RTeF CT: Benchmarking Segmentation Models on Diverse nalogue Media Damage

Daniela Ivanova
University of Glasgow, UK

Marco Aversa
Dotphoton, Switzerland

Paul Henderson
University of Glasgow, UK

John Williamson
University of Glasgow, UK

bstract

ccurately 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 RTeF CT, 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/ RTeF CT/ as the first-of-its-kind benchmark for analogue media damage detection and restoration.*

1. 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 [4, 7, 22]. 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 [6]. Previous work has only targeted one type of analogue media at a time and sourced all of their data from the same place [4, 30, 35, 58]. 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. We classify the images into 10 material categories to represent how damage manifests across different media based on material properties. Additionally, we use 4 content categories 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.

Figure 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.

- 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.

2. 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 [48], complicating conservators' decisions on restoration scope and methods [37]. 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 [1, 37]. However, this also necessitates skilled use of specialised software by restoration professionals [6–8]. 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 [7] categorizes such damage broadly into two groups: chemical degradations and mechanical degradations. While this classification found in Chambah's work pertains to film emulsion [6, 8], similar ontologies are commonly applied in the field of mathematical modelling of painting damage [4, 16, 17, 39].

Image restoration and segmentation. The task of image restoration typically involves reversing a degradation process x applied to the undamaged image I to produce the damaged image I . Common restoration tasks include denoising [9, 27, 28, 59, 60], superresolution [23, 28, 51], and colourisation [20, 26, 45, 61]. 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 [4, 6, 13, 16, 21, 22, 36, 47, 50].

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 [18, 40], GANs [57], and diffusion models [33,44,45,53,56]. Stable Diffusion [44] 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 [3, 19, 52] or U-Nets [22, 36]. State-of-the-art architectures like UPerNet [54] and SegFormer [55] could be trained for damage segmentation if annotations are available. Self-supervised models such as DINO [5], DINOv2 [38], SAM [25], and CLIP [42] 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 [44] have inspired methods such as DiffEdit [14], which uses pre-trained diffusion models to detect damaged areas based on text prompts, followed by inpainting. DiffSeg [49], 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 [15, 46] and semantic segmentation [11] in cultural heritage. More recent datasets pair artwork images with textual descriptions [2, 12], and even leverage text conditioning to synthesise artwork variants via diffusion models [10]. 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 [13, 47], illuminated manuscript miniatures [4], film emulsion scans [22], frescoes [35].

3. Damaged nalogue 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.

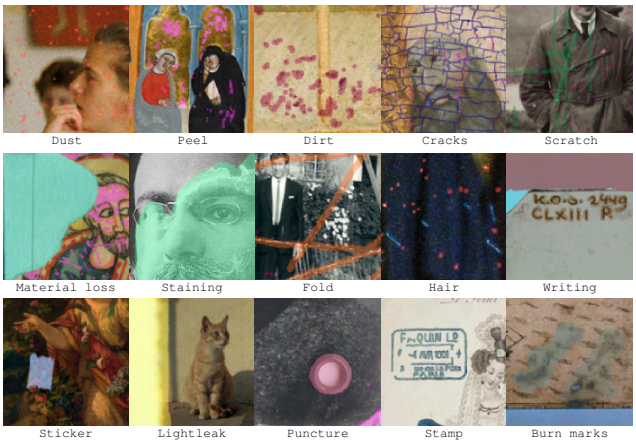


Figure 2. Damage types found in our dataset, demonstrating the variety of shape, scale and severity of damage.

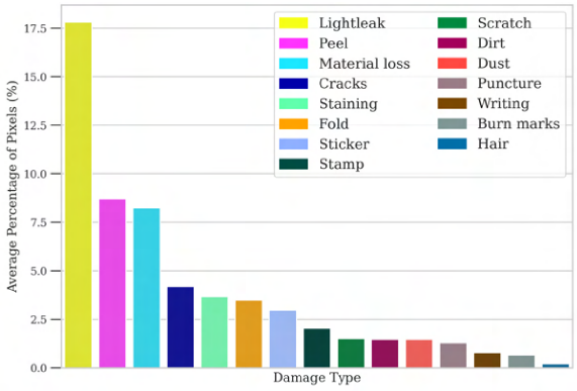
3.1. nnotation methodology

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 in the Supplementary.

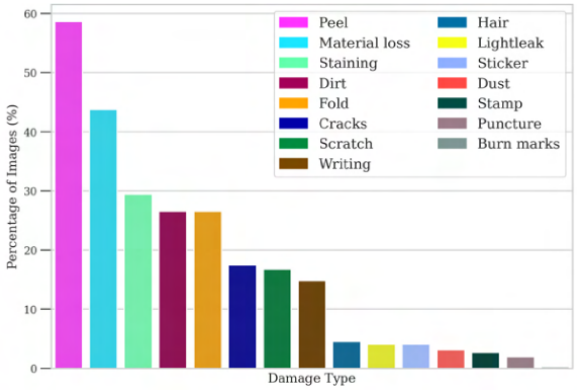
3.2. Damage Types

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



(a) Severity of Damage Types in Affected Images



(b) Prevalence of Damage Types in the Dataset

Figure 3. Overview of the prevalence and severity of different damage types in the Dataset.

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 3 shows the distribution and prevalence of each type of damage within the dataset. Figure 3a considers the extent of damage as a percentage of pixels covered. In Figure 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.

3.3. Image-level nnotations

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 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 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 [29]. 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 2. in the Supplementary.

4. 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.

Figure 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.

Table 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, but those do not need to be provided a priori.

Test Class	Point per class				BBox per class				Point per instance				BBox per instance				Point Grid				Grounding S M			
	cc	F1	mIoU	N	cc	F1	mIoU	N	cc	F1	mIoU	N	cc	F1	mIoU	N	cc	F1	mIoU	N	cc	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 rt	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
rt 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

4.1. Zero-shot Segmentation with S M

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.

4.1.1. 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 [43], 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 1.

Table 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.

Test Class	Segformer			UPerNet + Swin			UPerNet + ConvNeXt			DINOv2 + MLP			S M ViT-H + MLP		
	cc	F1	mIoU	cc	F1	mIoU	cc	F1	mIoU	cc	F1	mIoU	cc	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 rt	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
rt 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

4.2. 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 (multi-class).

4.2.1. 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.

4.2.2. 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** [54] 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** [32] backbone, and a transformer variant, using **Swin Transformer** [31].
- **SegFormer** [55] utilises a hierarchically structured Transformer encoder which outputs multiscale features, and combines those with a lightweight Multi-layer Perceptron (MLP).
- **DINOv2** [38] 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.
- **S M** [25] 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 [62] 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 [24], using the best learning rates and weight decay from the respective papers.

4.2.3. Data ugmentation

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 [41] before calculating the metrics.

Table 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			S M ViT-H + MLP		
	cc	F1	mIoU	cc	F1	mIoU	cc	F1	mIoU	cc	F1	mIoU	cc	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 rt	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
rt 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

4.3. Diffusion-based segmentation

In addition to the supervised methods, we also benchmark two recent diffusion-based segmentation approaches. Following DiffEdit [14], 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 [44]. For our second approach, we benchmark DiffSeg [49], which is an unsupervised segmentation model relying on self-attention features extracted from Stable Diffusion; we set the KL-divergence threshold to 5 in order to produce segments of higher granularity, and use an oracle to assign each predicted segment to the Damaged or Clean class, same as with SAM in 4.1.1. 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.

4.4. 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.

S M is missing the point. We found that SAM can produce good segmentation results only *if* prompted correctly (Figure 4), but doing so requires an impractical number of prompts. As shown in Table 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 doesn't guarantee all artefacts, especially fine ones, are sampled. The only method that provides adequate segmentation - achieving the highest average F1 score of 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 5 top row, but struggle with segmenting and classifying damage types in the multiclass setting for unseen contents or materials, regardless of pretraining or architecture, visualised in Figure 5 bottom row. Table 2 summarizes the binary segmentation results, where no model achieves an F1 score over 7. Notably, DINOv2 outperforms SAM despite being a smaller model. In the multiclass setting (Table 3), all models perform poorly, with no F1 score exceeding 3. Damage missclassifications are detailed in Section 3.1. of the Supplementary. DINOv2 performs best, likely due to its semantically meaningful features [38], aiding in differentiating damage types. SAM, however, struggles to distinguish between damage types and

Figure 5. Qualitative comparison for supervised binary (top row) and multiclass (bottom row) damage segmentation.

Table 4. Results across both Material and Content splits diffusion based text-guided (DiffEdit) and unsupervised (DiffSeg) binary damage segmentation.

Test class	DiffEdit			DiffSeg		
	cc	**F1**	**mIoU**	**cc**	**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 rt	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
rt Depiction	0.79	0.13	0.07	0.93	0.30	0.24

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.

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, 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 [34]. 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.

5. 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.

cknowledgements

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/R513222/1].

References

[1] J.H. Beck. *Reversibility, Fact Or Fiction?: The Dangers of rt Restoration*. Ars Brevis Foundation, 1999. 2

[2] Tibor Bleidt, Sedigheh Eslami, and Gerard de Melo. Artquest: Countering hidden language biases in artvqa. In *Proceedings of the IEEE/CVF Winter Conference on pplications of Computer Vision (W CV)*, pages 7326–7335, 2024. 3

[3] Nian Cai, Zhenghang Su, Zhineng Lin, Han Wang, Zhijing Yang, and Bingo Wing-Kuen Ling. Blind inpainting using the fully convolutional neural network. *The Visual Computer*, 33, 2017. 3

[4] Luca Calatroni, Marie d'Autume, Rob Hocking, Stella Panayotova, Simone Parisotto, Paola Ricciardi, and Carola-Bibiane Schönlieb. Unveiling the invisible: mathematical methods for restoring and interpreting illuminated manuscripts. *Heritage science*, 6, 2018. 1, 2, 3

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 3

[6] M. Chambah. Digital film restoration and image quality. In *IC -BELGIUM Colour Symposium*, Ghent, Belgium, 2019. 1, 2

[7] Majed Chambah, Christophe Saint-Jean, and Francois Helt. Image quality evaluation in the field of digital film restoration. *Proceedings of SPIE - The International Society for Optical Engineering*, 5668, 2005. 1, 2

[8] M. Chambah, C. Saint-Jean, F. Helt, and A. Rizzi. Further image quality assessment in digital film restoration. In Luke C. Cui and Yoichi Miyake, editors, *Image Quality and System Performance III*, volume 6059. International Society for Optics and Photonics, SPIE, 2006. 2

[9] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. *arXiv preprint arXiv:2204.04676*, 2022. 2

[10] Dario Cioni, Lorenzo Berlincioni, Federico Becattini, and Alberto Del Bimbo. Diffusion based augmentation for captioning and retrieval in cultural heritage. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3

[11] Nadav Cohen, Yael Newman, and Ariel Shamir. Semantic segmentation in art paintings. In *Computer graphics forum*, volume 41. Wiley Online Library, 2022. 3

[12] 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. 3

[13] Bruno Cornelis, Tijana Ružić, Emile Gezels, Ann Dooms, Aleksandra Pižurica, LJMMI Platiša, Jan Cornelis, Maximiliaan Martens, Marc De Mey, and Ingrid Daubechies. Crack detection and inpainting for virtual restoration of paintings: The case of the ghent altarpiece. *Signal Processing*, 93(3), 2013. 2, 3

[14] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 3, 7

[15] Elliot J. Crowley and Andrew Zisserman. The state of the art: Object retrieval in paintings using discriminative regions. In *British Machine Vision Conference*, 2014. 3

[16] Rui Dang, Baoping Wang, Xiangyang Song, Fenghui Zhang, and Gang Liu. The mathematical expression of damage law of museum lighting on dyed artworks. *Scientific Reports*, 11(1), 2021. 2

[17] P De Willigen. *mathematical study on craquelure and other mechanical damage in paintings*. Delft University Press, 1999. 2

[18] Cusuh Ham, Amit Raj, Vincent Cartillier, and Irfan Essa. Variational image inpainting. In *NeurIPS 2018 Workshops*, 2018. 3

[19] Amir Hertz, Sharon Fogel, Rana Hanocka, Raja Giryes, and Daniel Cohen-Or. Blind visual motif removal from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 3

[20] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 2

[21] 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 pplications - Volume 4: VIS PP,*. INSTICC, SciTePress, 2022. 2

[22] Daniela Ivanova, John Williamson, and Paul Henderson. Simulating analogue film damage to analyse and improve artefact restoration on high-resolution scans. *Computer Graphics Forum (Proceedings of Eurographics 2023)*, 42(2), 2023. 1, 2, 3

[23] 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. 2

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 3, 6

[26] Manoj Kumar, Dirk Weissenborn, and Nal Kalchbrenner. Colorization transformer. *arXiv preprint arXiv:2102.04432*, 2021. 2

[27] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2Noise: Learning image restoration without clean data. In Jennifer Dy and Andreas Krause, editors, *Proceedings*

of the 35th International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*. PMLR, 2018. 2

[28] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2

[29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 4

[30] L Liu, E Catelli, A Katsaggelos, G Sciutto, R Mazzeo, M Milanic, J Stergar, S Prati, and M Walton. Digital restoration of colour cinematic films using imaging spectroscopy and machine learning. *Scientific reports*, 12(1), 2022. 1

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021. 6

[32] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022. 6

[33] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3

[34] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *dvances in Neural Information Processing Systems*, 36, 2024. 8

[35] Fabio Merizzi, Perrine Saillard, Oceane Acquier, Elena Morotti, Elena Loli Piccolomini, Luca Calatroni, and Rosa Maria Dessì. Deep image prior inpainting of ancient frescoes in the mediterranean alpine arc. *Heritage Science*, 12(1), 2024. 1, 3

[36] 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 herit ge Contents*, 2020. 2, 3

[37] Andrew Oddy and Sara Carroll. Reversibility–does it exist? *Occasional Paper Number 135*, 1999. 2

[38] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3, 6, 7

[39] Ludovic Pauchard and Frédérique Giorgiutti-Dauphiné. Craquelures and pictorial matter. *Journal of Cultural Heritage*, 46, 2020. 2

[40] Jialun Peng, Dong Liu, Songcen Xu, and Houqiang Li. Generating diverse structure for image inpainting with hierarchical vq-vae. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 3

[41] Nicolas Pielawski and Carolina Wählby. Introducing hann windows for reducing edge-effects in patch-based image segmentation. *PloS one*, 15(3), 2020. 6

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 3

[43] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 5

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 3, 7

[45] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *CM SIGGR PH 2022 Conference Proceedings*, 2022. 2, 3

[46] Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv preprint arXiv:1505.00855*, 2015. 3

[47] B Sankar, Mukil Saravanan, Kalaivanan Kumar, and Siri Dubakka. 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. 2, 3

[48] Joyce Hill Stoner. Changing approaches in art conservation: 1925 to the present. *Scientific Examination of rt: Modern Techniques in Conservation and nalysis*, 2005. 2

[49] Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse attend and segment: Unsupervised zero-shot segmentation using stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3, 7

[50] Ziyu Wan, Bo Zhang, Dongdong Chen, Pan Zhang, Dong Chen, Jing Liao, and Fang Wen. Bringing old photos back to life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2

[51] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, 2018. 2

[52] Yi Wang, Ying-Cong Chen, Xin Tao, and Jiaya Jia. VCNet: A robust approach to blind image inpainting. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[53] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. *The Eleventh International Conference on Learning Representations*, 2023. 3

[54] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, 2018. 3, 6

[55] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transform-

ers. *dvances in Neural Information Processing Systems*, 34, 2021. 3, 6

[56] Peiqing Yang, Shangchen Zhou, Qingyi Tao, and Chen Change Loy. PGDiff: Guiding diffusion models for versatile face restoration via partial guidance. In *NeurIPS*, 2023. 3

[57] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 3

[58] Quan Yuan, Xiang He, Xiangna Han, and Hong Guo. Automatic recognition of craquelure and paint loss on polychrome paintings of the palace museum using improved u-net. *Heritage Science*, 11(1), 2023. 1

[59] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5728–5739, 2022. 2

[60] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021. 2

[61] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*. Springer, 2016. 2

[62] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 6