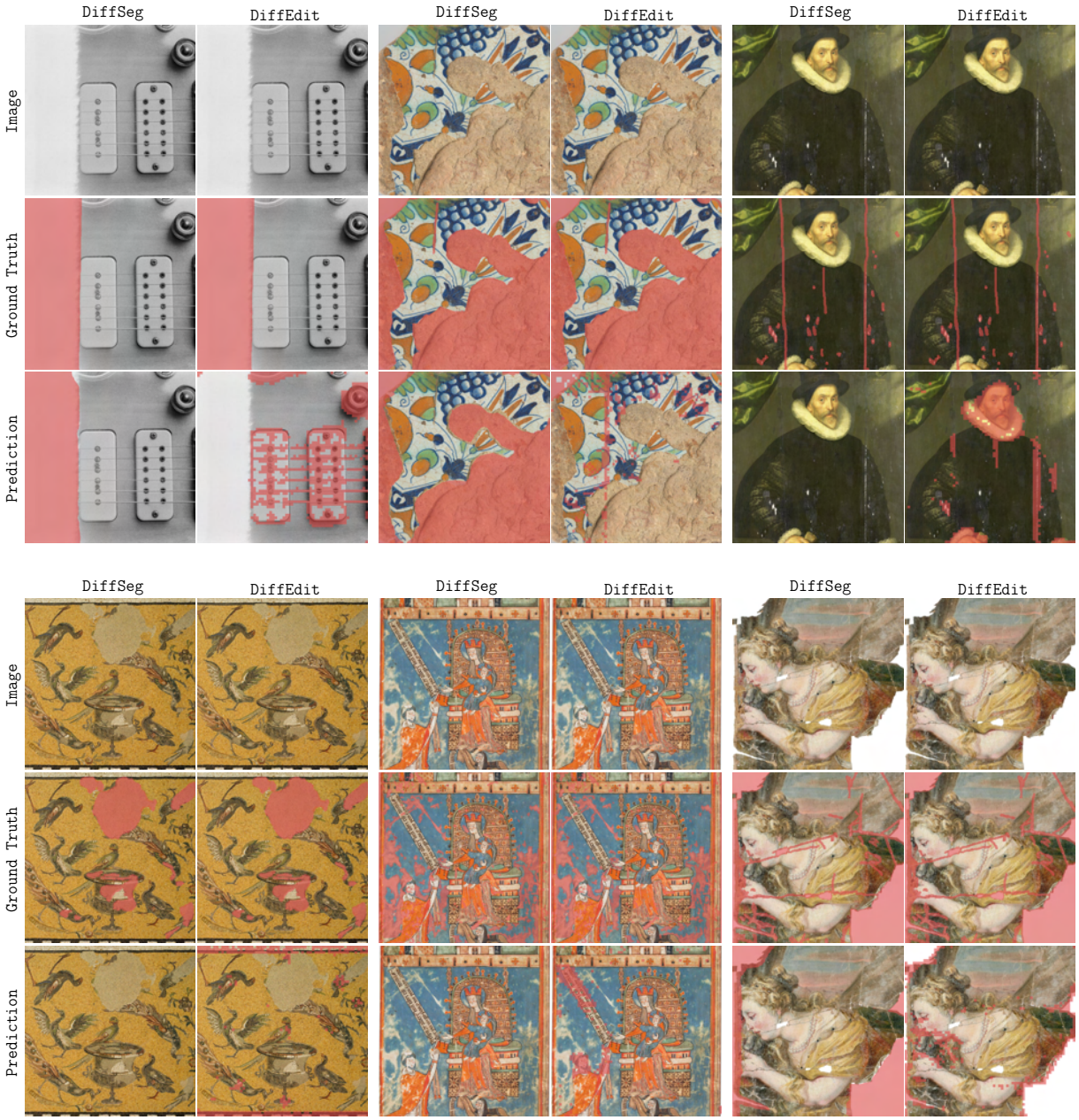


Figure 4.13: Qualitative comparison for diffusion-based damage segmentation between DiffSeg and DiffEdit.



Figure 4.14: Qualitative comparison for text-guided damage segmentation using DiffEdit over center-cropped images and in full resolution.



Figure 4.14: Qualitative comparison for text-guided damage segmentation using DiffEdit over center-cropped images and in full resolution.

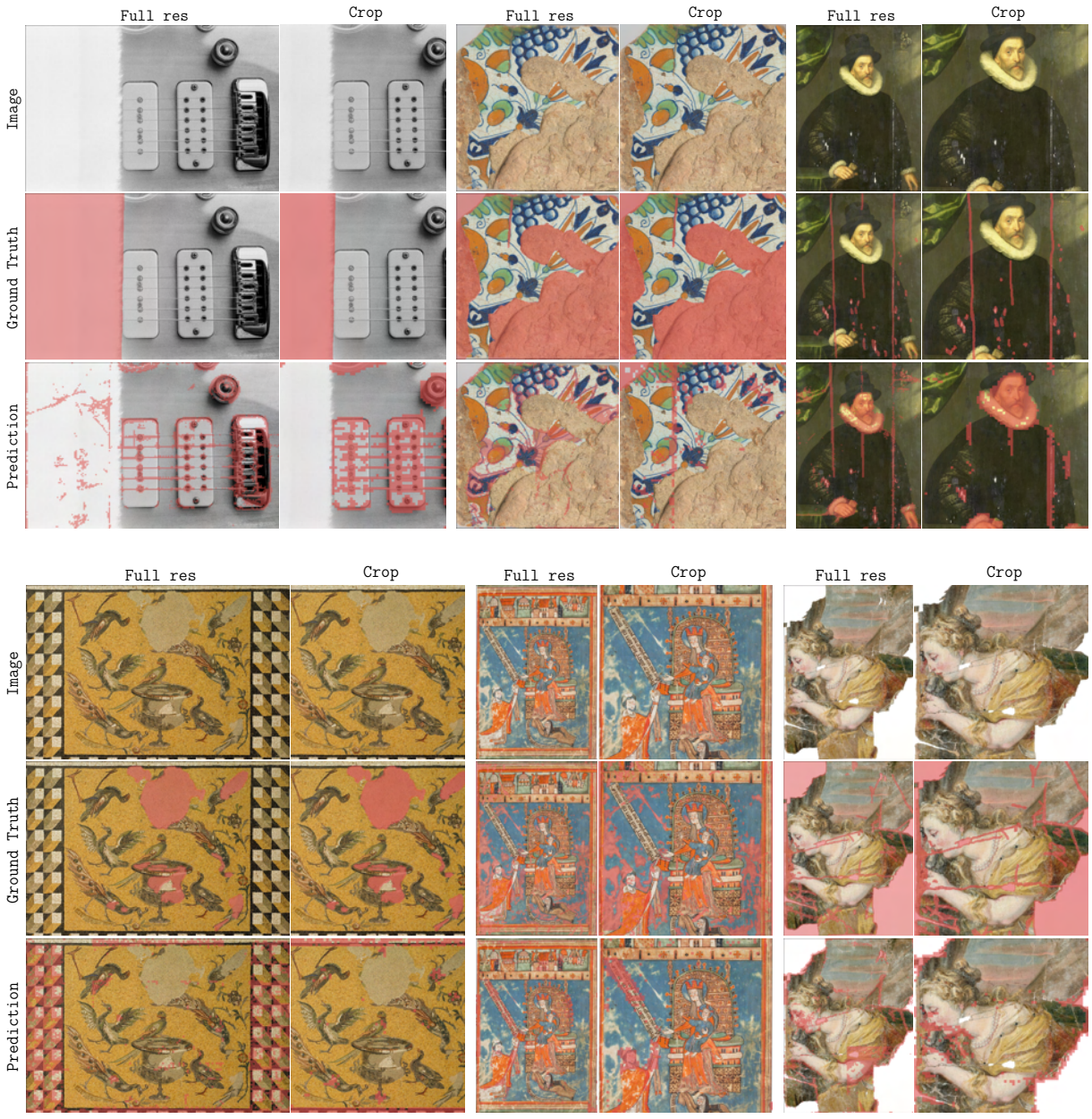


Figure 4.14: Qualitative comparison for text-guided damage segmentation using DiffEdit over center-cropped images and in full resolution.

Chapter 5

Unsupervised Segmentation by Diffusing, Walking and Cutting

The majority of this chapter (Section 5.3 through to Section 5.7) reproduces the author's version of the peer review manuscript submitted at the The IEEE/CVF Winter Conference on Applications of Computer Vision 2026 (WACV) as *Ivanova, D., Aversa, M., Henderson, P. and Williamson, J. "Unsupervised Segmentation by Diffusing, Walking and Cutting."* arXiv preprint arXiv:2412.04678 and does not violate the copyright of the publisher. The majority of the work presented below was undertaken by myself. I implemented the model self-attention feature extraction and Shi and Malik's Normalised Cuts algorithm. I formulated and carried out all experiments involving the benchmarking of unsupervised segmentation models, with the exception of verifying the results presented by EmerDiff, which was carried out by Marco Aversa. I drafted and rewrote the paper, with my supervisors Dr John Williamson and Dr Paul Henderson providing edits and corrections.

5.1 Summary

The limitations revealed through ARTeFACT motivated rethinking segmentation from first principles. Rather than developing another domain-specific approach, this chapter tackles the broader challenge of unsupervised semantic segmentation. The key insight is that text-to-image diffusion models learn rich semantic relationships between image regions through their attention mechanisms, without requiring any explicit segmentation training.

This chapter examines the rich information encoded in self-attention maps from diffusion models. These attention maps, which are naturally row-stochastic, can be directly interpreted as transition probabilities in a Markov random walk between image patches. This property enables us to modulate multi-hop relationships through matrix exponentiation, where P^k captures k-step transitions between patches. We then apply Normalised Cuts, which provides a principled approach to graph partitioning by minimizing the ratio of cross-connections to within-cluster as-

sociations, to cluster the resulting features. Our experiments demonstrate that higher exponents strengthen connections between semantically related elements at higher hierarchical levels, enabling the model to group lower-level components (such as ears, paws, eyes) into higher-level semantic entities (an entire cat).

Additionally, we adapt the Minimum Average Node Cut (MAN-C) algorithm to automatically determine the optimal threshold for terminating the recursive partitioning process. This adaptation eliminates the need for manual hyperparameter tuning, making the approach robust across diverse image types with varying complexity and content density.

The combination of enhanced random walk relationships, spectral clustering, and automatic threshold determination achieves state-of-the-art zero-shot segmentation results without any task-specific supervision, demonstrating how attention mechanisms naturally capture meaningful semantic groupings.

These state-of-the-art results in general semantic segmentation present the first step of a particularly promising direction for damage detection. The previous chapter showed how supervised training limited generalisation across different types of media damage: models struggled to move beyond the specific patterns they were trained on. In contrast, as shown by the experiments in this chapter, text-to-image diffusion models learn rich semantic relationships through their multimodal training, naturally grouping visually disjoint but semantically similar entities across different granularities, scales and spatial configuarions.

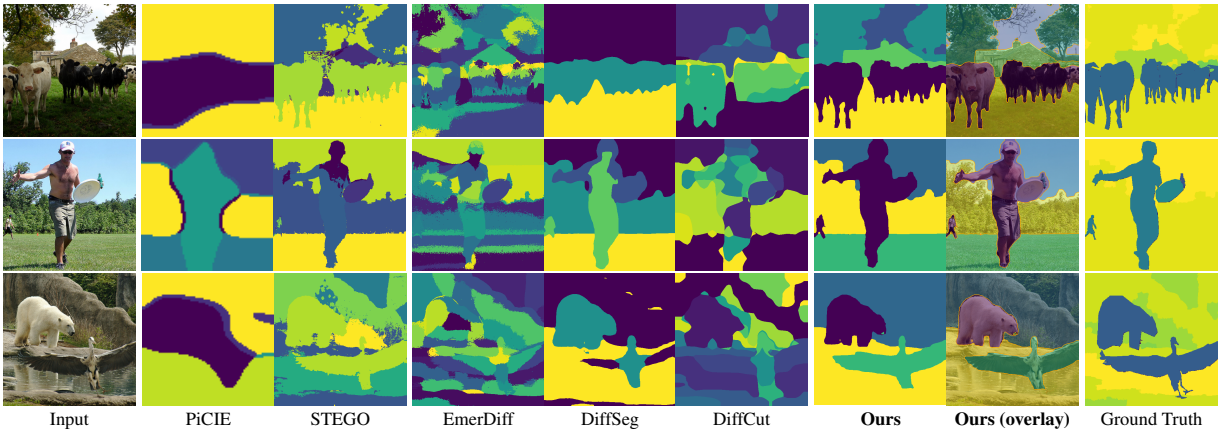


Figure 5.1: We perform unsupervised segmentation by applying Normalised Cuts [181] to self-attention features from Stable Diffusion [167] in a hyperparameter-free setting, and achieve superior performance on Coco-Stuff-27 [27] compared to trained and zero-shot methods.

5.2 Abstract

We propose a zero-shot unsupervised image segmentation method by utilising self-attention activations extracted from Stable Diffusion. We demonstrate that self-attention can directly be interpreted as transition probabilities in a Markov random walk between image patches. This property enables us to modulate multi-hop relationships through matrix exponentiation, which captures k-step transitions between patches. We then construct a graph representation based on self-attention feature similarity and apply Normalised Cuts to cluster them. We quantitatively analyse the effects of incorporating multi-node paths when constructing the NCuts adjacency matrix, showing that higher-order transitions enhance hierarchical relationships in the proposed segmentations. Finally, we describe an approach to automatically determine the NCut threshold criterion, avoiding the need to manually tune it. Our approach surpasses all existing methods for zero-shot unsupervised segmentation based on pre-trained diffusion models features, achieving state-of-the-art results on COCO-Stuff-27, Cityscapes and ADE20K.

5.3 Introduction

Segmentation of images into semantically-meaningful regions is of key interest in computer vision. *Unsupervised* segmentation aims to label images without explicit training to identify specific object types [181, 97, 81]. We focus on *unsupervised multi-class semantic segmentation*, where an image is labelled with many distinct and interpretable labels (“horse, sand, human, sky”) rather than e.g. binary partitioning into foreground/background as in [135, 208].

Unsupervised semantic segmentation is ultimately an ill-posed problem, as there can be no consistent *a priori* definition of what an “object” is [61]. Would the objects in a photograph of a crowd be people, a person, a face, an eye, an iris? Should we label a whole forest or each

tree in it? Yet, it is still possible to infer useful segmentations of images without supervision; in particular, inferring labellings which align well with human judgments of what coherent objects are. This implies, for example, that we want segmentations that have sharp boundaries, spatial coherence, consistency over similar images, and a consistent semantic level of segmentation within an image.

Modern approaches to unsupervised segmentation leverage the information inherently captured in the latent features of models trained for other tasks [31, 81, 194, 197, 52]. A particularly rich source of features is the self-attention maps of text-to-image generative models like Stable Diffusion [167]. These maps encode how much each (latent) pixel attends to every other pixel, and form a natural basis for partitioning an image. Because the self-supervised training process needs to parsimoniously account for e.g. spatial equivariance of rigid objects or deformation of flexible objects like human bodies, the latent features would be expected to give rise to similar embeddings for pixels across objects that are consistent under these types of transformations.

Leveraging the effectiveness of self-attention features from Stable Diffusion 1.4, as demonstrated in DiffSeg [197], we propose a novel unsupervised segmentation approach. Departing from DiffSeg's bottom-up, iterative merging strategy, we employ the classic Normalised Cuts algorithm [181] to iteratively partition image regions based on self-attention similarity. To address the challenge of manual tuning in Normalised Cuts, we propose a method that automatically determines the stopping criterion. This yields a tuning-free method that automatically estimates optimal thresholds for halting the iterative partitioning, surpassing the performance of existing zero-shot unsupervised segmentation techniques.

Furthermore, we interpret the self-attention maps as transition kernels for a random walk across the image space, effectively grouping latent pixels with high connectivity probability. By exponentiating these kernels, we enable random walks to take larger steps, providing a natural and controllable mechanism for selecting the desired semantic granularity of the segmentation. This mathematical interpretation not only provides theoretical grounding for our approach but also enables intuitive control over segmentation detail.

Our contributions are:

- A zero-shot unsupervised semantic segmentation algorithm that outperforms the state-of-the-art, including DiffSeg [197], EmerDiff [144], and DiffCut [52];
- A random walk interpretation of self-attention features used to build the Normalised Cuts adjacency matrix, which can modulate semantic granularity by simple matrix exponentiation;
- A novel approach to *automatically* selecting the critical threshold for Normalised Cuts on a per-image basis that eliminates the main "fiddle factor" in unsupervised segmentation.

5.4 Related Work

5.4.1 Unsupervised Segmentation via Generative Models

Internal features of GANs [75], VAEs [107] and other self-supervised models such as DINO [4, 31], have long been used for diverse downstream image tasks, achieving impressive results [46, 115, 201, 81, 200, 116, 37]. Recently, large-scale generative text-to-image models (e.g. Stable Diffusion [167]), have provided even richer internal features to be used for downstream tasks including depth estimation [41], semantic correspondence [129, 194, 233, 85], and segmentation [13, 215]. Unsupervised zero-shot segmentation is one example where methods leveraging diffusion model features have achieved state-of-the-art results [144, 197, 52].

EmerDiff [144] demonstrates that cross-attention query vectors encode semantic information at each spatial location. They first perform k-means clustering on these vectors at low resolution to obtain initial segments. To achieve high-resolution segmentation, they perturb the cross-attention activations corresponding to each segment, re-generate, and observe which pixels in the output are most affected. Although this method produces accurate segmentation boundaries, it requires two denoising passes through the diffusion model *per segment*. Additionally, the approach requires prior knowledge of the number of segments, meaning it is not truly unsupervised.

DiffSeg [197] instead uses self-attention maps from Stable Diffusion [167]. For an input image yielding N latent patches, they extract an N×N self-attention matrix where each row represents the attention distribution from one patch over all other patches. Inspired by SAM [108], they sample "anchor" points on a grid and iteratively merge regions based on the Kullback-Leibler divergence between their averaged attention distributions. This bottom-up approach does not require the number of segments to be provided *a priori*. However, it is computationally expensive due to both the grid sampling process and KL divergence calculations between large matrices; the latter must be calculated twice between each pair of patches/segments as it is not symmetric. Furthermore, the grid sampling strategy may miss small but significant image features if the anchors are not sampled densely enough, and the iterative merging requires pre-setting a threshold over KLD as well the number of iterative merging passes.

DiffCut [52], inspired by DiffSeg, extracts features (rather than raw self-attentions) from a Stable Diffusion XL model, and employs Normalised Cuts [181] iteratively with a manually tuned stopping criterion, and an exponentiation factor to increase the contrast of the adjacency matrix for easier partitioning. This follows from a line of recent works which has revitalized the use of spectral clustering methods by leveraging graphs constructed from neural network features. Deep Spectral Methods [135], for instance, utilize features extracted from a pre-trained self-supervised vision transformer (DINO) for salient object localization. Similarly, TokenCut [208] and MaskCut [207] employ Normalised Cuts on DINO features, with MaskCut iteratively partitioning the image into a pre-defined number of segments. More recently, Panoptic

Cut [101] extends this approach by applying Normalised Cuts for a fixed number of iterations on DINO features, generating segment proposals for open vocabulary segmentation. Notably, [238] first explored the effectiveness of spectral clustering over features from generative diffusion models, demonstrating they encode rich semantic and spatial information.

A common limitation across all the aforementioned segmentation methods is their reliance on either a single bipartition for salient object detection (Deep Spectral Methods, TokenCut) or iterative partitioning with manually specified parameters like the number of clusters (MaskCut), partitioning iterations (PanopticCut) or stopping criteria (DiffCut). Instead, our approach extracts and aggregates self-attention activations from Stable Diffusion 1.4 as in DiffSeg, to which we apply Normalised Cuts, concurrently to DiffCut. In addition, we propose a way to automatically halt the partitioning of each image, eliminating the need for manual tuning of NCut parameters, such as number of iterations or stopping threshold. Finally, we observe a special property of self-attention activations in particular, which allows us to control the partitioning for semantic granularity. Next, we review the theory behind the classic Normalised Cuts algorithm which informed our approach.

5.4.2 Normalised cuts

Spectral graph theory traces its roots to the 1970s [59, 64], but was first famously applied to computer vision by Shi and Malik [181] with the advent of Normalised Cuts.

Here, an image is represented as a weighted, undirected graph $G = (V, E, w)$ where pixels correspond to vertices, and edge weights represent some similarity measure between each pair of pixels. The graph can be partitioned, or cut, into two disjoint sets A and B, where the cost of cutting it is the sum of all edge weights that are removed:

$$\text{cut}(A, B) = \sum_{u \in A, v \in B} w(u, v). \tag{5.1}$$

The normalised cut (NCut) instead normalizes this cost relative to the total connections in each part of the graph:

$$\text{NCut}(A, B) = \frac{\text{cut}(A, B)}{\text{assoc}(A, V)} + \frac{\text{cut}(A, B)}{\text{assoc}(B, V)} \tag{5.2}$$

Here $\text{assoc}(A, V)$ measures A's total connections to all nodes. This normalization prevents the algorithm from isolating individual vertices, since cutting off a small set would result in a high NCut value. Instead the minimum-cost cut tends to split the image into two large but dissimilar regions.

The authors show that the normalised cut formulation can be solved through a generalized eigenvalue problem. Given an adjacency matrix A (representing the weighted graph G) containing edge weights between nodes and its corresponding degree matrix D (where each diagonal

entry is the sum of that row in A), the NCut optimisation is equivalent to solving:

$$Lx = \lambda Dx, \tag{5.3}$$

where L is the Laplacian $L = D - A$. The solution is found in the eigenvector x_2 corresponding to the second-smallest eigenvalue λ_2, also known as the Fiedler vector, which is used to describe a graph's algebraic connectivity [64]. This eigenvector assigns a value to each pixel; the algorithm finds a threshold in these values that minimizes the Normalised Cut, bipartitioning the image. Since the Fiedler vector takes on continuous values, the authors recommend testing several thresholds, and selecting the one with the smallest NCut cost. They also extend the algorithm to give more than two regions, by iteratively partitioning each resulting subgraphs until some condition, such as a target number of clusters, or a threshold NCut cost, is met.

Dynamic NCut threshold. A further extension to the NCuts algorithm is the Minimum Average Node Cut (MAN-C) [205], which aims to overcome the need for manually setting the number of clusters, iterations, or threshold NCut cost. It defines a "good" cluster based on the condition min-cut$(G) \geq \frac{nT}{2m}$, where n is the number of vertices, T is the total edge weight, and m is the number of edges. In MAN-C, the *minimum cut* for a partition is the smallest sum of edge weights needed to separate the graph into two parts. For each proposed Normalised Cuts partition, MAN-C checks if this minimum cut is lower than the threshold $\frac{nT}{2m}$ to determine when to halt the recursive partitioning.

Random walk interpretation. Meila and Shi [134] later gave a probabilistic interpretation of NCut by showing that spectral clustering can be viewed through the lens of random walks on graphs. They demonstrated that the degree-normalised adjacency matrix $P = D^{-1}A$ represents transition probabilities of a random walk, where $P_{i,j}$ represents the probability of moving from node i to node j. The eigenvectors of this transition matrix correspond exactly to the NCut solution, with the second largest eigenvector providing the optimal binary partition. This interpretation reveals that spectral clustering effectively groups nodes that have high probability of being connected by random walks and suggests that any matrix with random walk properties (row-stochastic, capturing similarity relationships) can be directly used in the NCut framework.

5.5 Methods

The success of DiffSeg [197] and EmerDiff [194] demonstrates that pre-trained text-to-image diffusion models encode rich semantic relationships in their internal representations. Our approach recognizes and exploits a fundamental mathematical property of these models' self-attention matrices. In this section, we present our method in five parts: first, we describe how we extract self-attention features from Stable Diffusion (Section 5.5.1); next, we explore two complementary approaches to leveraging these features—directly interpreting self-attention as

random walks (Section 5.5.2) and constructing adjacency matrices from attention feature similarity (Section 5.5.3); we then detail our application of Normalized Cuts to these representations (Section 5.5.4); propose an automatic threshold determination method (Section 5.5.5); and finally describe our upsampling process (Section 5.5.6).

5.5.1 Extracting diffusion features

Given the pre-trained diffusion model, we use DDIM Inversion [184, 58] to invert real images for a small number of steps (we use 10 out of 50 in our experiments, and ablate this in the Supplementary, Section 5.9.2), then extract self-attention from the denoising U-Net decoder from a single generation step. For Stable Diffusion, transformer activations exist in four resolutions across the decoder: 8^2, 16^2, 32^2 and 64^2. We extract self-attention maps $S^{(r)} \in \mathbb{R}^{N \times N}$ where $r \in \{0,1,2,3\}$ is the resolution level and N corresponds to the spatially flattened latent feature resolution 8^2, 16^2, 32^2 and 64^2. We sum across attention heads following [197, 238]. We aggregate over resolutions by upsampling to $64^2 \times 64^2$ followed by a proportional weighted sum, empirically defined in DiffSeg [197], producing an aggregated self-attention matrix $P = \sum_r w_r \cdot S^{(r)}$. Each matrix $S^(r)$ is row-normalized through softmax, and since the weights w_r sum to one, the aggregated matrix P maintains row-normalization.

5.5.2 Self-attention defines a random walk

While traditional NCut requires constructing an adjacency matrix A, from which the random-walk interpretation derives a degree-normalized (i.e. row-stochastic) transition matrix $P = D^{-1}A$, we observe that the aggregated self-attention P already satisfies the mathematical properties of a random walk transition matrix, albeit without being derived explicitly from some adjacency matrix A.

As described in the previous section, our aggregated self-attention matrix maintains row-normalization through the softmax operation in the attention mechanism and weighted aggregation. Consequently, each row in P represents a probability distribution, with elements $P_{i,j}$ giving the probability of transitioning from patch i to patch j in a random walk across the image.

It is therefore possible to directly apply Normalized Cuts theory in the Random Walk reframing. We use the transition matrix as the input to recursive NCut as proposed by Shi and Malik [181], as described in Section 5.5.4, by adapting the algorithm to use the eigenvectors of the transition probabilities P (c.f. [134]) instead of an adjacency matrix A. Since P is not necessarily symmetric, to estimate the second eigenvector, we apply power iteration with one deflation.

This formulation provides a principled approach to leveraging self-attention directly for segmentation, without requiring the construction of an intermediate adjacency matrix or similarity metric, while still benefiting from the theoretical guarantees of spectral clustering.

5.5.3 Adjacency matrices from feature similarity

Beyond directly using self-attention as transition probabilities, we can construct traditional adjacency matrices by treating each patch's attention pattern as a feature embedding. While DiffSeg uses Kullback-Leibler divergence between attention distributions, we employ either dot product or cosine similarity (empirically contrasted in Section 5.6.2), which can be computed more efficiently.

These constructions yield a positive-valued, symmetric adjacency matrix A suitable for the conventional NCut formulation. This allows for efficient direct eigendecomposition rather than power iteration.

At the same time, we can exploit the random-walk property of the self-attention matrix established in Section 5.5.2 prior to constructing the adjacency matrix. By computing P^k through matrix exponentiation, each entry $(P^k)_{ij}$ represents the probability of transitioning from patch i to j in exactly k steps [148]. This k-step transition probability aggregates all possible walks of length k between patches i and j, capturing information flow through intermediate patches.

The exponent k serves as a scale parameter: smaller values preserve local attention patterns, while larger values enable broader information diffusion across the image, connecting semantically related but spatially distant patches (see first row of Figure 5.4). We evaluate several values of k in Section 5.6.4, showing how this parameter effectively controls the hierarchical level of segmentation.

5.5.4 Applying normalised cuts (NCut)

So far, we have introduced two approaches to quantify semantic correlations across the image by (1) modeling the aggregated self-attention matrix as a k-step random walk among patches via matrix exponentiation, and (2) constructing an adjacency matrix from patch-level feature similarities. As discussed in Section 5.4.2, both introduced formulations allow us to apply Shi and Malik's recursive Normalised Cuts (NCut) criterion [181] to partition the image. Regardless of whether a random walk or adjacency matrix is used, the algorithm proceeds as follows. (1) Solve the generalised eigenvalue problem (equation 5.3). (2) Threshold the Fiedler vector across test values, selecting the bipartition with lowest NCut value. (3) If the NCut cost exceeds a predefined threshold, return the current partitions. (4) Otherwise, recursively apply the algorithm to each newly formed partition.

5.5.5 Adapting MAN-C

A key challenge in Normalized Cuts is determining when to stop recursive partitioning. Manual threshold selection is burdensome and image-dependent. To address this, we adapt the Minimum Average Node Cut (MAN-C) algorithm [205], detailed in Section 5.4. MAN-C works

well for sparse graphs, where there are few edges. However, our graphs derived from self-attention similarities have very high edge density, and in this case the MAN-C threshold reduces to $\frac{T}{n-1}$ i.e. it is scaled by the number of nodes—therefore, a minimum cut will almost always be larger, since it is not appropriately scaled w.r.t. the threshold anymore.

Hence, we propose and evaluate (Section 5.6.3) two alternative adjustments:

1. Directly compare NCut cost to the MAN-C threshold;
2. Compare to a *scaled* minimum cut, $\frac{\text{min-cut}(G)}{T}$, which scales the minimum cut relative to the graph's total edge weight T, to place it in the same scale as the MAN-C threshold.

Both approaches are designed to keep the adaptive nature of MAN-C while handling the dense nature of our similarity graphs.

5.5.6 Segmentation aggregation and upsampling

After recursive partitioning, we aggregate the NCut partitions into unique segments and calculate the average feature vector for each region, resulting in K averaged PDFs for K segments of shape $N \times N$, where N is the latent size (number of latent patches). We upsample these segmentations to the original image resolution by first applying bilinear interpolation to the original features—$512 \times 512 \times N^2$, then computing cosine similarity between each upsampled pixel feature and each segment's average distribution. Each pixel is assigned to its closest segment, producing the final segmentation map.

5.6 Experiments

Datasets. For our main experiments, we follow [42, 97, 197, 144] and use a subset [97] of COCO-Stuff-27 [27] validation data. It consists of 2174 images, combining 80 thing and 91 stuff categories from COCO-Stuff-171 into 27 classes. Our final approach is also evaluated on Cityscapes [49], following DiffSeg [197]. We additionally evaluate on ADE20K [240], and use it for our comparison with DiffCut [52].

Metrics. To quantitatively evaluate unsupervised segmentation, we report Mean Pixel Accuracy, Mean Intersection over Union (mIoU), as done in previous works [197, 144, 81], in addition to F1 score. Since class labels are not predicted, we use Hungarian matching to assign predicted segments to their closest Ground Truth segments, following standard practice [197, 144]. This penalizes oversegmentation, since when there are more predicted segments than Ground Truth ones, the remaining predicted segments are considered false positives. We note that due to the varying granularity of human annotated segmentation labels, this standard evaluation strategy is far from perfect: for example, in Figure 5.2, row 1, the Ground Truth has an annotation for the baseball glove as a separate object, while the human figures are considered

as one segment; however, our method either merges the glove with the humans segment with a lower NCut threshold (row 1, column 2), or separates the baseball glove, along with the hats, shoes, trousers, shirts, legs and arms into their own segments when the threshold is increased.

5.6.1 Random Walk Baseline

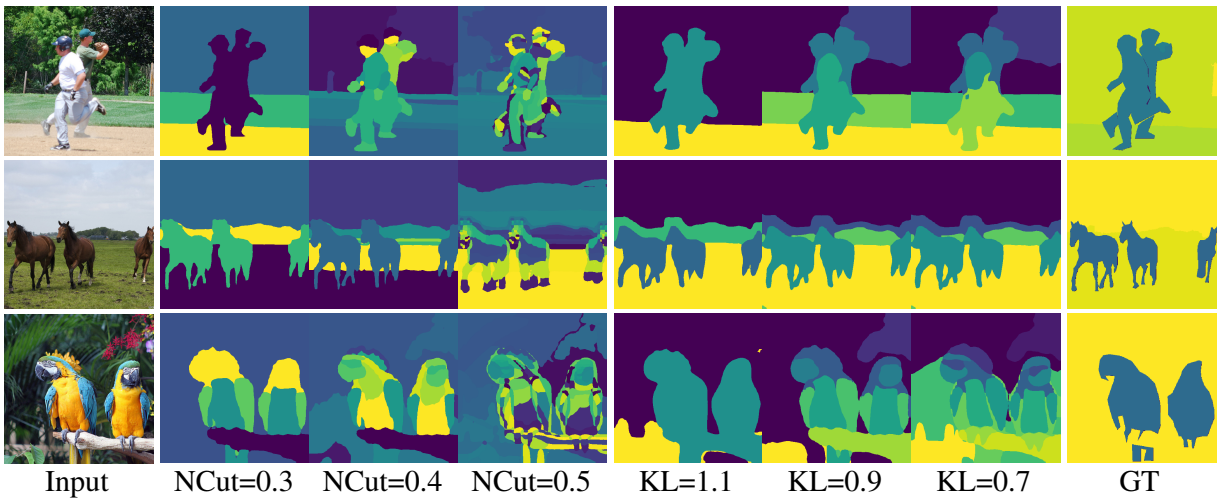
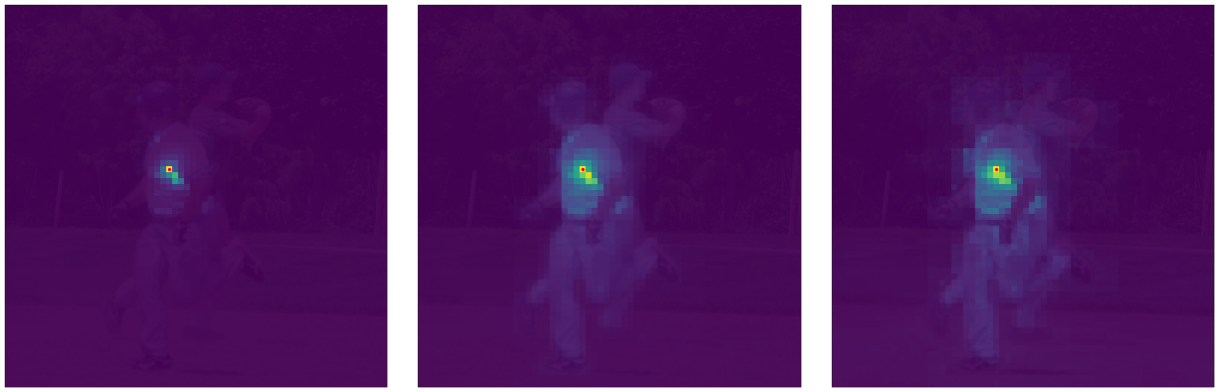


Figure 5.2: Qualitative comparison on COCO-Stuff-27 between our simple Random Walk approach for different NCut cost thresholds (columns 2-4) and DiffSeg for different Kullback-Leibler Divergence thresholds (columns 5-7).

As outlined in Section 5.5.2, our simple baseline approach uses the aggregated self-attention matrices directly as a random walk transition matrix and applies recursive Normalized Cuts over it. We show examples from several thresholds over the NCut cost as a stopping criterion in Figure 5.2, where we also compare it to results from DiffSeg for their best-perfoming KLD threshold (1.1, column 5) along with two lower values. Since Normalised Cuts is inherently a top-down segmentation approach, we note how a higher NCut threshold increases granularity in the predicted segments, which generally follow a hierarchical structure; for DiffSeg, lower KLD threshold has a similar effect, as the approach merges segments in a bottom-up fashion. Overall, our approach's predictions are more precise, and demonstrate a clear hierarchical semantic relationship between segments across thresholds, even in the case of extreme oversegmention—for instance, the horses legs, hooves, backs and faces are all grouped together in row 2, column 4. DiffSeg predictions, on the other hand, suffer from pronounced horizontal banding artefacts (row 2, 5). Quantitative results for NCut threshold set to 0.3 in Table 5.4 show that this simple approach already performs almost as well as the state-of-the-art DiffSeg, particularly with respect to mIoU and accuracy.



(a) Self-attention PDF (b) Dot product adjacency (c) Cosine similarity adjacency

Figure 5.3: Self-attention PDF (aggregated) (a) and corresponding dot product (b) and cosine similarity (c) adjacency values for a single patch.

5.6.2 Self-Attention adjacency

The previous experiment motivates a natural follow-up: how can we best construct a traditional NCuts adjacency matrix based on the similarity between self-attention activations? Here, we validate the two variants proposed in Section 5.5.3: constructing A as the dot-product between the self-attention PDFs, or as their cosine similarity (normalised dot product). Both variants result in a positive-valued symmetric adjacency matrix A, since self-attention is already probability-normalised. We compare the results of applying Normalised Cuts over each variant across three different NCut Thresholds in Table 5.1, where the Dot Product variant is the clear leader—note that when the NCut threshold is set to 0.4, it already outperforms state-of-the-art DiffSeg in terms of both mIoU and Accuracy.

Model Variant	NCut Threshold	Acc	F1	mIoU
Dot Product	0.4	74.0	58.4	44.0
	0.5	71.6	56.7	43.1
	0.6	60.0	45.2	32.1
Cosine Similarity	0.4	72.4	56.0	41.7
	0.5	71.5	56.0	42.2
	0.6	62.9	47.7	34.4

Table 5.1: Quantitative results on COCO-Stuff-27 between Normalised Cuts performed on adjacency matrices build as dot product vs cosine distance between patch self-attention across three different NCut thresholds.

Examining the spatial distribution plots of self-attention compared to the two adjacency variants for a given patch marked by red dot in Figure 5.3, we can see the intuition behind why dot product performs better. For a given query position q, the self-attention produces a PDF p_q that is sharply peaked at q and its immediate neighborhood $N(q)$, with near-zero values elsewhere, as shown in (a). The dot product $p_q^T p_r$ between two such distributions preserves this sparsity structure, showing high similarity only when neighborhoods overlap in (b). In contrast, cosine

similarity $\frac{p_q^T p_r}{||p_q||||p_r||}$ normalizes away this locality-preserving property, artificially inflating similarities between distributions with disjoint support, as evidenced by more distributed similarities in (c), which form a halo around the object of interest.

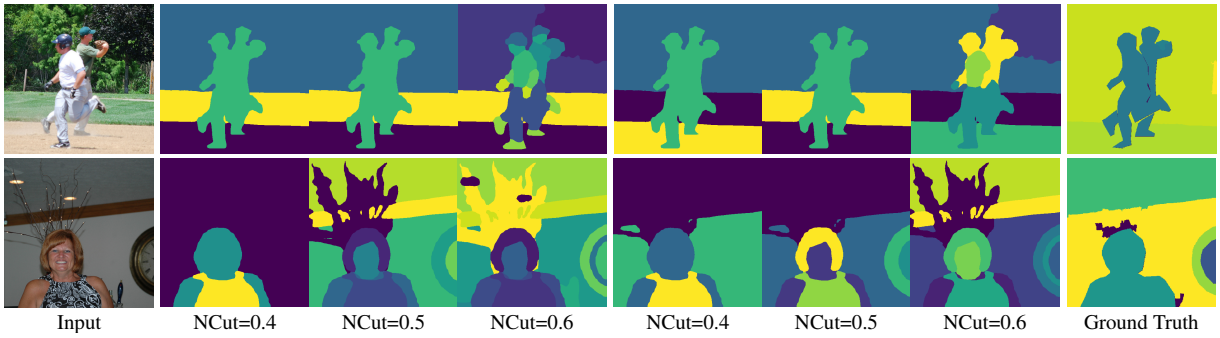


Figure 5.4: Qualitative comparison on COCO-Stuff-27 between NCut over a dot-product Adjacency matrix (columns 2-4) and over a cosine similarity Adjacency matrix (columns 5-7) across a range of NCut thresholds.

Hence, more uniformly distributed adjacency values would have a negative effect on Normalised Cuts, as it relies on having strong intra-segment connections and weak inter-segment connections; cosine similarity would therefore make the distinction between segments less pronounced and lead to arbitrary cuts, as seen in Figure 5.4.

5.6.3 Hyperparameter-free NCut

In Section 5.5.5 we proposed a way to perform Normalised cuts without setting an NCut cost as a stopping criterion, making the approach completely hyperparameter-free.

Specifically, we suggested two versions of the MAN-C algorithm with thresholds adapted to the specific graph structure that arises from our adjacency matrices. We compare these two automatic thresholding approaches to assess their effectiveness in partitioning the self-attention adjacency graphs, using the dot product construction from the previous experiments. While the quantitative results in Table 5.2 demonstrate that comparing the MAN-C criterion with NCuts is much more effective, we can see in the qualitative examples in Figure 5.5 that the segmentations proposed by thresholding with the scaled MinCut are not necessarily bad, as semantically meaningful groupings are clearly emerging; rather, they are too granular, and punished by the Hungarian matching due to the coarse annotations, as discussed earlier. Additionally, since the approach is now free of hyperparameters, we follow up with an ablation study on the self-attention resolution levels $8, 16, 32, 64$ by progressively removing each, starting from the lowest resolution. This allows us to evaluate how different self-attention scales contribute to segmentation quality. Removing lower resolutions increases the granularity of predictions for both methods, as seen in Figure 5.5, and decreases the metrics they achieve; yet the qualitative examples demonstrate preserved semantic coherence even fore the most granular predictions, such as those made with the Scaled MinCut threshold. Overall, our quantitative results show that we

outperform DiffSeg, which optimises several dataset-specific hyperparameters (merging level, number of anchors, and KLD merging threshold) in a completely hyperparameter-free setting.

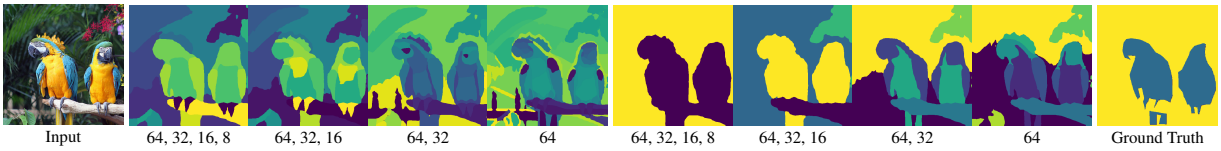


Figure 5.5: Qualitative comparison of segmentation results for our two automatic thresholding approaches: Scaled MinCut (columns 2-5) and NCut (columns 6-9) across different self-attention resolution levels.

Approach	Resolution Levels				Acc	F1	mIoU
	8	16	32	64			
Random Walk	✓	✓	✓	✓	72.2	56.9	43.3
MAN-C w/ NCut	✓	✓	✓	✓	74.7	60.6	<u>46.1</u>
		✓	✓	✓	74.1	60.8	**46.6**
			✓	✓	70.1	56.4	42.7
				✓	63.9	49.6	36.4
MAN-C w/ Scaled MinCut	✓	✓	✓	✓	45.9	36.1	23.8
		✓	✓	✓	42.0	32.0	20.6
			✓	✓	37.2	27.6	17.0
				✓	36.1	25.0	15.3
DiffSeg	✓	✓	✓	✓	72.5	58.5	43.6
DiffCut			✓		59.6	44.5	31.3

Table 5.2: Comparison of our automatic threshold methods across different attention resolution combinations on Coco-Stuff-27. MAN-C NCut consistently outperforms both Scaled MinCut and state-of-the-art DiffSeg. Random Walk baseline has NCut threshold set to 0.3.

5.6.4 Power Walks

While our previous experiment demonstrates a fully automated approach for zero-shot unsupervised segmentation, removing the need for pre-setting the NCut threshold means we do not have a way to control the granularity of segmentations. As established in Section 5.5.3, self-attention matrices can be interpreted as random walk transition probabilities. This allows us to modulate multi-hop relationships through matrix exponentiation prior to constructing our adjacency matrix. Our quantitative results in Table 5.3 show that considering higher order walks ($k = 2,3,4$) substantially improves the performance of the Scaled MinCut threshold approach against the annotations of COCO-Stuff-27. Qualitatively (Figure 5.6), we see that higher exponents have the effect of hierarchically merging the more granular segments produced at $k = 1$:

the heads, wings, feet and bodies of the parrots are merged into a larger "parrot" segment, which is closer to the human-annotated labels provided in the ground truth.

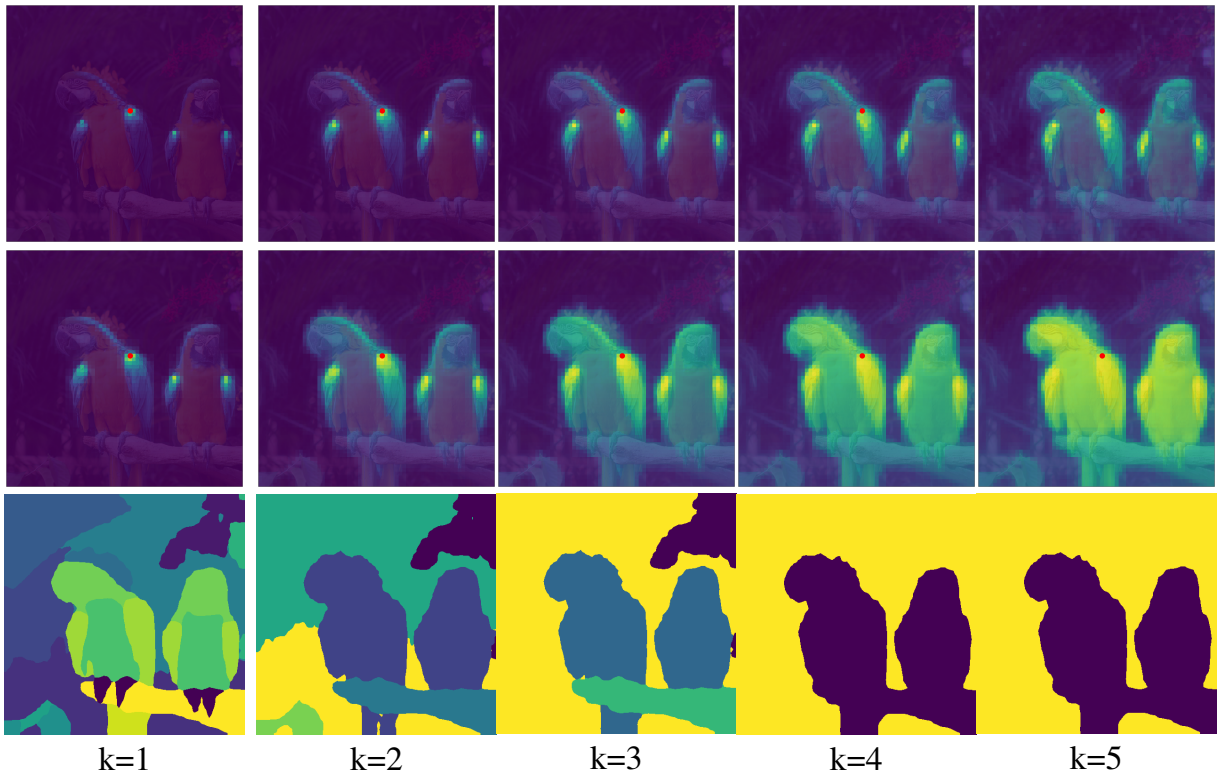


Figure 5.6: Power Walks: Self-attention extracted from Stable Diffusion for an arbitrary patch marked by red dot (row 1) framed as a random walk transition probability matrix and raised to the k^{th} power, allows us to build adjacency matrices which increasingly capture higher-order patch-similarities (row 2). The resulting unsupervised segmentations (row 3) produced by our hyperparameter-free NCut method demonstrate that as long-range relationships are enhanced, the granularity of segmentation decreases, while still extracting semantically-meaningful segments.

steps k	1	2	3	4	5
Acc	45.9	67.0	71.2	70.7	70.0
F1	36.1	54.4	56.9	55.8	54.2
mIoU	23.8	40.7	42.4	41.1	39.6

Table 5.3: Comparison for a range exponents (random walk steps k) applied to the self-attention transition matrix P before constructing the adjacency matrix A. We employ Scaled MinCut for automatic threshold selection, considering all available resolutions.

5.6.5 Comparison with state-of-the-art

In Table 5.4 we summarise the results of our method alongside several recent approaches to unsupervised segmentation—DiffSeg [197], DiffCut [52] and EmerDiff [144], which are the closest to it conceptually, being zero-shot unsupervised segmentation models which rely on Stable Diffusion features and represent state-of-the-art, and also PiCIE [42] and STEGO [81].

	Trained		Zero-shot			
	PiCIE	**STEGO**	**EmerDiff**	**DiffSeg**	**DiffCut**	**Ours**
COCO-Stuff	13.8	28.2	26.6	<u>43.6</u>	31.3	**46.6**
Cityscapes	12.3	21.0	12.0	<u>21.2</u>	18.1	**22.0**
ADE20K	-	-	-	37.7	<u>40.1</u>	**41.2**

Table 5.4: Quantitative comparison of mIoU between our approach and current state-of-the-art on COCO-Stuff-27, Cityscapes and ADE20K. Best in bold, second best underscored.

For DiffSeg, we use their suggested best hyperparameters, i.e. 1.1 and 0.9 KLD threshold for COCO-Stuff and Cityscapes respectively, and report additional ADE20K results from [52]. For EmerDiff, we evaluate with Hungarian matching for a fair comparison against the other methods, and set the mask proposal parameter $K = 30$, as the authors suggest. While both DiffSeg and EmerDiff rely on Stable Diffusion 1.4 as a feature extractor, DiffCut report their main results on SSD 1B [79], a distilled variant of Stable Diffusion XL[159], and utilise a post-processing step to refine segmentation predictions—PAMR [7]. For fairness, we exclude PAMR, and evaluate the method on Stable Diffusion 1.4 using hyperparameters $\tau = 0.2$ and $alpha = 10$, optimised for SD1.4 by the authors in their ablation. We include the best-performing variant of our method, which uses three self-attention resolutions, and moreover does not rely on any hyperparameters as it leverages our dynamic NCut thresholding scheme; this variant improves on the state-of-the-art DiffSeg by 3.0% mIoU on COCO-Stuff and 0.8% on Cityscapes. In fact, some of the weaker variants of our approach still outperform DiffSeg, see Tables 5.1 & 5.2. Furthermore, our dynamic threshold approach outperforms DiffCut, despite it using NCut hyperparameters specifically optimised for Stable Diffusion 1.4.

Model	**Method**	**w/o PAMR**	**w/ PAMR**
SSD 1B	DiffCut	46.1	49.1
	Ours	41.3	42.6
SD 1.4	DiffCut	31.3	45.2
	Ours	46.6	47.1

Table 5.5: Comparison with DiffCut (mIoU) on COCO-Stuff-27.

We perform further ablation against DiffCut Table 5.5. For fair comparison, we benchmark our approach against DiffCut using both SD 1.4 and SSD 1B. We also ablate the effect PAMR postprocessing by adding it to our pipeline as well. We use the best hyperparameters suggested by the authors, i.e. $\alpha = 10$ and $\tau = 0.5$ for SSD 1B or $\tau = 0.2$ for SD 1.4.

With SD 1.4, our method achieves 46.6 mIoU on Coco-Stuff-27 without PAMR post-processing outperforming DiffCut's 45.2 mIoU *with* PAMR and 31.3 without. In fact, without PAMR, our SD 1.4 results (46.6 mIoU) outperform DiffCut even with the newer, larger SSD 1B model.

Unlike SD 1.4, SSD 1B lacks self-attention at 8×8 and 16×16 resolutions, and has an

asymmetric encoder-decoder structure with additional decoder transformer blocks (34 in total), hence the DiffSeg feature aggregation strategy does not translate to this setting. While our initial results with this backbone show promise (Table 5.5), optimization of feature aggregation strategies for SSD 1B might improve performance.

Qualitatively, both our coarser and finer predictions (Figure 5.1) are more precise than those of DiffSeg and DiffCut; additional visual comparisons are given in the Supplementary 5.9.3.

5.7 Conclusions

Unsupervised image segmentation is a challenging and ambiguous task where even the definition of "correct" remains debatable [61]. We demonstrate that we can exploit the naturally emerging patterns in diffusion self-attentions to discover meaningful segmentations, by exploring the activations through the lens of random walks. By framing self-attention as transition probabilities in a Markov process, we establish a principled approach for clustering image regions based on their semantic relationships, while offering intuitive control over segmentation granularity through walk length. Our approach combines the strengths of both top-down spectral clustering methods (like DiffCut) and bottom-up merging strategies (like DiffSeg) while addressing their limitations: we eliminate the need for manual threshold tuning in recursive partitioning through our automatic threshold determination, while avoiding the grid sampling and iterative merging inefficiencies of bottom-up approaches through our power walk exponentiation, which naturally handles the "merging" of semantically related regions. This balanced methodology produces hierarchically consistent segmentations that align with human perceptual judgments across diverse image types.

Our approach reveals intrinsic object groupings in learned representations, specifically self-attention layers, while achieving superior quantitative results. Though implemented for Stable Diffusion 1.4, our methodology is model agnostic and applicable to any model with attention features, as demonstrated with SSD 1B in our ablation against DiffCut.

Variants that most closely matched human annotation granularity naturally produced the best quantitative results. However, our qualitative analysis reveals that even lower-scoring variants maintain consistent semantic groupings. Future work should investigate the discrepancies between human annotations and unsupervised segmentations from large models—potentially revealing insights about human versus machine perception patterns.

5.8 Afterword

The work in this chapter demonstrates how classical spectral methods can leverage the rich representations learned by T2I generative diffusion to achieve state-of-the-art unsupervised segmentation. It provides empirical evidence that semantic understanding—specifically, the ability to group visually distinct entities based on their semantic relationships—can emerge naturally through large-scale multimodal generative pretraining, without explicit task-specific supervision.

The work in this chapter solves a general semantic segmentation problem and is thus quantitatively evaluated on standard benchmarks, such as Coco-Stuff-27 [27] and ADE20K [240]. In the Supplementary, Section 5.9.3, additional qualitative results on the ARTeFACT dataset from Chapter 4 are provided. While damage artefacts are correctly segmented, determining which segement is the damage, e.g., *damage detection* is a task which must be solved next.

For damage detection, the success in unsupervised segmentation presents both opportunities and challenges. While the framework effectively partitions images into semantically coherent regions without supervision, the critical remaining task involves semantic disambiguation—identifying which segments correspond to damage. This raises an important question about the representational capacity of foundation models: *can the same multimodal architectures that enable semantic clustering also support semantic labeling?*

More precisely, given a model capable of distinguishing between film scratches and cat whiskers through unsupervised clustering, can we leverage the alignment between visual and textual modalities to assign semantic labels to these clusters? This represents a challenging cross-modal grounding problem. The interaction between text conditioning, self- and cross-attention, and emergent semantic grouping could potentially provide a principled approach to damage identification that circumvents the domain generalisation limitations inherent in supervised methods demonstrated in Chapter 4.

These open questions establish promising research directions, explored in the final Chapter 6, Section 6.3. Understanding mechanisms for querying and extracting damage-specific semantic knowledge from foundation models could lead to detection systems that generalze robustly across different damage manifestations.

5.9 Supplementary

5.9.1 A note on evaluation strategies

Zero-shot unsupervised segmentation presents unique evaluation challenges compared to traditional supervised approaches. This section discusses three complementary evaluation strategies that better capture different aspects of zero-shot performance.

Traditional Global Averaging vs. Per-Image Evaluation While previous non-zero-shot methods typically use global averaging (accumulating statistics across all images before calculating metrics) [42, 81], this may not be ideal for zero-shot models. Global averaging emphasizes performance on frequent classes, but zero-shot models have no prior knowledge of class distributions. Instead, per-image averaging, which computes metrics independently for each image before averaging, treats each instance of class discovery equally—better reflecting the zero-shot nature of the task.

The Challenge of Granularity A key challenge in evaluating zero-shot models is their tendency to discover more fine-grained segmentations than present in the ground truth. For example, in Figure 5.7, row 8, where the ground truth groups a bear and bird into a single "animal" class, models trained on the CocoStuff dataset e.g. STEGO [81] also group them together; in contrast, our model distinguishes them as separate classes. While this more detailed segmentation may be semantically meaningful, it is penalized by both global and per-image metrics.

Oracle-Merged Evaluation To address this, we introduce an oracle-merged evaluation strategy that uses ground truth to merge oversegmented areas based on their primary class overlap. This is conceptually similar to EmerDiff's approach [144], though they merge regions based on embedding similarity. However, EmerDiff requires pre-specifying the number of clusters (K=30), which leads to extreme oversegmentation as shown in Figure 5.9. While this hurts their performance under standard metrics, they benefit significantly from merged evaluation strategies, as discussed below.

Our extended evaluation framework combines all three approaches. As shown in Table 5.6, our method outperforms other diffusion-based zero-shot baselines across all metrics. EmerDiff achieves higher accuracy only in the oracle-merged setting, likely due to their cross-attention modulating upsampling approach. The qualitative examples in Figure 5.9 demonstrate why our approach still outperforms EmerDiff in the more strict F1 and mIoU metrics—despite benefiting from a more sophisticated upsampling strategy, EmerDiff produces in noisy clusters. Furthermore, even our "oversegmented" versions produce high-quality, semantically meaningful segments, without the need to specify the number of clusters or any other hyperparameter a priori, unlike EmerDiff and DiffSeg [197].

Importantly, this extended evaluation framework further underscores that metrics don't necessarily measure segmentation "correctness" (which is inherently subjective), but rather alignment with human-labeled ground truths—which can vary in granularity across datasets.

	Globally averaged (traditional)			Per-image averaged			Merged and per-image averaged		
	Acc	F1	mIoU	Acc	F1	mIoU	Acc	F1	mIoU
EmerDiff	28.8	22.8	13.3	28.9	47.8	0.1	92.5	34.9	24.0
DiffSeg	72.5	58.5	43.6	72.6	71.1	34.1	79.2	73.0	60.6
Ours	74.1	60.8	46.6	74.0	73.3	38.0	83.6	76.0	64.5

Table 5.6: We calculate metrics on predictions from Coco-Stuff-27 validation set under three different settings: globally averaged, per-image averaged, and per-image averaged with an oracle which merges oversegmented regions; our approach excels in all three.

5.9.2 Ablation over timestep for feature extraction

Regarding DDIM inversion and choice of noising step, we conduct ablation experiments, shown in Table 5.7. Our method performs robustly across different timesteps [30,45], and maintains performance even without DDIM inversion.

Step	25	30	35	40	45	40 no inv.
mIoU	43.1	45.7	46.6	46.6	45.3	46.6

Table 5.7: Ablation over noising step and importance of DDIM Inversion (mIoU) on Coco-Stuff-27.

5.9.3 Additional qualitative examples

Besides qualitative comparison for all methods on the two standard segmentation benchmark datasets CocoStuff-27 [27] in Figure 5.7 and Cityscapes [49] in Figure 5.8, we also visualize segmentations obtained via our proposed method on anomalous objects from MVTec [17] in Figure 5.10; Brain MRI scans from BraTS 2017 [136] in Figure 5.11, damaged analogue media [93] in Figure 5.12 and real-life images captured in the wild in Figure 5.13.



Figure 5.7: Qualitative comparison of segmentation results on Coco-Stuff-27 [27] Validation across multiple methods. All approaches are unsupervised; PICIE and STEGO are fit on the training set of Coco-Stuff-27, while EmerDiff, DiffSeg, DiffCut and Ours are all zero-shot and rely only on features extracted from Stable Diffusion. Our approach, in contrast to DiffSeg, DiffCut and EmerDiff, is also hyperparameter-free.

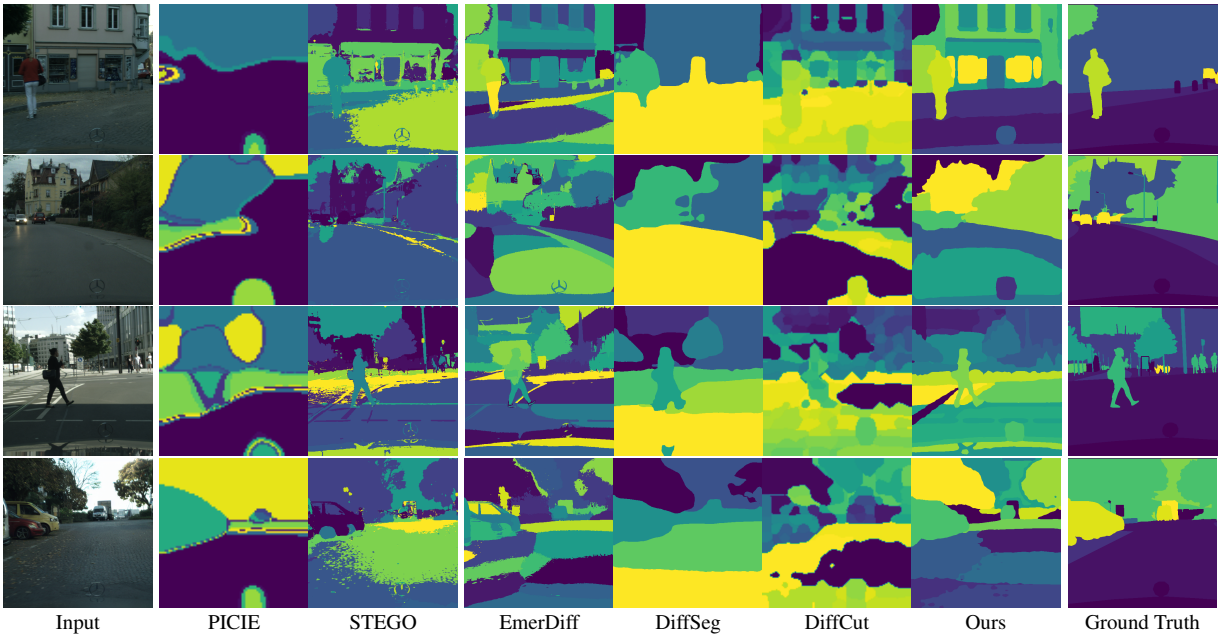


Figure 5.8: Qualitative comparison of segmentation results on Cityscapes [49] Validation across multiple methods. All approaches are unsupervised; PICIE and STEGO are fit on the training set of Cityscapes, while EmerDiff, DiffSeg, DiffCut and Ours are all zero-shot and rely only on features extracted from Stable Diffusion. Our approach, in contrast to DiffSeg, DiffCut and EmerDiff, is also hyperparameter-free.

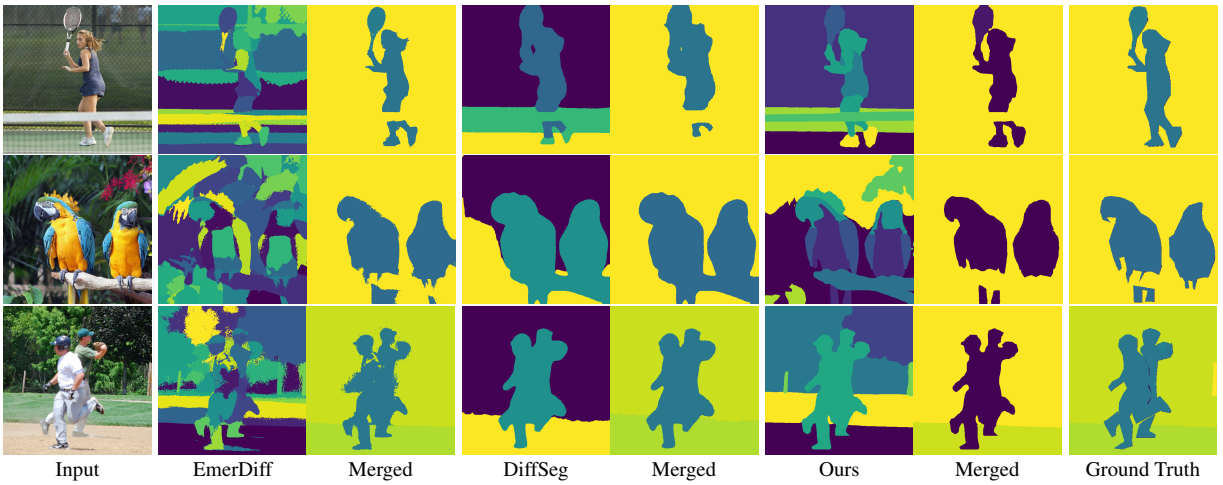


Figure 5.9: Examples of predictions used in the traditional evaluation setting (Hungarian matching) from the zero-shot diffusion feature-based approaches (columns 2, 4, and 6), and the corresponding merged predictions (columns 3, 5, and 7), designed to account for the models segmenting more granular classes compared to what is given in the ground truth. Notice that predictions from our model can be even more precise (baseball players example, row 3) or more accurate (tennis net covering the player's legs, row 1).

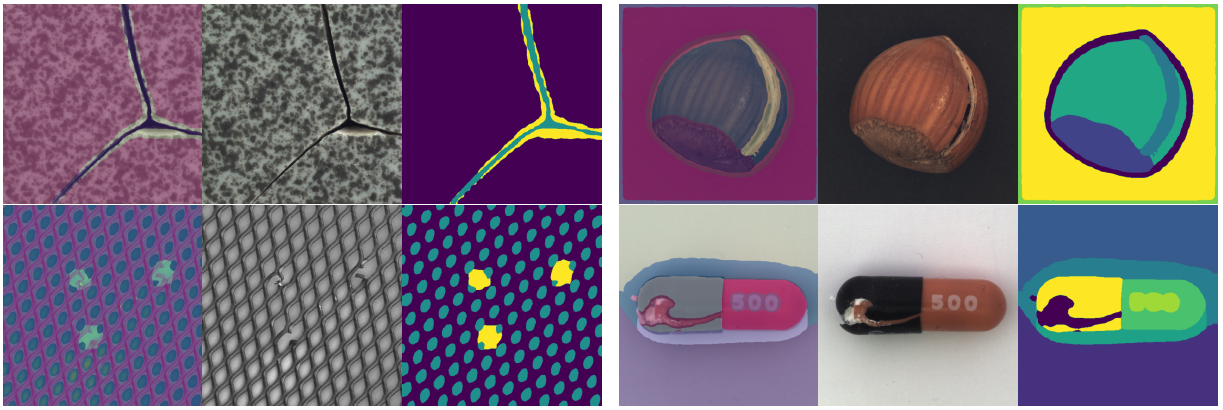


Figure 5.10: Segmentations of various types of anomalous objects from the MVTec Dataset [17].

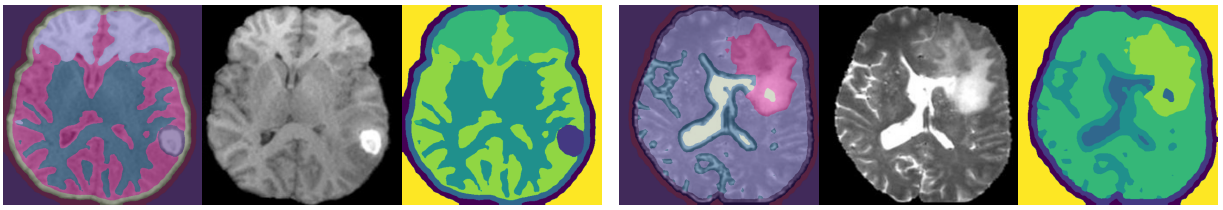


Figure 5.11: Segmentations of Brain MRI scans from the BraTS 2017 Dataset [136].



Figure 5.12: Segmentations of varying granularities on different types of damaged analogue media [93].

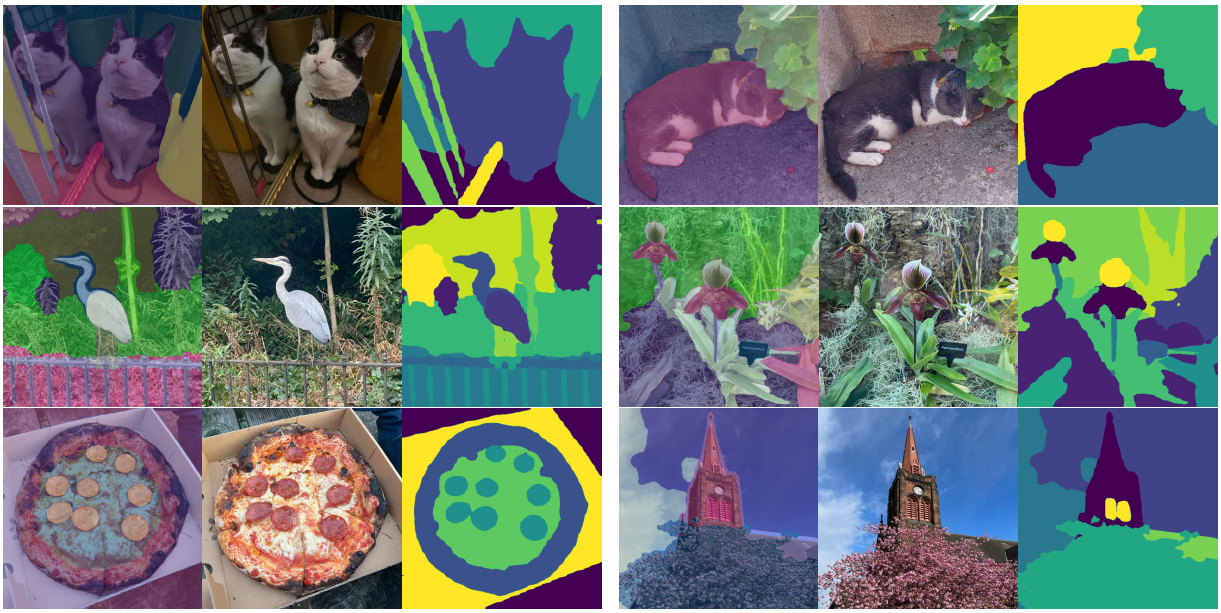


Figure 5.13: Segmentations of varying granularities of in-the-wild images captured by a smartphone: overlay (left), input (middle), and segmentation (right).

Chapter 6

Conclusion

This chapter reviews the contributions, findings, and implications of this thesis's investigation into using machine learning for the detection of damage in analogue media. It begins with a concise summary of the key contributions organized by research questions. The chapter then revisits the evolution of computational approaches to damage detection, from traditional DSP methods through supervised machine learning to foundation vision and text-to-image models, contextualizing how each contribution addresses previously identified challenges, and outlining new ones. Next, the chapter acknowledges methodological limitations related to dataset constraints, evaluation metrics, and generalization capabilities. This is followed by identifying promising directions for future work, and a discussion of practical applications. The chapter concludes by connecting the technical findings to broader themes in conservation philosophy and image restoration.

6.1 Summary of Contributions

This thesis investigated whether machine learning techniques can effectively detect and segment damage in analogue media. The work was guided by three research questions, each contributing distinct advances to the field:

RQ1: How can we develop statistically rigorous models of analogue media damage to enable robust synthetic data generation for supervised learning?

Chapter 3 addressed this question through the development of a comprehensive statistical model of film damage. By manually annotating over 12,000 individual artifacts from high-resolution film scans and extracting detailed statistics regarding size, shape, density, and spatial distribution, we created an algorithm capable of producing highly realistic synthetic damage, and validated its realism through two large-scale perceptual studies ($n = 251$ and $n = 78$, respectively). This approach enabled the generation of large-scale training data with accurate ground truth, overcoming the fundamental data scarcity limitation in supervised learning for damage detection. The statistical model's effectiveness was demonstrated through improved

performance in both artifact segmentation and restoration tasks compared to previous synthetic data generation approaches. Models were benchmarked on a novel dataset of authentically damaged and expertly restored film scans across multiple types of emulsions (e.g. positive, negative and black and white film).

RQ2: To what extent can supervised machine learning approaches generalize across different types of analogue media, damage, and contents?

Chapter 4 explored this question through the creation and analysis of ARTeFACT, the first comprehensive dataset spanning multiple media types with over 11,000 pixel-accurate annotations covering 15 damage categories, as well as textual descriptions of image contents and image-level classification based on material type (10 classes) and content type (4 classes). Systematic evaluation of state-of-the-art supervised segmentation approaches in a Leave-one-out cross-validation setting revealed fundamental limitations in cross-domain generalization. Models trained on all but one material or content class performed poorly when evaluated on the held-out class in a binary setting (damaged vs non-damaged areas), and even worse when the damage class was broken down into the 15 damage categories in a multi-class segmentation setting. Experiments with the Segment Anything Model (SAM) further revealed that even models trained on billions of masks lack the semantic understanding to conceptualize damage across diverse media. While capable of producing precise segmentations given the correct prompting technique, SAM essentially outsorces the semantic prior to the user, requiring each damage segment to be manually selected via a bounding box. The results from the Stable Diffusion-based DiffEdit and DiffSeg, however, suggested that text-to-image generative models, which are cable of generating images of any content or style, can group image regions semantically, and segment regions which most closely align to a text prompt. Still, these were not directly applicable to damage detection: DiffSeg would fail to segment damage of varying scales w.r.t to image features, and DiffEdit entangled damage with other salient features in images, such as faces.

RQ3: How can unsupervised methods leveraging multimodal foundation models overcome the limitations of supervised approaches in distinguishing damage from visually similar semantic content?

Chapter 5 began to address the issues outlined in the previous chapter with respect to multimodal text-to-image generative diffusion models by exploring how the rich internal representations of such models could be leveraged for zero-shot unsupervised semantic segmentation in a general setting. We improved upon DiffSeg by leveraging the same self-attention features for semantic segmentation, but with a top-down clustering approach: Normalized Cuts with a dynamic NCut threshold. We revealed that self-attention features can be viewed as random walks transition probabilities between patches, and utilised that to controllably to hierarchically group semantically related image regions through matrix exponentiation. This is a particularly useful mechanisum for semantic segmentation in settings where the regions we are interested in vary in scale w.r.t. other image features, i.e. analogue damage segmentation, where segment-

ing across multiple granularities while preserving semantic grouping is crucial. Developing a state-of-the-art segmentation approach by utilising a T2I generative model in Chapter 5 provides the foundation for exploring the overall RQ3, and addresses its first part by outperforming both existing unsupervised and supervised methods. Additional work, however, is required to address the latter part of this RQ, namely: how do we distinguish damage amongst the segmented clusters? Preliminary experiments and suggestions for immediate future work and outstanding challenges are detailed in Section 6.3.

The research conducted for this thesis advanced through three distinct phases: developing statistical damage models for synthetic training data generation, comprehensively evaluating supervised approaches across diverse media, and leveraging multimodal foundation model features to capture semantic understanding without explicit supervision. This progression yielded several tangible contributions: two novel benchmark datasets, two large-scale perceptual studies, a taxonomy of analogue damage for digital restoration, and a state-of-the-art zero-shot unsupervised semantic segmentation method. By synthesizing computational image restoration with analogue media conservation principles, we precisely situated analogue damage detection within the literature. This framework established both the challenges and potential solutions for computational damage detection while revealing insights about how machine learning models represent abstract concepts like damage in visually complex settings. The following section examines this in greater detail.

6.2 Evolution of Computational Damage Detection

6.2.1 DSP Limitations: Single-image, explicit modeling

As discussed in Section 2.2, digital signal processing approaches to image restoration have historically excelled at global, well-defined, mathematically tractable degradations such as blur, noise, and compression artifacts. When applied to more complex tasks like periodic noise removal—as shown in Figure 2.6—handcrafted degradation modeling could not account for a fundamental DSP limitation: its image and degradation priors are typically derived from individual images rather than learned from large datasets. Without rich priors, useful image features cannot be differentiated from degradations with similar visual signatures. Analogue media damage presents an even greater challenge as it does not admit simple mathematical models for either image content or degradation. Degradation characteristics vary significantly not only across different media types but even within a single damaged artifact.

6.2.2 Machine Learning Advances: Intra-medium generalization

Machine learning offers an advantage over DSP for image restoration: ML models simultaneously learn both a prior over images (what constitutes natural content) and a prior over damage

operators (how damage typically manifests). This dual learning capacity makes ML particularly attractive as a tool for analogue damage restoration, as it could distinguish between content and damage in ways explicit models cannot. One significant obstacle remained: lack of training data. As standard ML practice is to generate training data by applying degradation transforms to high-quality images (e.g., grayscale transformations for colorization tasks), the preliminary work in Section 2.4 relied on naïvely applying film damage overlays to clean scans to generate both training and testing data. This directly inspired the initial contributions in this thesis: acquiring authentic damaged-restored test data and developing more realistic synthetic training data.

Chapter 3 constrained the problem to a single medium. We developed a sophisticated damage generator using statistical modeling techniques to capture the distribution of five classes of real film damage. Additionally, we curated a novel benchmark dataset pairing real damaged film scans with expert restorations, enabling rigorous evaluation of both damage restoration and detection. In this intra-medium context, supervised learning proved effective: models successfully generalized across different damage instances, film emulsions, and image contents, significantly outperforming previous techniques.

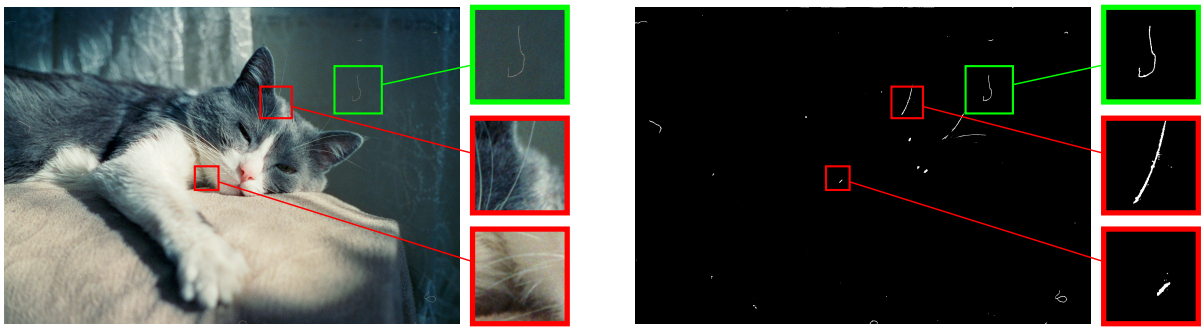


Figure 6.1: Edge case: the best performing damage segmentation model from Chapter 3, applied to a film scan in the wild. While they may *look* similar, Cooper's whiskers are *not* damage!

The fixed size of 35mm film frames (35mm by 24mm) enabled accurate capture of the scale, shape and spatial distributions of dust, scratches, dirt and hairs. Importantly, damage artifacts maintain consistent physical dimensions regardless of film format—a hair or dust speck will be the same size whether it adheres to 35mm or large format emulsion. This allowed our statistical model to generate unique overlays per data point that could be extrapolated to medium or large format film while preserving realistic damage characteristics. The generated damage not only fooled human participants in our perceptual study, but also produced training data with visual properties sufficiently aligned with real damage to enable effective learning. This consistency of damage scale is absent across other media types: paintings might fit within the palm of a hand [166] or cover an entire wall in The Doge's Palace [198]. Both could exhibit similar small-scale damage, while the larger work would additionally present large-scale damage absent in

smaller pieces. The photographic content of film also provided a more consistent visual domain compared to the diverse artistic styles and scales found across other media types.

The intra-medium generalization represented a significant advance over DSP approaches, as models could process previously unseen images without manual parameter tuning. A reflection of this was the improved performance in both segmentation and restoration with models trained on our synthetic data. Yet even in this data-driven setting, relying solely on visual features proved problematic when image content details resembled damage artifacts, as in Figure 6.1.

6.2.3 Cross-Media Challenges & The limits of visual learning

The extension to multiple media types, categorized by material and content in Chapter 4, explicitly demonstrated the limitations of purely visual learning approaches. The twofold challenge discussed earlier became more complex: first, damage exhibits tremendous visual diversity, with different damage types having distinct visual signatures even within the same medium, and these signatures vary further across different media types; second, damage can visually resemble intentional content features, a problem more pronounced in stylized depictions like paintings, drawings, or decorative elements.



a) BBox prompts per damage instance (GT annotations)

b) SAM predictions w/ BBox and Oracle

c) DiffEdit prompted with <*damaged*>

Figure 6.2: SAM requires an infeasible amount of bounding boxes (a), effectively asking the user to do the heavy lifting in damage detection to produce accurate segmentations (b). In contrast, DiffEdit, leveraging the visual-semantic features of StableDiffusion, can detect damage via a single prompt. However, it also detects salient features (faces) and objects which *look* like damage, like the string the cherub is holding.

The supervised ML approaches that (mostly) succeeded in Chapter 3 needed to model not only the diverse manifestations of damage—conceptually similar but visually different across media types—but also multiple visual styles beyond natural photographic images. The ARTe-

FACT dataset, along with ten material-based classes, further categorized these varying content types into four classes: photographic, artistic, abstract, and geometric depictions. Leave-one-out cross-validation experiments demonstrated that supervised models failed to generalize when the constraints on material or content are relaxed. This underscored that models were learning medium-specific visual patterns rather than capturing the conceptual essence of *damage* as a disruption of intended meaning.

In a similar vein, SAM, while capable of precise boundary segmentation when correctly prompted, treated each damage instance (e.g., individual cracks) as a separate entity requiring its own prompt, rather than understanding “cracks” as a semantic category encompassing multiple instances. This further underscored that purely visual learning, even at massive scale and across diverse data, remains insufficient for the conceptual understanding required to identify damage across diverse media contexts.



Figure 6.3: Text-to-image generative models are able to generate convincing "ancient" frescoes, without explicitly mentioning the presence of damage. The prompt for this image is "Ancient roman fresco of a man presenting a modern laptop, closeup photograph, pompeii –ar 3:2 –v 5.1 –q 2". Reproduced from "A well-preserved Roman fresco recently unearthed in Pompeii".

Initial experiments with diffusion models in Chapter 4 provided the first hints of a potential alternative. Benchmarking DiffEdit demonstrated that Stable Diffusion, capable of generating images of any content or style (including specific media like “painting” or “fresco”, as in Figure 6.3), could group image regions semantically based on text prompts. The detected regions were not perfectly disentangled from style and content (highlighting salient features like faces)

and somewhat noisy (detecting high frequency detail). In Section 2.3.7, amongst other data callenges, we discuss how large digitised collections of analogue media lack specific damage annotations, and sometimes the presence of damage is not explicitly stated even in the metadata of such images. In Figure 6.3, a T2I model prompted to generate a fresco can successfully add plausible damage—without "damage" explicitly having been part of the prompt. We can see this as evidence that while such models can indeed reproduce analogue damage, the concept of it is entangled with the concept of the medium itself.

Yet, DiffEdit could connect prompts such as "damaged" with corresponding damaged areas in images, as seen in Figure 6.2. Despite the lack of supervision from textual descriptions explicitly describing deterioration in analogue media, the sheer amount of text-image pairs consumed during the training of such models suggests suggests an emergent understanding of damage as a semantic concept rather than merely a visual pattern.

6.2.4 Semantic Understanding: The promise of multiple modalities

The DiffEdit and DiffSeg results from Chapter 4 inspired the exploration of Stable Diffusion features as a means for capturing damage as a visual-semantic concept.

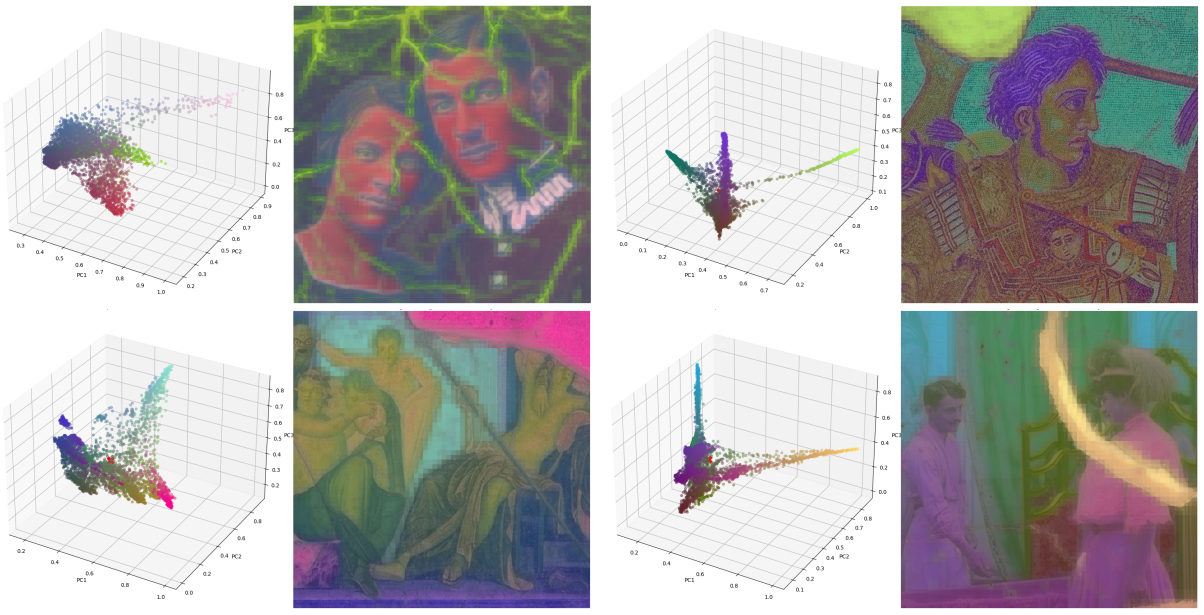


Figure 6.4: Self-attention activations extracted from Stable Diffusion for example damaged images, with Principle Component Analysis applied for dimensionality reduction. Showing the first 3 components, matched to RGB channels.

Initial investigation of self-attention features as conducted in which Principle Component Analysis was applied to self-attention features, shown in Figure 6.4, similar to how it had been done to DINO features [153]. It confirmed that these features not only grouped similar visual entities (e.g. people) within the latent space of the model, but that this emergent behaviour also extended to damage. The authors of DINO further used thresholding along the first PCA

component to identify the background in images; other works, similarly, used the Fiedler vector (eigenvector corresponding to the smallest, non-zero eigenvalue of an adjacency graph build from image features, see Chapter 5) to identify salient objects in images [135, 208]. In the case of damage, however, depending on the medium, type and prominence of damage, and visual content, the damage "clusters" could either emerge clearly within a separate component (e.g. green, corresponding to the second principle component in the RGB-based visualization), or as a combination of multiple principle components (e.g. yellow, i.e. first and second components in RGB). In other words, unlike salient objects, it was not guaranteed that damage would consistently emerge along a certain component, or a set of components, and simple thresholding along a damage/non-damage axis was not a viable solution. Hence, the problem was approached in two stages: first, segment the image such that visual entities of various granularities, including damage, are properly separated; then, select the segment(s) which corresponds to damage.

Chapter 5 focused on solving the first task, by improving upon DiffSeg. It utilized the same multimodal self-attention features, but in a top-down approach using Normalized Cuts to account for varying visual granularity. The analysis of self-attention mechanisms revealed that they naturally encode semantic relationships between image regions regardless of spatial proximity or visual consistency. The matrix exponentiation technique demonstrated how higher-order walks could modulate the hierarchical level of segmentation.



Figure 6.5: Applying the approach developed in Chapter 5 to examples from the ARTeFACT dataset from Chapter 4.

This approach achieved state-of-the-art results in unsupervised segmentation on standard semantic segmentation benchmark datasets such as Coco-Stuff-27 [27] and ADE20K [240], without any training. It addressed limitations of previous bottom-up methods like DiffSeg, providing more consistent segmentations while accommodating varying levels of visual granularity.

6.3 Outlook

Figure 6.5 qualitatively demonstrates that the method developed in Chapter 5 can successfully segment damage along with other semantic entities. Selecting which segment corresponds to damage without an oracle remains the last step in a complete damage detection pipeline.

Here, we return to ideas which follow a similar thread to DiffEdit [51]. Open-Vocabulary Attention Maps (OVAM) [132] is a recently published state-of-the-art technique which can generate text-attribution maps from Stable Diffusion cross-attention activations based on open vocabulary descriptions. It also allows for text embeddings to be optimized on a single image and target mask, and then applied for segmentation. As part of this thesis, OVAM was adapted to real images via DDIM inversion. Figure 6.6 visualizes unoptimised and optimised OVAM attention attributions on real images, using an optimized token embedding for the word <*cat*> provided by the authors.

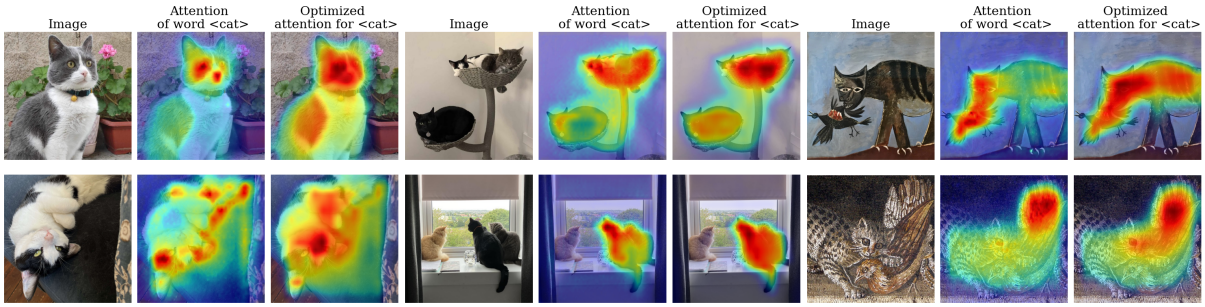


Figure 6.6: OVAM applied to various real images of cats, showing raw and optimized attention attribution maps for the <*cat*> token. First column has single cat instances, second column shows multiple cat instances, and third column shows cats in stylised depictions (mosaic and painting).

While impressive, limitations of the method are evident in column 2 and 3: even with concepts with consistent visual signatures such as cats, OVAM struggles when multiple cats are present in the image, and is inaccurate in stylized depictions, merging the cats with their pray.

The aforementioned issues are exacerbated with damage, which varies in shape, size, spatial density and prevalence a lot more than cats do. Figure 6.7 visualizes OVAM embeddings for <*damaged*>, either unoptimized or optimized for one random sample per material class from the ARTeFACT dataset, and then applied across all classes. In the case of unoptimized tokens, only `Paper` and `Wood` show attributions which actually focus on the damage, and not on the most salient object in the image. With the optimized embeddings, while detection capabilities do translate between media occasionally, like with `Film emulsion` and `Paper`, or `Wood` and `Textile`, overall, the embeddings fail to generalize. Similarly to the general case shown in Figure 6.6, even optimized OVAM embeddings struggle with multiple spatially disjoint instances.

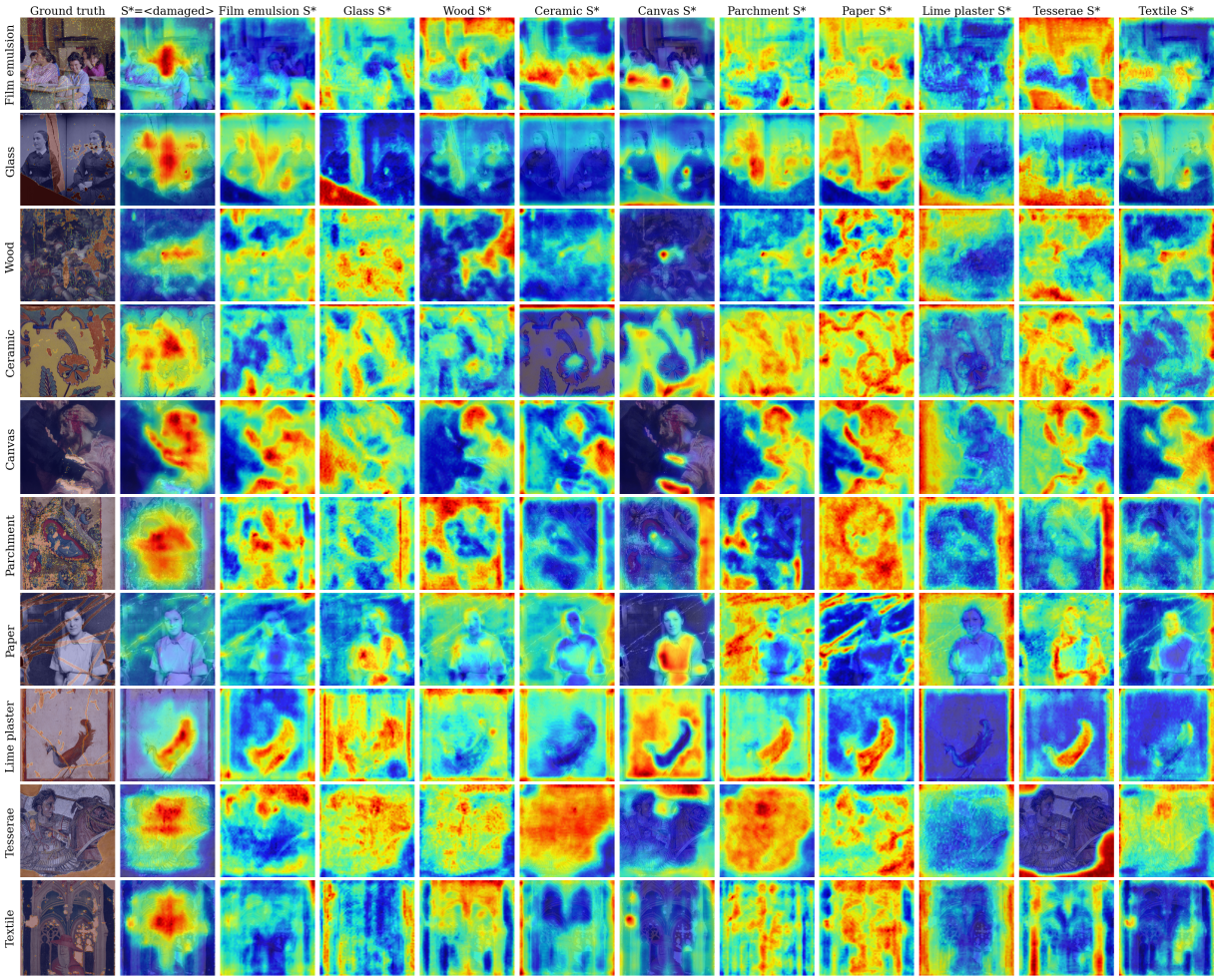
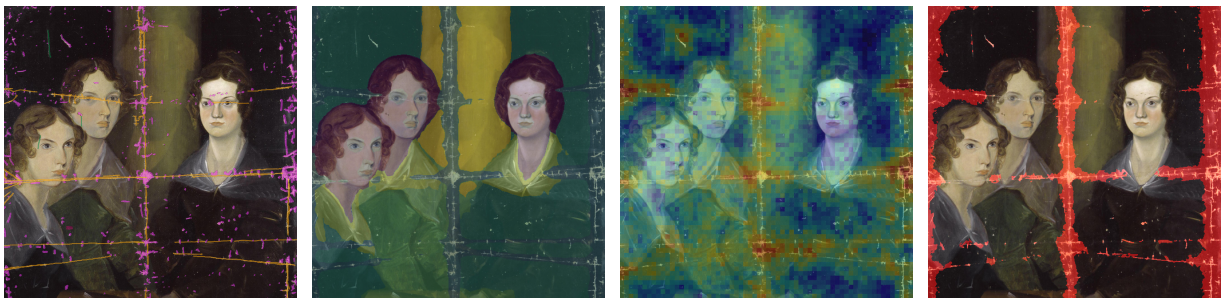


Figure 6.7: OVAM applied to examples from the ARTeFACT dataset from Chapter 4 (random sample per material class). Ground truth binarized masks in first column; unoptimized <*damaged*> token OVAM [132] attribution maps in second column; the remaining columns show the token attribution maps optimized for a given material class, applied across all classes in the dataset. Best viewed zoomed-in.

Provided the cross-attention attribution maps were accurate, they could be used to extend the Normalized Cuts approach from Chapter 5 to bias the cut [131]. Instead of partitioning the NCuts graph via the Fiedler vector (as detailed in Chapter 5), a new vector can be constructed as a weighted combination of all eigenvectors, where the weight of each is proportional to the "correlation" with the bias. In our case, this is represented via cross-attention attribution maps for a query embedding, as shown in Figure 6.8. The embedding in this case has not been optimized, yet it is sufficient for biasing the Normalised Cuts approach to specifically segment damage. This is not always the case, as shown with unoptimized embeddings for <*damage*> in Figure 6.7. An immediate line of future work presents itself in producing optimized embeddings which can generalize across different media and types of damage.



a) Ground truth (from ARTeFACT dataset) b) Unsupervised segmentation via NCuts c) Attention attribution for <*damaged*> d) Normalized Cuts biased with <*damaged*>

Figure 6.8: We can extend the approach described in Chapter 5 by using the OVAM attribution maps as bias vectors in an extension to the Normalized Cuts algorithm.

6.4 Limitations

Despite the advances presented in this thesis, several important limitations must be acknowledged. Some reflect standard challenges inherent to machine learning, such as the need for large annotated datasets, the ability of trained models to generalise on unseen data, or evaluation fairness, all discussed later in this section. Others are specific to analogue media restoration, including the scarcity of authentic paired data, the ambiguity in defining a "ground truth" former condition, and the difficulty of disentangling damage from intentional features in diverse image mediums and style. The collection and curation of new datasets was therefore imperative for any further progress. The work done in this thesis towards that end enabled the benchmarking of existing segmentation models, while the a novel unsupervised segmentation framework was developed in parallel. However, the automated selection of the damaged segments produced by our method (as well as others) remains. Immediate future work in line with addressing this is discussed in Section 6.3, with additional research directions in Section 6.5.

6.4.1 Data quality & biases

Our research datasets present several inherent constraints. The film damage benchmark dataset from Chapter 3, while encompassing diverse damage types, film emulsions, and camera configurations, was confined to the 35mm film format from a single source, the personal photography of Dmitri Tcherbadji. This approach ensured consistent digitization quality due to single-source expertise but restricted the diversity contents, geographical locations, and historical periods. Conversely, ARTeFACT from Chapter 4 sourced data from various institutions and collections, spanning multiple historical periods, geographic locations, materials, visual styles and contents. As a result, variations in digitization quality were unavoidable despite efforts to maintain high resolution. This limitation excludes three-dimensional visual media such as sculptures or pottery, which are equally subject to physical damage but would require more

complex digitization processes and data representations (e.g., voxels, NeRFs [138], or Gaussian splats [104]).

6.4.2 Processing real-world data

In both cases, a significant amount of engineering work targeted adapting machine learning models—restoration, segmentation, and generative models alike—to accommodate training and evaluation on images with much larger resolution and with varying aspect ratios compared to conventional image datasets in the field, which are usually square and have standard input sizes such as 224×224 or 512×512 pixels. For instance, DiffEdit was evaluated twice in Chapter 4: once with data square-cropped and resized to 512×512 (standard input size for Stable Diffusion), and a second time with preserving the resolution and aspect ratio of images, and using Hann windows for patch-wise processing to mitigate the loss of detail affecting small-scale damage artefacts [158]. This results in yet another trade-off: when artefact scale is incredibly small with respect of the imaging plane (as in the case of film emulsion), resizing to standard input size can make them virtually undetectable; on the contrary, processing at full resolution via patching minimizes the impact of global context, which is important in determining whether the detected line is a scratch, or a cat's whisker (see Figure 6.1).

6.4.3 Statistical modeling

The statistical modeling approach in Chapter 3, while effective for film emulsion, may not transfer directly to all analogue media types. Media like canvas or parchment experience damage through different physical mechanisms, potentially requiring more complex modeling approaches that account for material properties like directionality, tensile strength, or chemical composition, or interactions between the content and damage (e.g. dense ink might be more prone to damage). Additionally, our approach modeled different film damage types independently, whereas in reality, damage types across different media often interact: cracks can be a precursor for and develop into peels and material loss over time.

6.4.4 Evaluation

Related to our problem formulation, the evaluation methodology for damage segmentation also has inherent limitations. The binary framing of pixels as either "damaged" or "undamaged" simplifies a more nuanced reality, where damage exists along a spectrum of severity. Conservation decisions about what constitutes damage versus acceptable aging or patina involve subjective judgments that vary across cultural contexts and conservation philosophies. This binary framing may oversimplify the task and fail to capture these nuances.

For supervised approaches, our reliance on Hungarian matching or oracles to align predicted

segments with ground truth introduces methodological challenges. This approach penalizes models that segment at a different granularity level than human annotations, even when those segmentations might be semantically meaningful. This is an inherent problem of the data labeling process, where human annotators must commit to a single annotation granularity, although multiple correct ones exist, as discussed in the Supplementary for Chapter 5, Section 5.9.1. In conservation practice, different granularities of damage identification might be appropriate for different purposes, target media types, or intervention levels. This leads to another limitation born out of the misalignment of goals between conservation science and segmentation tasks. While reference-based metrics like IoU and F1 score provide quantitative assessment of segmentation accuracy, they necessitate ground truth labels, which usually capture only one possible level of semantic granularity. For instance, however, conservationists may be interested in identifying both cracks and craquelure (networks of cracks) [26, 155] as damage entities. These aspects of damage assessment may require different evaluation frameworks and data benchmarks that more closely align with conservation priorities.

6.4.5 Damage as a semantic concept

The unsupervised segmentation approach presented in Chapter 5, while demonstrating impressive general segmentation capabilities, only solves the first stage the damage detection problem. Without explicit damage-related supervision, these models can segment images into semantically meaningful regions but lack the specific capability to distinguish damage from content. Bridging this gap remains a significant challenge, with particular issues related to inaccurate text prompt visual attribution for abstract concepts such as damage discussed in Section 6.3.

6.5 Future Directions

6.5.1 Damage-Specific Semantic Segmentation via Text

A primary direction for future research is bridging the gap between general semantic segmentation capabilities demonstrated in Chapter 5 and damage-specific identification.

Cross-attention maps in text-to-image diffusion models reveal how specific words influence different image regions during generation. The qualitative analysis in Section 6.3 lays the groundwork for future work in this direction by outlining current limitations of state-of-the-art approaches such as OVAM with respect to damage detection.

While OVAM [132] demonstrates that one-shot embedding optimization generalizes well for visually consistent concepts in standard benchmarks, few-shot learning approaches may offer a solution for visually complex concepts such as damage. Providing a small number of annotated examples for specific damage types could adapt models' semantic understanding to particular

conservation contexts. This would align with practical workflows, where conservators might provide a limited set of examples specific to their collection or media type of interest.

Another interesting direction is exploring the relationship between damage detection and image editing capabilities of T2I diffusion models. Works like DiffEdit [51] have shown the ability to edit specific regions or attributes in images using text-guided diffusion. Improved cross-attention attribution could facilitate the framing of damage removal as a text-guided image editing task—for instance, by modulating or even suppressing specific model activations which are found to be most responsive to the presence of damage.

6.5.2 Technical advancements

Several potential technical advancements could be developed by expanding upon the approaches described in this thesis.

Integration with state-of-the-art inpainting methods represents a natural extension of this work. While the focus primarily fell on damage detection, combining the segmentation approaches with recent advances in inpainting could create end-to-end restoration systems that maintain the advantages of decomposed processing while offering seamless user experiences. Exploring controllable inpainting approaches that allow conservators to specify the degree of intervention or stylistic parameters would be particularly valuable; for example, it would accommodate studies on human perception of damage and restoration in artworks, like the one conducted by Tate Britain and discussed in Section 2.1, at a much larger scale.

Extending the ideas and methodologies in this thesis to temporal data for motion picture restoration presents unique challenges and opportunities. Motion picture film damage often has temporal content consistency, while individual artifacts can only be present within a single frame. Leveraging this information could improve detection accuracy and conceptual disentanglement of damage from meaningful visual content.

Multi-spectral imaging approaches offer potential for improved damage detection in challenging cases. Conservation science already utilizes infrared, ultraviolet, and X-ray imaging to reveal features not visible under normal lighting conditions. Extending machine learning approaches to incorporate multi-spectral data could enhance damage detection capabilities, particularly for damage that shares visual characteristics with intentional content in visible spectrum imaging. These methods are not without their drawbacks, however. Multi-spectral imaging often requires specialist equipment, can be costly to deploy at scale, and in some cases may expose artefacts to potentially harmful radiation or light levels. In this context, purely RGB-based framings, such as the one developed in this thesis, remain valuable as they are non-invasive, scalable, and compatible with the vast majority of existing digitised collections.

6.5.3 Theoretical advancements

The problem formulation proposed in this thesis (Section 2.3.1) necessarily oversimplifies the complex reality of analogue damage by representing it as a binary mask. Our treatment of damage as either present or absent disregards aleatoric uncertainty, i.e. the inherent ambiguity in the degree of damage present. Physical properties such as variable opacity, in which damage is spatially entangled with image content, exist on a continuum that binary classification fundamentally cannot represent. Simultaneously, our models encounter epistemic uncertainty when distinguishing damage from visually similar content features, especially in unseen media context. Future work could model both uncertainty types: probabilistic maps would express the inherent variability in damage through continuous values rather than binary classifications, while confidence maps would highlight regions where the model potentially lacks knowledge about unseen media. These approaches would transform damage detection into a nuanced probabilistic assessment better aligned with conservation practice.

The active inference framework offers another unexplored direction for damage detection research. The Tate Britain study discussed in Section 2.1.3 revealed how damaged areas in paintings attracted disproportionate visual attention, suggesting damage creates prediction errors in human perception. Within this framework, damage can be conceptualized as regions that violate expectations of what "should be there" based on surrounding undamaged content. This perspective aligns with research framing visual search as an active inference scheme, where saccadic eye movements function as experiments to resolve uncertainty [69].

This framework brings value in three important ways. First, it offers a potential solution to the labor-intensive process of manual damage annotation, creating a semi-automated labeling pipeline based on where humans naturally look, which in turn, can be calibrated with traditionally labeled images, like the ones from ARTeFACT in Chapter 4. Second, by comparing expert and novice gaze patterns (as done by Billi et al. [20]), we could develop models that distinguish between features that draw attention due to damage versus those that are artistically significant—addressing the fundamental challenge of disentangling damage from content. Third, a computational approach inspired by this could potentially harness model priors over image content to treat damage detection a process of identifying anomalous regions that maximize surprise or prediction error, as damage disrupts the continuity and coherence of the intended content.

6.5.4 Cross-Disciplinary Applications

Future research should emphasize large-scale collaboration with conservation professionals to ensure that computational approaches align with conservation needs and practices. Co-designing systems with conservators and conducting user studies to evaluate how well computational tools integrate into existing workflows would enhance their practical utility. Developing

interfaces that allow conservators to interactively refine computational predictions and incorporate domain expertise would support more effective human-AI collaboration.

Integration with physical conservation techniques presents another promising direction. Digital damage detection could inform physical intervention planning by identifying areas requiring structural stabilization or chemical treatment.

Development of tools for conservation education and training could also extend the impact of this research. Interactive systems that visualize different damage types across media categories could support conservator training. Simulations that demonstrate how damage progresses over time under different environmental conditions could aid preventive conservation planning.

Exploring how damage detection approaches might transfer to adjacent domains like archaeological documentation, architectural heritage preservation, or natural history collection management could further build upon the frameworks presented in this research. Many of these fields face similar challenges in documenting and preserving physical artifacts with complex damage patterns that require both visual assessment and contextual knowledge.

Analogue media exists beyond cultural heritage. For instance, medical whole-slide slides which contain thin, dyed tissue specimens prepared for microscopic examination, function as another form of analogue medium. The tissue itself can often be chemically died, and the slides are annotated or marked by medical professionals; they are also subject to physical wear and tear. Applying the framework and approaches developed in this thesis could aid in cleaning and preparing such historical datasets, which remain abundant in older medical records.

6.6 Practical Applications

The technical advances presented in this thesis offer potentially impactful applications for cultural heritage preservation, access, and engagement.

6.6.1 Accelerating Conservation Workflows

Automated damage detection could significantly accelerate traditional conservation workflows. Large cultural institutions like museums, archives, and libraries often manage millions of analogue media items with limited conservation staff. Even partial automation of damage identification would enable rapid triage of large collections to prioritize urgent conservation needs, systematic condition assessment across entire collections rather than representative sampling, quantitative tracking of damage progression over time for preventive planning, and consistent documentation of damage types and locations before intervention. By handling routine identification tasks, conservators could focus their expertise on more complex decision-making regarding restoration approaches, historical context assessment, and treatment planning.

6.6.2 Interactive Access to Cultural Heritage

Beyond conservation workflows, this research enables new forms of public engagement with cultural heritage through interactive interfaces. Digital platforms could allow viewers to toggle between original damage states and algorithmically restored versions, creating dialogue about authenticity, historical change, and artistic intent. Interactive systems could highlight and explain different damage types, their causes, and historical context, transforming damage from distraction to educational opportunity. Augmented reality applications could overlay restored content onto physically damaged objects in museum settings, preserving material authenticity while enhancing visual comprehension. Such applications transform damage from a purely negative attribute into a meaningful aspect of an object's provenance that connects viewers to its material history while still allowing appreciation of its original appearance.

6.6.3 Digital Surrogates with Enhanced Metadata

The computational understanding of damage developed in this thesis, particularly the move towards a unified taxonomy of analogue damage for the digital domain, could enrich digital surrogates of analogue media with enhanced metadata. Detailed damage maps could accompany digital records, providing scholars with information about an object's physical condition without requiring direct access. Damage classifications could be included in search and filtering systems, allowing researchers to identify patterns of deterioration across collections, time periods, or geographic regions. Quantitative damage assessments could inform digitization priorities and preservation strategies. These enhanced metadata layers would make damage information more accessible and interpretable for both specialists and general audiences.

6.6.4 Democratizing Access to Restoration

Finally, the approaches developed in this thesis could democratize access to restoration capabilities. Cultural institutions with limited conservation resources could employ computational tools to develop preliminary restoration plans for damaged materials. Community archives and smaller museums could offer basic digital restoration of their collections without specialized expertise. Individual owners of family photographs or historical documents could utilize accessible restoration tools to preserve personal heritage. Creators currently working with analogue media such as film could benefit from cheaper, hardware-independent approaches to detecting damage artifacts as part of the film scanning process. By bridging the gap between conservation science and computational image processing, this work contributes to a more inclusive approach to heritage preservation that extends beyond major institutions to encompass diverse collections and communities.

6.7 Final Remarks

This thesis aimed to bridge conservation and computing by aligning digital image restoration with analogue media preservation. We established the analogue media restoration task within the digital domain, reconciling definitions of image, damage, and restoration.

Initially, we employed machine learning to model film damage, generalizing well on similar photographic content and film types. However, limitations emerged when image features resembled damage, indicating a need for broader understanding beyond visual patterns. Extending to diverse media and content confirmed that standard supervised ML learned material-specific patterns, and did not capture the essence of damage as a generalized concept.

To address this, we explored multimodal foundation models, leveraging text-to-image diffusion models. Our unsupervised segmentation demonstrated semantic grouping of visually disparate elements. As future work we propose using text prompts to identify damage within these segments, enabling semantic, rather than purely visual, damage understanding. This aligns with human expert practice: recognizing damage not merely as a visual pattern but as a disruption of intended meaning—the very understanding that human experts apply in conservation practice.

In returning to Vermeer's "The Love Letter" and the layers of transformation it underwent—from artistic creation, to physical damage, to digital capture, to computational analysis—we see how each stage reveals a different aspect of what might be obscuring the "true image". Unlike physical restoration, which necessarily commits to a single interpretive path, computational approaches to analogue damage detection enable multiple possible restorations to coexist digitally. Perhaps the most profound insight of this work is that the elusive *true image* is not static; it consists of all of these layers across time, media and imaging methods. Despite philosophical disagreements about what damage is, or how it should be dealt with, the seemingly effortless perceptual ability of humans to identify its presence represents one of the most challenging frontiers for machine learning. It requires models to understand not just patterns but meaning, to recognize that similar visual features can represent either intentional content or unwanted damage depending on their semantic context, or lack thereof.

The difficulty of this task, as demonstrated throughout this thesis, stems from the abstract nature of *damage* itself. Unlike standard object recognition tasks where visual consistency exists across examples, damage is defined by its relationship to intended content rather than visual characteristics only.

The capability to understand damage is valuable beyond the practical applications in heritage preservation. It forces us to develop models that can reason about relationships between content and medium, corras humans do. Whether fully replicating these sophisticated forms of visual understanding is possible with machine learning remains to be seen.

Bibliography

[1] Panos Achlioptas et al. “Learning representations and generative models for 3d point clouds”. In: *International Conference on Machine Learning*. PMLR. 2018, pp. 40–49.

[2] Eleni Aggelakopoulou and Asterios Bakolas. “What were the colors of the Parthenon? Investigation of the entablature’s cornice blocks”. In: *Journal of Archaeological Science* 140 (2022), p. 105553.

[3] Jourdan Aldredge. *Why ’Dune’ Was Shot on Digital, Transferred to 35mm, Then Scanned to Digital*. https://nofilmschool.com/Dune-Digital-Film-Process. Accessed: 2023-03-19. 2021. URL: https://nofilmschool.com/Dune-Digital-Film-Process.

[4] Shir Amir et al. “Deep ViT Features as Dense Visual Descriptors”. In: *arXiv preprint arXiv:2112.05814* (2021).

[5] Harry C Andrews. “Digital image restoration: A survey”. In: *Computer* 7.5 (1974), pp. 36–45.

[6] Jason Antic. *DeOldify: A Deep Learning based project for colorizing and restoring old images (and video!)* 2020. URL: https://github.com/jantic/DeOldify.

[7] Nikita Araslanov and Stefan Roth. “Single-stage semantic segmentation from image labels”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 4253–4262.

[8] Jonathan Ashley-Smith. “Definitions of damage”. In: *Text of a talk given in the session" When conservator and collections meet" at the Annual Meeting of the Associaton of Art Historians. London*. 1995.

[9] Jonathan Ashley-Smith. “The Ethics of Conservation”. In: *The Conservator* 6.1 (1982), pp. 1–5. DOI: 10.1080/01400096.1982.9635616.

[10] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. “SegNet: A deep convolutional encoder-decoder architecture for image segmentation”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.12 (2017), pp. 2481–2495.

[11] Arian Bakhtiarnia, Qi Zhang, and Alexandros Iosifidis. “Efficient High-Resolution Deep Learning: A Survey”. In: *arXiv preprint arXiv:2207.13050* (2022).

[12] Mark R Banham and Aggelos K Katsaggelos. "Digital image restoration". In: *IEEE signal processing magazine* 14.2 (1997), pp. 24–41.

[13] Dmitry Baranchuk et al. "Label-Efficient Semantic Segmentation with Diffusion Models". In: *International Conference on Learning Representations*. 2021.

[14] Mauro Barni, Franco Bartolini, and Vito Cappellini. "Image processing for virtual restoration of artworks". In: *IEEE multimedia* 7.2 (2000), pp. 34–37.

[15] J.H. Beck. *Reversibility, Fact Or Fiction?: The Dangers of Art Restoration*. Ars Brevis Foundation, 1999. URL: https://books.google.co.uk/books?id=8FGAQwAACAAJ.

[16] John Berger. *Ways of seeing*. Penguin UK, 2008.

[17] Paul Bergmann et al. "MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.

[18] Jorge Bernal, Javier Sánchez, and Fernando Vilarino. "Towards automatic polyp detection with a polyp appearance model". In: *Pattern Recognition* 45.9 (2012), pp. 3166–3182.

[19] Jorge Bernal et al. "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians". In: *Computerized medical imaging and graphics* 43 (2015), pp. 99–111.

[20] Eliana Billi, Alessandra Maria Genovese, and Stefano Sdoia. "From Mutilated to Complete Image. Lacunae in Paintings Through the Eyes of Restorers, Art Historians and Ordinary Viewers". In: *International and Interdisciplinary Conference on Image and Imagination*. Springer. 2021, pp. 496–503.

[21] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Vol. 4. 4. Springer, 2006.

[22] Tibor Bleidt, Sedigheh Eslami, and Gerard de Melo. "ArtQuest: Countering Hidden Language Biases in ArtVQA". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2024, pp. 7326–7335.

[23] Raphaël Bornard. "Probabilistic approaches for the digital restoration of television archives". PhD thesis. Ecole Centrale Paris, 2002.

[24] Cesare Brandi. *Teoria del restauro*. La Nave di Teseo Editore spa, 1963.

[25] Antoni Buades, Bartomeu Coll, and J-M Morel. "A non-local algorithm for image denoising". In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. Vol. 2. Ieee. 2005, pp. 60–65.

[26] Spike Bucklow. “The description of craquelure patterns”. In: *Studies in conservation* 42.3 (1997), pp. 129–140.

[27] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. “Coco-stuff: Thing and stuff classes in context”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 1209–1218.

[28] Nian Cai et al. “Blind inpainting using the fully convolutional neural network”. In: *The Visual Computer* 33 (2017), pp. 249–261.

[29] Luca Calatroni et al. “Unveiling the invisible: mathematical methods for restoring and interpreting illuminated manuscripts”. In: *Heritage science* 6 (2018).

[30] Kristel Van Camp. “Damage Atlas for Photographic materials. Analogue objects”. In: *CeROArt. Conservation, exposition, Restauration d’Objets d’Art*. EGG 1. Association CeROArt asbl. 2010.

[31] Mathilde Caron et al. “Emerging Properties in Self-Supervised Vision Transformers”. In: *Proceedings of the International Conference on Computer Vision (ICCV)*. 2021.

[32] M. Chambah. “Digital film restoration and image quality”. In: *ICA-BELGIUM Colour Symposium*. Ghent, Belgium, 2019. URL: https://hal.archives-ouvertes.fr/hal-02998573.

[33] M. Chambah et al. “Further image quality assessment in digital film restoration”. In: *Image Quality and System Performance III*. Ed. by Luke C. Cui and Yoichi Miyake. Vol. 6059. International Society for Optics and Photonics. SPIE, 2006, 60590S. DOI: 10.1117/12.642414. URL: https://doi.org/10.1117/12.642414.

[34] Majed Chambah, Christophe Saint-Jean, and Francois Helt. “Image quality evaluation in the field of digital film restoration”. In: *Proceedings of SPIE - The International Society for Optical Engineering* 5668 (Jan. 2005). DOI: 10.1117/12.586738.

[35] Tony F Chan and Luminita A Vese. “Active contours without edges”. In: *IEEE Transactions on image processing* 10.2 (2001), pp. 266–277.

[36] Hanting Chen et al. “Pre-trained image processing transformer. In 2021 IEEE”. In: *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 12294–12305.

[37] Jiaqi Chen et al. “Generative semantic segmentation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 7111–7120.

[38] Liangyu Chen et al. “Simple Baselines for Image Restoration”. In: *European Conference on Computer Vision*. 2022, pp. 17–33.

[39] Tianrun Chen et al. “Sam-adapter: Adapting segment anything in underperformed scenes”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 3367–3375.

[40] Xiaoran Chen et al. “Unsupervised lesion detection via image restoration with a normative prior”. In: *Medical image analysis* 64 (2020), p. 101713.

[41] Yida Chen, Fernanda Viégas, and Martin Wattenberg. “Beyond surface statistics: Scene representations in a latent diffusion model, 2023”. In: *NeurIPS 2023 Workshop on Diffusion Models* (2023).

[42] Jang Hyun Cho et al. “Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 16794–16804.

[43] Dario Cioni et al. “Diffusion Based Augmentation for Captioning and Retrieval in Cultural Heritage”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

[44] Lucia Cipolina-Kun, Simone Caenazzo, and Gaston Mazzei. “Comparison of CoModGans, LaMa and GLIDE for Art Inpainting Completing MC Escher’s Print Gallery”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 716–724.

[45] Nadav Cohen, Yael Newman, and Ariel Shamir. “Semantic segmentation in art paintings”. In: *Computer graphics forum*. Vol. 41. Wiley Online Library. 2022.

[46] Edo Collins et al. “Editing in style: Uncovering the local semantics of gans”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 5771–5780.

[47] Marcos V Conde and Kerem Turgutlu. “CLIP-Art: Contrastive pre-training for fine-grained art classification”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

[48] International Committee for the Conservation of Mosaics. Conference. *The Conservation and Presentation of Mosaics: at what Cost?: Proceedings of the 12th ICCM Conference, Sardinia, October 27-31, 2014*. Getty Conservation Institute, 2017.

[49] Marius Cordts et al. “The Cityscapes Dataset for Semantic Urban Scene Understanding”. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[50] Bruno Cornelis et al. “Crack detection and inpainting for virtual restoration of paintings: The case of the Ghent Altarpiece”. In: *Signal Processing* 93.3 (2013).

[51] Guillaume Couairon et al. “Diffedit: Diffusion-based semantic image editing with mask guidance”. In: *arXiv preprint arXiv:2210.11427* (2022).

[52] Paul Couairon et al. “Zero-Shot Image Segmentation via Recursive Normalized Cut on Diffusion Features”. In: *arXiv preprint arXiv:2406.02842* (2024).

[53] Elliot J. Crowley and Andrew Zisserman. “The State of the Art: Object Retrieval in Paintings using Discriminative Regions”. In: *British Machine Vision Conference*. 2014.

[54] George Cybenko. “Approximation by superpositions of a sigmoidal function”. In: *Mathematics of control, signals and systems* 2.4 (1989), pp. 303–314.

[55] Rui Dang et al. “The mathematical expression of damage law of museum lighting on dyed artworks”. In: *Scientific Reports* 11.1 (2021).

[56] P De Willigen. *A mathematical study on craquelure and other mechanical damage in paintings*. Delft University Press, 1999.

[57] Aditya Deshpande et al. “Learning diverse image colorization”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 6837–6845.

[58] Prafulla Dhariwal and Alexander Nichol. “Diffusion models beat gans on image synthesis”. In: *Advances in Neural Information Processing Systems* 34 (2021).

[59] William E Donath and Alan J Hoffman. “Lower bounds for the partitioning of graphs”. In: *IBM Journal of Research and Development* 17.5 (1973), pp. 420–425.

[60] Chao Dong et al. “Image Super-resolution Using Deep Convolutional Networks, vol. 38”. In: (2014).

[61] Mark Eramian et al. “Benchmarking human performance in semi-automated image segmentation”. In: *Interacting with Computers* 32.3 (2020), pp. 233–245.

[62] Miguel Espinosa et al. “There is no samantics! exploring sam as a backbone for visual understanding tasks”. In: *arXiv preprint arXiv:2411.15288* (2024).

[63] Haoqiang Fan, Hao Su, and Leonidas J Guibas. “A Point Set Generation Network for 3D Object Reconstruction from a Single Image”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 605–613.

[64] Miroslav Fiedler. “Algebraic connectivity of graphs”. In: *Czechoslovak mathematical journal* 23.2 (1973), pp. 298–305.

[65] Gabriel Fielding. *Digital ICE: Defect Detection and Correction Using Infrared-enabled Scanners*. Technical Report. Available at: https://www.kodak.com/motion/hub/itp/dice/default.htm. Eastman Kodak Company, 2008.

[66] Andreas Fischer. “IAM-HistDB a dataset of handwritten historical documents”. In: *Handwritten historical document analysis, recognition, and retrieval-state of the art and future trends* (2020).

[67] B Fisher. “Digital restoration of snow white: 120,000 famous frames are back”. In: *Advanced Imaging* (1993), pp. 32–36.

[68] Giovanna Fossati. *From Grain to Pixel: The Archival Life of Film in Transition*. Vol. 1. Framing Film. Retrieved 25 August 2012. Amsterdam, Netherlands: Amsterdam University Press, 2010. ISBN: 9789089641397.

[69] Karl Friston et al. "Perceptions as hypotheses: saccades as experiments". In: *Frontiers in psychology* 3 (2012), p. 151.

[70] Leonardo Galteri et al. "Deep generative adversarial compression artifact removal". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 4826–4835.

[71] Robert G Gann. *Image scanner and method for detecting a defect in an image to be scanned*. (Google Patents Entry) //patents.google.com/patent/US7355159B2/. Apr. 2008.

[72] Ioannis Giakoumis, Nikos Nikolaidis, and Ioannis Pitas. "Digital image processing techniques for the detection and removal of cracks in digitized paintings". In: *IEEE Transactions on Image Processing* 15.1 (2005), pp. 178–188.

[73] Rafael C Gonzalez. *Digital image processing*. Pearson education india, 2009.

[74] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. URL: http://www.deeplearningbook.org.

[75] Ian J. Goodfellow et al. "Generative adversarial nets". In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'14. Montreal, Canada: MIT Press, 2014, pp. 2672–2680.

[76] JoAnn Greco. "Saving Old Movies". In: *Distillations* 4.3 (Nov. 2018). Retrieved April 23, 2020, pp. 36–39.

[77] Hillary K. Grigonis. "Kodak shares surprising statistics on the rise of film as factory reopens after shutdown". In: *Digital Camera World* (Jan. 2025). URL: https://www.digitalcameraworld.com/cameras/film-cameras/kodak-shares-surprising-statistics-on-the-rise-of-film-as-factory-reopens-after-shutdown (visited on 01/24/2025).

[78] Varun Gupta et al. "Restoration of artwork using deep neural networks". In: *Evolving Systems* 12.2 (2021), pp. 439–446.

[79] Yatharth Gupta et al. *Progressive Knowledge Distillation Of Stable Diffusion XL Using Layer Level Loss*. 2024. arXiv: 2401.02677 [cs.CV].

[80] Cusuh Ham et al. "Variational Image Inpainting". In: *NeurIPS 2018 Workshops*. 2018.

[81] Mark Hamilton et al. "Unsupervised Semantic Segmentation by Distilling Feature Correspondences". In: *International Conference on Learning Representations*. 2022. URL: https://openreview.net/forum?id=SaKO6z6Hl0c.

[82] Fengxiang He, Tongliang Liu, and Dacheng Tao. "Why resnet works? residuals generalize". In: *IEEE transactions on neural networks and learning systems* 31.12 (2020), pp. 5349–5362.

[83] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[84] Kaiming He et al. "Mask r-cnn". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961–2969.

[85] Eric Hedlin et al. "Unsupervised semantic correspondence using stable diffusion". In: *Advances in Neural Information Processing Systems* 36 (2024).

[86] Martin Heidegger. *The origin of the work of art*. University of Waterloo, 1963.

[87] Amir Hertz et al. "Blind Visual Motif Removal from a Single Image". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 6858–6867.

[88] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. "Multilayer feedforward networks are universal approximators". In: *Neural networks* 2.5 (1989), pp. 359–366.

[89] Satoshi Iizuka and Edgar Simo-Serra. "DeepRemaster: Temporal Source-Reference Attention Networks for Comprehensive Video Enhancement". In: *ACM Transactions on Graphics (Proc. of SIGGRAPH ASIA)* 38.6 (2019), pp. 1–13.

[90] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. "Let there be Color!: Joint End-to-end Learning of Global and Local Image Priors for Automatic Image Colorization with Simultaneous Classification". In: *ACM Transactions on Graphics (Proc. of SIGGRAPH 2016)* 35.4 (2016), 110:1–110:11.

[91] Phillip Isola et al. "Image-to-image translation with conditional adversarial networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1125–1134.

[92] Daniela Ivanova, John Williamson, and Paul Henderson. "Simulating analogue film damage to analyse and improve artefact restoration on high-resolution scans". In: *Computer Graphics Forum (Proceedings of Eurographics 2023)* 42.2 (2023). DOI: 10.1111/cgf.14749.

[93] Daniela Ivanova et al. "State-of-the-Art Fails in the Art of Damage Detection". In: *European Conference on Computer Vision (ECCV) Workshop on VISART* (2024).

[94] Daniela Ivanova., Jan Siebert., and John Williamson. "Perceptual Loss based Approach for Analogue Film Restoration". In: *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP,* INSTICC. SciTePress, 2022, pp. 126–135. ISBN: 978-989-758-555-5. DOI: 10.5220/0010829300003124.

[95] Jonathan Janson. *Vermeer Thefts: 1971 - The Love Letter*. Accessed: 2025-03-10. 2023. URL: https://www.essentialvermeer.com/fakes_thefts_school_of_delft_lost_sp/vermeer_theft_01.html.

[96] Jongheon Jeong et al. "Winclip: Zero-/few-shot anomaly classification and segmentation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 19606–19616.

[97] Xu Ji, Joao F Henriques, and Andrea Vedaldi. "Invariant information clustering for unsupervised image classification and segmentation". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 9865–9874.

[98] Xin Jin et al. "Rethinking Exposure: End-to-End Low-light Image Enhancement". In: *IEEE Transactions on Neural Networks and Learning Systems* (2023).

[99] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. "Perceptual Losses for Real-Time Style Transfer and Super-Resolution". In: *European Conference on Computer Vision*. Springer. 2016, pp. 694–711.

[100] Laurent Joyeux, Samia Boukir, and Bernard Besserer. "Film line scratch removal using Kalman filtering and Bayesian restoration". In: *Proceedings Fifth IEEE Workshop on Applications of Computer Vision*. IEEE. 2000, pp. 8–13.

[101] Dahyun Kang and Minsu Cho. "In Defense of Lazy Visual Grounding for Open-Vocabulary Semantic Segmentation". In: *European Conference on Computer Vision*. 2024.

[102] Tero Karras et al. "Training generative adversarial networks with limited data". In: *Advances in neural information processing systems* 33 (2020), pp. 12104–12114.

[103] Bahjat Kawar et al. "Denoising Diffusion Restoration Models". In: *Advances in Neural Information Processing Systems*. 2022.

[104] Bernhard Kerbl et al. "3D Gaussian splatting for real-time radiance field rendering." In: *ACM Trans. Graph.* 42.4 (2023), pp. 139–1.

[105] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. "Accurate image super-resolution using very deep convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 1646–1654.

[106] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[107] Diederik P. Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2014. URL: http://arxiv.org/abs/1312.6114.

[108] Alexander Kirillov et al. "Segment anything". In: *arXiv preprint arXiv:2304.02643* (2023).

[109] Wolfgang Köhler. "Gestalt psychology". In: *Psychologische forschung* 31.1 (1967), pp. XVIII–XXX.

[110] L Kuiper, Patricia Wardle, and Alan Griffiths. "Restauratieverslag van Vermeer's Liefdesbrief/Report on the Restoration of Vermeer's Love Letter". In: *Bulletin van het Rijksmuseum* 20.Afl. 3 (1972), pp. 147–167.

[111] Manoj Kumar, Dirk Weissenborn, and Nal Kalchbrenner. "Colorization transformer". In: *arXiv preprint arXiv:2102.04432* (2021).

[112] Christian Ledig et al. "Photo-realistic single image super-resolution using a generative adversarial network". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4681–4690.

[113] Yunseung Lee and Pilsung Kang. "AnoViT: Unsupervised anomaly detection and localization with vision transformer-based encoder-decoder". In: *IEEE Access* 10 (2022), pp. 46717–46724.

[114] Jaakko Lehtinen et al. "Noise2Noise: Learning Image Restoration without Clean Data". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 2965–2974. URL: https://proceedings.mlr.press/v80/lehtinen18a.html.

[115] Daiqing Li et al. "Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 8300–8311.

[116] Feng Li et al. "Mask dino: Towards a unified transformer-based framework for object detection and segmentation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 3041–3050.

[117] Juncheng Li et al. "All-in-One Image Restoration for Unknown Degradation". In: *arXiv preprint arXiv:2204.03965* (2022).

[118] Xiaoming Li et al. "Enhanced Blind Face Restoration with Multi-Exemplar Images and Adaptive Spatial Feature Fusion". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 2706–2715.

[119] Yawei Li et al. "Efficient and Explicit Modelling of Image Hierarchies for Image Restoration". In: *International Journal of Computer Vision* (2023), pp. 1–22.

[120] Jingyun Liang et al. "SwinIR: Image restoration using swin transformer". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 1833–1844.

[121] Bee Lim et al. “Enhanced deep residual networks for single image super-resolution”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2017, pp. 136–144.

[122] Don J Lindler. “Block iterative restoration of astronomical images from the Hubble Space Telescope”. In: *The Restoration of HST Images and Spectra; proceedings of a workshop held at the Space Telescope Science Institute, Baltimore, Maryland, 20–21 August 1990*. Space Telescope Science Institute, Baltimore, MD. 1991, p. 47.

[123] Guilin Liu et al. “Image inpainting for irregular holes using partial convolutions”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 85–100.

[124] Haotian Liu et al. “Visual Instruction Tuning”. In: *NeurIPS*. 2023.

[125] L Liu et al. “Digital restoration of colour cinematic films using imaging spectroscopy and machine learning”. In: *Scientific reports* 12.1 (2022).

[126] Ze Liu et al. “Swin transformer: Hierarchical vision transformer using shifted windows”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.

[127] Zhuang Liu et al. “A convnet for the 2020s”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

[128] Andreas Lugmayr et al. “RePaint: Inpainting using Denoising Diffusion Probabilistic Models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 11461–11471.

[129] Grace Luo et al. “Diffusion Hyperfeatures: Searching Through Time and Space for Semantic Correspondence”. In: *Advances in Neural Information Processing Systems*. 2023.

[130] Sarah MaiSey et al. “Recovering from destruction: the conservation, reintegration and perceptual analysis of a flood-damaged painting by John Martin”. In: (2011).

[131] Subhransu Maji, Nisheeth K Vishnoi, and Jitendra Malik. “Biased normalized cuts”. In: *CVPR 2011*. IEEE. 2011, pp. 2057–2064.

[132] Pablo Marcos-Manchón et al. “Open-Vocabulary Attention Maps with Token Optimization for Semantic Segmentation in Diffusion Models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2024, pp. 9242–9252.

[133] Marshall McLuhan. *Understanding media: The extensions of man*. MIT press, 1994.

[134] Marina Meila and Jianbo Shi. “Learning Segmentation by Random Walks”. In: *Advances in Neural Information Processing Systems*. Ed. by T. Leen, T. Dietterich, and V. Tresp. Vol. 13. MIT Press, 2000.

[135] Luke Melas-Kyriazi et al. “Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

[136] Bjoern H Menze et al. “The multimodal brain tumor image segmentation benchmark (BRATS)”. In: *IEEE transactions on medical imaging* 34.10 (2014), pp. 1993–2024.

[137] Fabio Merizzi et al. “Deep image prior inpainting of ancient frescoes in the Mediterranean Alpine arc”. In: *Heritage Science* 12.1 (2024).

[138] Ben Mildenhall et al. “Nerf: Representing scenes as neural radiance fields for view synthesis”. In: *Communications of the ACM* 65.1 (2021), pp. 99–106.

[139] S. Minaee et al. “Image Segmentation Using Deep Learning: A Survey”. In: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 44.07 (2022), pp. 3523–3542. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2021.3059968.

[140] Ionuţ Mironică. “A Generative Adversarial Approach with Residual Learning for Dust and Scratches Artifacts Removal”. In: *Proceedings of the 2nd Workshop on Structuring and Understanding of Multimedia heritAge Contents*. 2020, pp. 15–22.

[141] Fernando Moral-Andrés et al. “Can artificial intelligence reconstruct ancient mosaics?” In: *Studies in Conservation* 69.5 (2024), pp. 313–326.

[142] Salvador Muñoz-Viñas. *Contemporary theory of conservation*. Routledge, 2012.

[143] M. R. Nagel. “NASA Seminar on the Evaluation of Motion Degraded Images, Cambridge, 3–5 December 1968”. In: *Appl. Opt.* 8.5 (May 1969), pp. 939–949. DOI: 10.1364/AO.8.000939. URL: https://opg.optica.org/ao/abstract.cfm?URI=ao-8-5-939.

[144] Koichi Namekata et al. “EmerDiff: Emerging Pixel-level Semantic Knowledge in Diffusion Models”. In: *The Twelfth International Conference on Learning Representations*. 2024. URL: https://openreview.net/forum?id=YqyTXmF8Y2.

[145] NASA Science. “Hubble’s Mirror Flaw”. In: *NASA Science* (n.d.). Accessed: 10 March 2025. URL: https://science.nasa.gov/mission/hubble/observatory/design/optics/hubbles-mirror-flaw/.

[146] Kamyar Nazeri et al. “Edgeconnect: Generative image inpainting with adversarial edge learning”. In: *arXiv preprint arXiv:1901.00212* (2019).

[147] Barry Neild. “Botched Restoration of Ecce Homo Fresco Shocks Spain”. In: *The New York Times* (Aug. 2012). Accessed: March 12, 2025. URL: https://www.nytimes.com/2012/08/24/world/europe/botched-restoration-of-ecce-homo-fresco-shocks-spain.html (visited on 03/12/2025).

[148] Mark Newman. *Networks*. Page 137, Eq: 6.31. Oxford University Press, 2018.

[149] Alex Nichol et al. “Glide: Towards photorealistic image generation and editing with text-guided diffusion models”. In: *arXiv preprint arXiv:2112.10741* (2021).

[150] Konstantinos Ntirogiannis, Basilis Gatos, and Ioannis Pratikakis. “ICFHR2014 competition on handwritten document image binarization (H-DIBCO 2014)”. In: *2014 14th International conference on frontiers in handwriting recognition*. IEEE. 2014, pp. 809–813.

[151] Amy-Fern Nuttall. “Analogue Resurgence Continues with New Film Cameras”. In: *The Royal Photographic Society* (Mar. 2023). Analogue Group Web Editor.

[152] Andrew Oddy and Sara Carroll. “Reversibility–Does it Exist?” In: *Occasional Paper Number 135* (1999).

[153] Maxime Oquab et al. “Dinov2: Learning robust visual features without supervision”. In: *arXiv preprint arXiv:2304.07193* (2023).

[154] Deepak Pathak et al. “Context encoders: Feature learning by inpainting”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2536–2544.

[155] Ludovic Pauchard and Frédérique Giorgiutti-Dauphiné. “Craquelures and pictorial matter”. In: *Journal of Cultural Heritage* 46 (2020).

[156] Jialun Peng et al. “Generating diverse structure for image inpainting with hierarchical VQ-VAE”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 10775–10784.

[157] Ken Perlin. “An Image Synthesizer”. In: *Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH ’85. New York, NY, USA: Association for Computing Machinery, 1985, pp. 287–296. ISBN: 0897911660. DOI: 10.1145/325334.325247. URL: https://doi.org/10.1145/325334.325247.

[158] Nicolas Pielawski and Carolina Wählby. “Introducing Hann windows for reducing edge-effects in patch-based image segmentation”. In: *PloS one* 15.3 (2020), e0229839.

[159] Dustin Podell et al. “SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis”. In: *The Twelfth International Conference on Learning Representations*.

[160] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: 2103.00020 [cs.CV].

[161] Aditya Ramesh et al. *Zero-Shot Text-to-Image Generation*. 2021. arXiv: 2102.12092 [cs.CV].

[162] James M Reilly. “Stability Problems of 19th and 20th Century Photographic Materials (1980)”. In: *Issues in the Conservation of Photographs* 3 (2010), p. 40.

[163] Dongwei Ren et al. “Progressive image deraining networks: A better and simpler baseline”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 3937–3946.

[164] Tianhe Ren et al. “Grounded sam: Assembling open-world models for diverse visual tasks”. In: *arXiv preprint arXiv:2401.14159* (2024).

[165] *restoration, n.* In: *OED Online*. Oxford University Press, Mar. 2025. DOI: 10.1093/OED/1486514540. URL: https://doi.org/10.1093/OED/1486514540 (visited on 07/21/2025).

[166] Rembrandt Harmensz van Rijn. *Self-Portrait*. One of only five known Rembrandt paintings on copper, and one of three with gold leaf ground. Stockholm, Sweden, 1630. URL: https://artsandculture.google.com/asset/selfportrait-rembrandt-harmensz-van-rijn/WAGlk3EYZbNjQQ (visited on 03/18/2025).

[167] Robin Rombach et al. “High-Resolution Image Synthesis with Latent Diffusion Models”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, June 2022, pp. 10674–10685. DOI: 10.1109/CVPR52688.2022.01042.

[168] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

[169] Azriel Rosenfeld. “Picture processing by computer”. In: *ACM Computing Surveys (CSUR)* 1.3 (1969), pp. 147–176.

[170] Matt Rothman. “Disney, Kodak bow restored ’Snow’”. In: *Variety* (June 1993). Accessed: March 5, 2025. URL: https://variety.com/1993/film/news/disney-kodak-bow-restored-snow.

[171] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. “Learning representations by back-propagating errors”. In: *nature* 323.6088 (1986), pp. 533–536.

[172] John Ruskin. “The lamp of memory (1849)”. In: *Historical Perspectives on Preventive Conservation* 6.2 (2013).

[173] Chitwan Saharia et al. “Image Super-Resolution via Iterative Refinement”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[174] Chitwan Saharia et al. “Palette: Image-to-image diffusion models”. In: *ACM SIGGRAPH 2022 Conference Proceedings*. 2022, pp. 1–10.

[175] Babak Saleh and Ahmed Elgammal. “Large-scale classification of fine-art paintings: Learning the right metric on the right feature”. In: *arXiv preprint arXiv:1505.00855* (2015).

[176] B Sankar et al. “Transforming Pixels into a Masterpiece: AI-Powered Art Restoration using a Novel Distributed Denoising CNN (DDCNN)”. In: *2023 International Conference on Emerging Techniques in Computational Intelligence (ICETCI)*. IEEE. 2023.

[177] Jorge L. C. Sanz, Fritz Merkle, and Kwan Y. Wong. “Automated digital visual inspection with dark-field microscopy”. In: *J. Opt. Soc. Am. A* 2.11 (1985), pp. 1857–1862. DOI: 10.1364/JOSAA.2.001857. URL: http://opg.optica.org/josaa/abstract.cfm?URI=josaa-2-11-1857.

[178] Jorge LC Sanz, Fritz Merkle, and Kwan Y Wong. “Automated digital visual inspection with dark-field microscopy”. In: *Journal of the Optical Society of America A* 2.11 (1985), pp. 1857–1862.

[179] Thomas Schlegl et al. “f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks”. In: *Medical image analysis* 54 (2019), pp. 30–44.

[180] Thomas Schlegl et al. “Unsupervised anomaly detection with generative adversarial networks to guide marker discovery”. In: *International conference on information processing in medical imaging*. Springer. 2017.

[181] Jianbo Shi and Jitendra Malik. “Normalized cuts and image segmentation”. In: *IEEE Transactions on pattern analysis and machine intelligence* 22.8 (2000), pp. 888–905.

[182] Shot On What. *Celluloid Acquisition*. Accessed: 2023-03-19. 2023. URL: https://shotonwhat.com/acquisition/celluloid-acquisition.

[183] Eliana Siotto and Paolo Cignoni. “Digital methods and techniques for reconstructing and visualizing ancient 3D polychromy–An overview”. In: *Journal of Cultural Heritage* 68 (2024), pp. 59–85.

[184] Jiaming Song, Chenlin Meng, and Stefano Ermon. *Denoising Diffusion Implicit Models*. 2021.

[185] Jouwon Song et al. “AnoSeg: Anomaly segmentation network using self-supervised learning”. In: *arXiv preprint arXiv:2110.03396* (2021).

[186] *Staff/FAQ*. https://www.film-foundation.org/faq. Retrieved 22 August 2020. The Film Foundation.

[187] Donald J. Stavely et al. *Film scanner with dust and scratch correction by use of dark-field illumination*. (Google Patents Entry) //patents.google.com/patent/US5969372A/. Oct. 1999.

[188] Philip Steadman. *Vermeer's camera: uncovering the truth behind the masterpieces*. Oxford University Press, 2002.

[189] Joyce Hill Stoner. "Changing approaches in art conservation: 1925 to the present". In: *Scientific Examination of Art: Modern Techniques in Conservation and Analysis* (2005).

[190] Gjorgji Strezoski and Marcel Worring. "Omniart: a large-scale artistic benchmark". In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14.4 (2018), pp. 1–21.

[191] David Strubel, Blanchon Marc, and David Fofi. "Deep learning approach for artefacts correction on photographic films". In: *Fourteenth International Conference on Quality Control by Artificial Vision*. Ed. by Cudel et al. Vol. 11172. Proceedings of SPIE UNSP 111720M. Mulhouse, France: SPIE-INT SOC OPTICAL ENGINEERING, 1000 20TH ST, PO BOX 10, BELLINGHAM, WA 98227-0010 USA, May 2019, p. 35. DOI: 10.1117/12.2521421. URL: https://hal-univ-bourgogne.archives-ouvertes.fr/hal-02369128.

[192] Roman Suvorov et al. "Resolution-robust large mask inpainting with fourier convolutions". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022, pp. 2149–2159.

[193] Hossein Talebi and Peyman Milanfar. "Learning to Resize Images for Computer Vision Tasks". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 497–506.

[194] Luming Tang et al. "Emergent correspondence from image diffusion". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 1363–1389.

[195] The New York Times. "Vermeer Painting Is Stolen in Brussels". In: *The New York Times* (Sept. 1971). Accessed: 2025-03-10. URL: https://www.nytimes.com/1971/09/25/archives/vermeer-painting-is-stolen-in-brussels.html.

[196] The US National Archives. *Documerica Project*. (Digitised Flickr Collection), https://www.flickr.com/photos/usnationalarchives/collections/72157620729903309/. 1977.

[197] Junjiao Tian et al. "Diffuse, Attend, and Segment: Unsupervised Zero-Shot Segmentation using Stable Diffusion". In: *CVPR* (2024).

[198] Jacopo Tintoretto. *Paradise*. Considered one of the largest paintings on canvas in the world. Venice, Italy, 1588–1599. URL: https://dogespalacevenice.com/tintoretto-paradise/ (visited on 03/18/2025).

[199] Carlo Tomasi and Roberto Manduchi. "Bilateral filtering for gray and color images". In: *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*. IEEE. 1998, pp. 839–846.

[200] Nontawat Tritrong, Pitchaporn Rewatbowornwong, and Supasorn Suwajanakorn. "Repurposing gans for one-shot semantic part segmentation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 4475–4485.

[201] Narek Tumanyan et al. "Splicing vit features for semantic appearance transfer". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10748–10757.

[202] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. "Deep image prior". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 9446–9454.

[203] Ziyu Wan et al. "Bringing Old Films Back to Life". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 17694–17703.

[204] Ziyu Wan et al. "Bringing old photos back to life". In: *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 2747–2757.

[205] Jiabing Wang et al. "A graph clustering algorithm based on minimum and normalized cut". In: *Computational Science–ICCS 2007: 7th International Conference, Beijing, China, May 27-30, 2007, Proceedings, Part I 7*. Springer. 2007, pp. 497–504.

[206] Xintao Wang et al. "ESRGAN: Enhanced super-resolution generative adversarial networks". In: *The European Conference on Computer Vision Workshops (ECCVW)*. Sept. 2018.

[207] Xudong Wang et al. "Cut and learn for unsupervised object detection and instance segmentation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 3124–3134.

[208] Yangtao Wang et al. "Tokencut: Segmenting objects in images and videos with self-supervised transformer and normalized cut". In: *IEEE transactions on pattern analysis and machine intelligence* (2023).

[209] Yi Wang et al. "Vcnet: A robust approach to blind image inpainting". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*. Springer. 2020.

[210] Yinhuai Wang, Jiwen Yu, and Jian Zhang. "Zero-Shot Image Restoration Using Denoising Diffusion Null-Space Model". In: *The Eleventh International Conference on Learning Representations* (2023).

[211] Zhendong Wang et al. “Uformer: A General U-Shaped Transformer for Image Restoration”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022), pp. 1–18.

[212] Zhou Wang et al. “Image quality assessment: from error visibility to structural similarity”. In: *IEEE Transactions on Image Processing* 13.4 (2004), pp. 600–612. DOI: 10.1109/TIP.2003.819861.

[213] Simon Welker, Henry N Chapman, and Timo Gerkmann. “DriftRec: Adapting diffusion models to blind JPEG restoration”. In: *IEEE transactions on image processing* (2024).

[214] R White and C Burrows. “The HST spherical aberration and its effects on images”. In: *The Restoration of HST Images and Spectra* (1991), p. 2.

[215] Weijia Wu et al. “Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 1206–1217.

[216] Julian Wyatt et al. “Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 650–656.

[217] Tete Xiao et al. “Unified perceptual parsing for scene understanding”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018.

[218] Enze Xie et al. “SegFormer: Simple and efficient design for semantic segmentation with transformers”. In: *Advances in Neural Information Processing Systems* 34 (2021).

[219] Zishan Xu et al. “A comprehensive dataset for digital restoration of Dunhuang murals”. In: *Scientific Data* 11.1 (2024), p. 955.

[220] Jie Yang, Yong Shi, and Zhiquan Qi. “Dfr: Deep feature reconstruction for unsupervised anomaly segmentation”. In: *arXiv preprint arXiv:2012.07122* (2020).

[221] Peiqing Yang et al. “PGDiff: Guiding Diffusion Models for Versatile Face Restoration via Partial Guidance”. In: *NeurIPS*. 2023.

[222] Xincheng Yao et al. “Focus the discrepancy: Intra-and inter-correlation learning for image anomaly detection”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 6803–6813.

[223] Jiahui Yu et al. “Free-form image inpainting with gated convolution”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019.

[224] Jiahui Yu et al. “Generative image inpainting with contextual attention”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 5505–5514.

[225] Ke Yu et al. “Crafting a Toolchain for Image Restoration by Deep Reinforcement Learning”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 2443–2452.

[226] Ke Yu et al. “Deep convolution networks for compression artifacts reduction”. In: *arXiv preprint arXiv:1608.02778* (2016).

[227] Quan Yuan et al. “Automatic recognition of craquelure and paint loss on polychrome paintings of the Palace Museum using improved U-Net”. In: *Heritage Science* 11.1 (2023).

[228] Syed Waqas Zamir et al. “Multi-stage progressive image restoration”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 14821–14831.

[229] Syed Waqas Zamir et al. “Restormer: Efficient Transformer for High-Resolution Image Restoration”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 5728–5739.

[230] Syed Waqas Zamir et al. “Restormer: Efficient Transformer for High-Resolution Image Restoration”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 5728–5739.

[231] Lujun Zhai et al. “A comprehensive review of deep learning-based real-world image restoration”. In: *IEEE Access* 11 (2023), pp. 21049–21067.

[232] Hui Zhang et al. “DiffusionAD: Norm-guided one-step denoising diffusion for anomaly detection”. In: *arXiv preprint arXiv:2303.08730* (2023).

[233] Junyi Zhang et al. “A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence”. In: *Advances in Neural Information Processing Systems* 36 (2024).

[234] Kai Zhang et al. “Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising”. In: *IEEE Transactions on Image Processing* 26.7 (2017), pp. 3142–3155.

[235] Kai Zhang et al. “Plug-and-Play Image Restoration with Deep Denoiser Prior”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.10 (2021), pp. 6360–6376.

[236] Richard Zhang, Phillip Isola, and Alexei A Efros. “Colorful image colorization”. In: *European conference on computer vision*. Springer. 2016, pp. 649–666.

[237] Richard Zhang et al. “The Unreasonable Effectiveness of Deep Features as a Perceptual Metric”. In: *CVPR*. 2018.

[238] Xiao Zhang, David Yunis, and Michael Maire. "Deciphering'What'and'Where'Visual Pathways from Spectral Clustering of Layer-Distributed Neural Representations". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 4165–4175.

[239] Shengyu Zhao et al. "Large scale image completion via co-modulated generative adversarial networks". In: *arXiv preprint arXiv:2103.10428* (2021).

[240] Bolei Zhou et al. "Scene parsing through ade20k dataset". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 633–641.

[241] Qihang Zhou et al. "Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection". In: *arXiv preprint arXiv:2310.18961* (2023).

[242] Yang Zou et al. "Spot-the-difference self-supervised pre-training for anomaly detection and segmentation". In: *European Conference on Computer Vision*. Springer. 2022, pp. 392–408.