Progress Report I: Film damage restoration using diffusion with temporal bias

Pol Riubrogent Comas
Supervisor: Ramon Baldrich Caselles

10/03/2025

1 Introduction

The aim of this final project is to explore the field of image restoration, specifically focusing on old scans of film reels and slides. Image restoration plays a crucial role in preserving and reviving historical media, as many archival films suffer from various forms of degradation over time. These degradations include physical damage such as scratches and dust, and other artifacts introduced during storage.

Traditional film restoration relies heavily on manual labor, with experts cleaning frames individually or using rule-based automated processes. While these methods have been effective in the past, they often require weeks or months to restore just reel of a film. Deep learning and diffusion-based models offer a promising alternative by learning to reconstruct damaged sections while preserving the integrity of the original material. Generative models have demonstrated their ability to recover missing information in images, making them a suitable approach for film restoration.

This project aims to develop a deep learning-based solution tailored to the specific characteristics of 1960s 35mm and 16mm film scans, as well as home video formats from the same period. The goal is to create a model that not only removes dirt and scratches but also ensures that the restored images retain their original texture and grain. To achieve this, the project will leverage state-of-the-art diffusion models and explore ways to condition them on temporal information from adjacent frames, allowing for more coherent and context-aware restorations.

Ultimately, this project seeks to bridge the gap between traditional restoration techniques and modern AI-driven approaches, providing a tool that is both effective and accessible. In addition to developing and evaluating restoration models, the project will also focus on usability, designing an intuitive interface that allows users to visualize and interact with the restoration pipeline easily. Through these efforts, the project aims to contribute to the growing field of AI-assisted film preservation, offering a solution that can be extended to other film formats and historical archives in the future.

2 Objectives

ID	Task	Priority
O1	Propose a solution to **restore damaged scanned film reels.** The solution has to be able to restore damage like dirt or scratches on the film, but keep the characteristic grain of film images and videos. The solution will focus specifically on 1960s 35/16mm movie scans, as well as home videos of the same decade.	**Main objective**
O2	Obtain a usable dataset to train restoration models based on common and real film damage.	**Essential**
O3	Said solution has to be easy to use, as well as visually appealing.	**Essential**
O4	Generalize the model as to be able to restore any type of film scans, not only the ones presented on O1.	**Not essential**

Table 1: Summary of the objectives defining the project

2.1 Tasks

ID	Task	Objective
T1	**Explore dataset options**. Research different resources to be used as dataset (ground truth or testing). By the end of this task there should be a trainable dataset.	*O2*
T2	**Create synthetic ground truth dataset** using real scanned film damage.	*O2*
T3	**Segment** the damaged parts of a frame, without prior knowledge of said film. The proposed model should be able to segment all damaged parts of the film.	*O1*
T4	Propose a model to segment the damaged parts of a frame **having context of other frames** to bias the segmentation model.	*O1*
T5	Research about inpainting models and implement one to start testing the specific use case.	*O1*
T5	Modify an existing inpainting diffusion model in order to, providing context of other frames as a prompt, bias the generation to better adequate said generation to the ground truth.	*O1*
T6	Explore different inpainting architectures to compare performance with the original chosen one.	*O1*
T7	Implement a **graphical user interface** in order to provide an easy and catchy representation of all the parts of the final pipeline and showcase the results of the project in a tidy and usable way.	*O3*

Table 2: Summary of the tasks defining the project

3 Methodology

This project will be developed by what in software development would be called agile development. This type of development is characterized with short work cycles, with predefined objectives (tickets) that have a clear deliverable in mind. For each work cycle, the objectives, as well as the deliverables, will be predefined in this initial report. Subsequently in the following Reports of Progress, a small report shall be written for each work cycle, detailing weather the objectives and deliverables have been met, with a pertinent reasoning in case of failure to do so. These work cycles will be marked by a weekly meeting with the tutor of the project, however this may not coincide with each start of work cycle, as the meeting schedule will be adjusted on a weekly basis.

4 State of the Art

Diffusion-based models have emerged as a powerful tool in image generation. In the defined task for this project, image generation is a key aspect, since in order to restore an image, we need to generate the missing data from the image. Existing solutions include papers like DiffIR [10].

4.1 DiffIR

DiffIR (Diffusion-based Image Restoration) introduces a diffusion-based image restoration pipeline, outperforming traditional CNNs by effectively handling various degradations, such as noise, blur and compression artifacts. To effectively achive this, DiffIR proposes a solution consisting of a compact image restoration prior extraction network (CPEN), which extracts a compact image restoration representation which encapsulates relevant priors for the restoration, a dynamic IR transformer (DIRformer), which restores low quality images using the prior representation given by the CPEN, and a de-noising network.

4.2 U-Net

U-Net is a convolutional neural network architecture designed for precise biomedical image segmentation [9]. It uses a symmetric architecture consisting of an encoder and a decoder. The key idea is to combine abstract features with fine-grained spatial information by introducing skip connections between corresponding kayers of the encoder and decoder. This allows the network to retain high-resolution information lost during downsampling, improving segmentation accuracy, especially around object boundaries. This is crucial for this project's application, since most artifacts needed to be detected may be around 10 pixels wide.

Attention U-Net

Attention U-Net builds upon the original U-Net architecture by integration attention gates into the skip connections, enabling the network to automatically learn *where to focus* in an input image. These attention gates allow the model to suppress irrelevant regions while enhancing features that are useful for the segmentation task [8].

This mechanism is especifically valuable for this project since many of the artifacts to be detected are very small, and may be easily overlooked. In these cases, irrelevant background activations may interfere with the detection of fine details.

Experimental results in the original paper show that Attention U-Net outperforms the standard U-Net in abdominal CT segmentation, particularly in terms of recall and surface accuracy, even when trained with fewer examples. This suggests that AttU-Net is able to generalize better and maintain performance with limited data, an important consideration in this restoration task where ground-truth annotations can be scarce.

RU-Net

RU-Net, or Recurrent U-Net, extends the original U-Net architecture by introducing recurrent convolutional layers (RCLs) within both the encoder and decoder paths [2]. These layers allow the network to refine its feature representations over discrete time steps, effectively increasing the network's depth without increasing the number of parameters.

This recurrent mechanism is particularly beneficial in segmentation tasks where capturing fine structures and contextual dependencies is crucial—conditions that closely align with the goals of this project, especially when dealing with temporally and spatially coherent image artifacts across degraded film frames. By allowing iterative feature accumulation, RU-Net can better delineate small or subtle artifacts that may otherwise be obscured.

Experimental results show RU-Net consistently outperforms standard U-Net and ResU-Net across multiple medical image segmentation benchmarks, including retina vessel segmentation and skin lesion detection, with superior performance in sensitivity and Dice scores even with fewer parameters. This makes RU-Net especially suitable for restoration scenarios where training data is limited, and model generalization is paramount.

R2U-Net

R2U-Net combines the strengths of both residual learning and recurrent convolutional operations by embedding residual connections into the recurrent convolutional layers of RU-Net. This architecture—Recurrent Residual U-Net—enables efficient gradient flow and stable training of deeper models, while further enhancing feature refinement through temporal recurrence [2].

For the task of fine-grained artifact segmentation in degraded films, R2U-Net's architecture is highly advantageous. It not only captures spatial and contextual details over multiple iterations, but the residual connections help prevent vanishing gradients, making the model robust even on small datasets. The ability to extract intricate, low-level features makes R2U-Net well-suited for identifying and segmenting defects such as scratches, dirt, and frame tears in scanned analog films.

R2U-Net achieved top performance across several datasets used in the original study, including STARE and DRIVE, with the highest AUC and Dice coefficients among tested models. This demonstrates its strong generalization and segmentation fidelity, aligning perfectly with the challenges of restoration where target regions are often irregular and dispersed.

4.3 RePaint

RePaint is a state-of-the-art inpainting approach based on Denoising Diffusion Probabilistic Models (DDPMs), designed to fill in missing or corrupted regions of an image in a semantically meaningful and visually consistent manner [**?**]. Unlike conventional methods that are typically trained on specific mask types, RePaint conditions a pretrained unconditional diffusion model through a clever modification of the reverse diffusion process. This conditioning strategy enables it to generalize to arbitrary mask shapes without the need for task-specific retraining.

For this restoration project, RePaint offers several compelling advantages. Many of the image degradations in analog film—scratches, tears, dust spots—are irregular, sparsely located, and vary in scale. RePaint is uniquely capable of addressing such free-form degradation due to its mask-agnostic formulation. Moreover, when used in conjunction with a segmentation model like Attention U-Net (to localize defects), RePaint can be applied only to damaged areas, preserving the integrity of the undistorted regions.

The iterative nature of DDPMs allows RePaint to generate harmonized and semantically coherent content, using a resampling strategy that alternates between denoising and re-noising. This enhances integration between restored and original regions—particularly useful for subtle textures and complex scenes in historic footage.

Experimental results in the original paper demonstrate that RePaint outperforms both GAN-based and autoregressive inpainting models across diverse datasets and challenging mask settings. It produces visually plausible reconstructions even under extreme occlusions and has shown strong results in both perceptual realism (user studies) and learned perceptual similarity (LPIPS). This robustness makes it especially valuable in domains like film restoration, where ground truth is absent and subjective visual quality is paramount.

5 Dataset

Since the focus of the project is restoring film damage while keeping the characteristic grain of developed film, I cannot use the typical image restoration datasets, like LSDIR (reference) since it is focused on a super-resolution restoration, defeating one of the main, self-imposed, restrictions of my project. To achive the objective, I would need a dataset centred around film damage (dirt, hairs, scratches, etc.). Online there is no robust existing dataset for said solution.

The only solution left would be to create my own dataset. There are two methodologies. The first one would be, given original and HQ damaged film scans, to label myself a dataset based on real images. The main problem with this solution is time, since it is very limited for my project. The second solution, which I have chosen, is to create a synthetic dataset using HQ non-damaged film scans. To create the synthetic dataset first I would need HQ isolated film damage to superimpose to the HQ images. FILM-AA [4] provides synthetic damage extracted from 4k scans of damaged film. This solution allows me to create synthetic masks to add to HQ images to create a suitable dataset for the task.

Apart from the already stated benefits of said solution, it gives me the ability to tinker with what films I want to restore, so I can focus on the original objective.

5.1 Synthetic dataset

The dataset used in this project is composed of three distinct parts: a synthetic dataset for training and validation, a pretraining dataset, and a real dataset for qualitative testing. Since the segmentation model requires accurate ground truth masks indicating the location of artifacts, and manual annotation of real-world footage is beyond the scope of this project, all training and validation have been carried out using fully synthetic data that I generated.

This synthetic dataset was created using the publicly available FILM-AA library [4], previously described. It currently consists of over 2,000 high-resolution frames extracted from digitally restored transfers of 35mm films from the 1960s and 1970s. Corresponding to these frames are more than 6,000 synthetic artifact masks, generated using the FILM-AA framework. These masks are randomly assigned and composited onto the film frames at training time, serving as a form of dynamic data augmentation. This strategy ensures both variability and scalability while maintaining tight control over the ground truth, which is crucial for effective supervised learning.

Figure 1: Original clean frame extracted from a restored 35mm copy of IL BUONO, IL BRUTTO, IL CATTIVO (1966)

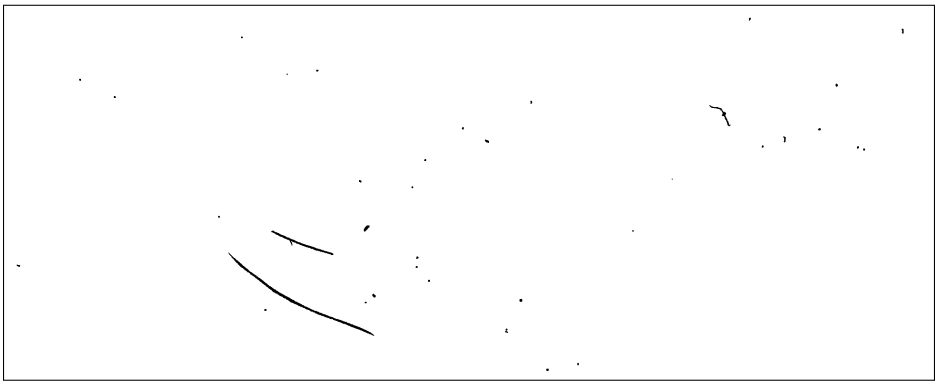

Figure 2: Artefacts mask generated with the custom library

The first dataset generated, was just superimposing the black mask to the image. This gave semi-realistic results, but comparing with real damaged frames, it still seemed fake. To improve the performance of the model, and the realism of the dataset **edge smoothing** has been introduced around the annotated regions. This was achieved by applying a Gaussian blur at the boundaries of the segmented artefacts, thereby mitigating the model's tendency to focus solely

on high-contrast transitions.

Figure 3: Image generated superimposing the mask into the original frame.

Figure 4: Image generated adding gaussian blur around the artefacts in the mask.

5.2 Pretraining dataset

In order to improve convergence and provide a better initialization for the segmentation model, I also leveraged a separate synthetic dataset provided to me through direct communication with the lead author of the FILM-AA paper. This new dataset—created with the same methodology as my primary synthetic dataset—contains over XXX annotated images. However, instead of being based on restored film transfers, these frames were derived from high-resolution scans of 35mm still images from various (not provided) sources.

Although this dataset is not perfectly aligned with the final application domain of degraded motion picture footage, it offers valuable priors on the visual texture and artifact distribution commonly seen in film. As such, it was used for pretraining the segmentation network, which was later fine-tuned on my custom dataset to adapt it to the specific degradation patterns present in cinematic material.

Figure 5: Example image of the Documartica dataset

5.3 Testing dataset

To qualitatively evaluate the performance of the entire restoration pipeline on real-world data, I will use samples from the *Hearst Metrotone News Collection* [1]. This collection, owned by the Regents of the University of California and curated by the UCLA Film & Television Archive since 1981, comprises a vast archive of historical newsreels originally recorded on 35mm film. The archive contains approximately 25 million feet of 35mm footage and an additional 2 million feet of 16mm film, accompanied by rich contextual documentation.

The materials selected for testing have been digitized using a high-resolution, high dynamic range (HDR) film scanner, ensuring the preservation of fine-grained image details and subtle degradations. This allows for a realistic and visually meaningful evaluation of the model's ability to identify and restore artifacts in archival footage.

Figure 6: Example image of the testing dataset

6 Time Schedule

Week	Date	Task Name	Deliverable
1	27/01 - 02/02	Read and research SOTA solutions	-
2	03/02 - 09/02	Explore available online datasets (T1)	Dataset
3	10/02 - 16/02	Implement and create a synthetic dataset (T2)	-
4	17/02 - 23/02	Explore segmentation solutions to detect film damage (T3)	-
5	24/02 - 02/03	T3	Initial segmentation model results
6	03/03 - 09/03	Prepare and redact the initial report delivery	Initial Report
10/03/2025		**DUE INITIAL REPORT**	
7	10/03 - 16/03	Improve the synthetic dataset, T3, T4	Result comparison
8	17/03 - 23/03	Explore inpainting solutions and implement a starting model (T5)	-
9	24/03 - 30/03	T5	Initial inpainting model results
10	31/03 - 06/04	Develop a solution to incorporate temporal bias into the inpainting diffusion model (T6)	-
11	07/04 - 13/04	Prepare a dataset to train *T6* model and train the model	Result comparison
12	14/04 - 20/04	Prepare and redact Progress Report I	Progress Report I
20/04/2025		**DUE PROGRESS REPORT I**	
13	21/04 - 27/04	Explore different architectures for inpainting restoration (T7), T6	-
14	28/04 - 04/05	Train and/or test T7 models	Result comparison
15	05/05 - 11/05	Prepare models for GUI (T8)	Model inference service
16	12/05 - 18/05	T7, T8	Final GUI
17	19/05 - 25/05	Prepare and redact Progress Report II	Progress Report II
25/05/2025		**DUE PROGRESS REPORT II**	

Week	Date	Task Name	Deliverable
18	26/05 - 01/06	Prepare and redact final report proposal	-
19	02/06 - 08/06	Prepare and redact final report proposal	-
20	09/06 - 15/06	Prepare and redact final report proposal	Final report proposal
15/06/2025		**DUE FINAL REPORT PROPOSAL**	
21	16/06 - 22/06	Prepare presentation slides	Presentation
20/06/2025		**DUE PRESENTATION PROPOSAL**	
22	23/06 - 29/06	Prepare final dossier	Dossier
29/06/2025		**DUE FINAL DOSSIER**	

Table 3: Weekly Planning

6.1 Review

In this section, I provide a week-by-week summary of the project's progress, aligned with the scheduled tasks for each period. This overview is meant to give a general sense of the work completed without going into technical or implementation-specific details, which will be covered in the following section. The objective here is to offer a clear and concise account of the project's development over time.

1	27/01 - 02/02	Read and research SOTA solutions	-
The first week focused on reviewing relevant literature and surveying existing solutions related to the project. The primary outcome was the selection of an initial architecture consisting of two core components: a segmentation model and an inpainting model. Specific implementation details will be discussed in the following section.			

2	03/02 - 09/02	Explore available online datasets (T1)	Dataset
The focus this week was identifying a suitable dataset for the task. A general explanation has already been included in the *Initial Report*; for clarity, that explanation remains in this report in blue font (Section 5). Additional information is provided in standard text throughout the document.			

3	10/02 - 16/02	Implement and create a synthetic dataset (T2)	-
This week was dedicated to customizing the FILM-AA library to dynamically generate synthetic datasets using custom images. Additionally, I implemented the necessary classes and functions required to train models using these datasets.			

4	17/02 - 23/02	Explore segmentation solutions to detect film damage (T3)	-

This week focused on identifying a suitable segmentation architecture. After reviewing several papers, I selected U-Net as the baseline model. Although transformer-based segmentation models were considered, they were ultimately discarded. The U-Net implementation was adapted to fit the project's requirements and to simplify training. Code details are documented in the source code via comments and in the *README.md* file of the repository.

5	24/02 - 02/03	T3	Initial segmentation model results

This week continued the work on segmentation. The previously chosen U-Net model was finalized and adapted further. Again, implementation details are documented within the code and in the repository's *README.md*.

7	10/03 - 16/03	Improve the synthetic dataset, T3, T4	Result comparison

Although improvements to the synthetic dataset were planned for this week, much of this work had already been completed during earlier training sessions with the U-Net. This week primarily involved preparing an inference workflow to facilitate the comparison of results across different trained models. Detailed evaluations are presented in the development section.

8	17/03 - 23/03	Explore inpainting solutions and implement a starting model (T5)	-

This week was dedicated to researching diffusion-based inpainting frameworks. I ultimately selected RePaint, as detailed in the State of the Art section. Parallel to this, I continued testing ways to improve the U-Net segmentation models.

9	24/03 - 30/03	T5	Initial inpainting model results

While this week was intended to yield initial results from the inpainting model, most of the time was spent adapting the RePaint implementation, constructing a test dataset, and continuing with pretraining experiments from the previous week. As a result, the planned work extended into the following week.

10	31/03 - 06/04	Develop a solution to incorporate temporal bias into the inpainting diffusion model (T6)	-
11	07/04 - 13/04	Prepare a dataset to train *T6* model and train the model	Result comparison
Unfortunately, I was unable to complete the tasks for these weeks as planned. Week 10 was spent performing test runs with the RePaint model. By week 11, I had achieved end-to-end results for the full pipeline. These outcomes are discussed later in this report.			

7 Development

7.1 Pipeline

Before entering into model implementation, code development and preliminary results, it is essential to establish a clear architectural framework that will guide this project. Given the complexity of film restoration, particularly when working with analog formats such as 35mm film, the architecture must accound for both the visual characteristics of film and the nature of degradations we aim to correct.
To this end, the proposed pipeline is composed of two main stages: segmentation and restoration. First, a U-Net-based segmentator identifies and isolates regions of damage such as dust, scratches and other artifacts. This damage mask is then passed as input to a diffusion-based inpainting model, which reconstructs the missing or corrupted regions while preserving the original film's texture, tone and grain.
This two-step approach allows for more targeted restorations, as the inpainting model can focus solely on the areas marked as damaged, reducing the risk of altering intact portions of the image. Furthermore, the architecture is designed to incorporate temporal information from adjacent frames, helping ensure consistency and coherence across time; Although this has not yet been implemented, it is one of the objectives for the following weeks. By explicitly separating damage detection from content restoration, the architecture remains modular and interpretable, paving the way for further improvements and extensions.

7.2 Segmentation

This section outlines the different stages of experimentation, from initial baselines to advanced pretraining strategies,encompassing architectural experimentation, dataset enhancement, and strategic training techniques. The process is structured into several phases, described below.

Baseline Architectures

The initial experimentation involved evaluating a series of established segmentation architectures:

- The original U-Net architecture [9]
- The Attention U-Net (AttU-Net) [8]
- The Recurrent Residual U-Net (R2U-Net) and R2AttU-Net [2]

These models were trained from scratch on a synthetic dataset consisting of 600 images, created during Week 5. As anticipated, the results were suboptimal, primarily due to the limited size of the dataset. One of the main problems all these current model have is having a tendency to detect high-contrast transitions as artefacts.

Dataset Expansion and Preprocessing Enhancements

To address the limitations posed by data scarcity, a new dataset of 3000 synthetic artefacts masks was generated. This expansion significantly improved model performance. in addition, the dataset was further refined by introducing **edge smoothing** around the annotated regions. This was achieved by applying a Gaussian blur at the boundaries of the segmented artefacts, thereby mitigating the model's tendency to focus solely on high-contrast transitions.

Despite these improvements, the models still exhibited limited generalization capability. One of the hypothesis for these lack of performance was thought to be due to the randomly initialized weights. In the next section this problem will be tackled.

Pretraining Approaches

To enhance both convergence speed and final model performance, a series of pretraining strategies were explored:

- **Denoising Pretraining [3]:**
  A denoising task was employed as a proxy pretraining objective using approximately 10,000 images. This allowed the model to learn useful low-level features prior to fine-tuning on the segmentation task. Initial experiments with the U-Net architecture showed clear qualitative improvements in segmentation results.
- **Transfer of Pretraining Across Architectures:** Upon validating the benefits of denoising pretraining with the U-Net, the same prodecure was extended to AttU-Net, R2U-Nett and R2AttU-Net. All architectures demonstrated improved performance when initialized with pretrained weights.
- **Multi-Stage Pretraining:** In a subsequent refinement, the models were additionally pretrained on the **Documartica dataset**, a corpus generated similarly to the target dataset but lacking task-specific structures (mainly the frames not being from 35mm scanned films). This intermediate step further improved the model's repesentational capacity. Final fine-tuning was performed using the custom dataset of 3000 images.

Model Simplification for Computational Efficiency

To optimize computational efficiency, the number of channels in all layers of the network was reduced by half. This modification led to a significant reduction in training time without compromising the segmentation quality, as evaluated qualitatively.

Final Configuration and Observations

The best qualitative performance was achieved using the **Attention U-Net** architecture, following the full pipeline:

1. Denoising pretraining
2. Intermediate pretraining on the Documartica Dataset
3. Final fine-tuning on the task-specific dataset
4. Reduced-channel model configuration

While formal quantitative metrics are not yet available, the observed improvements where consistent across visual inspections. The implementation of a quantitative evaluation protocol is a key objective for the second phase of this project.

7.3 InPainting with RePaint

RePaint [7] is an inpainting method based on diffusion models that extends the standard denoising diffusion probabilistic models (DDPMs) by introducing a novel sampling strategy. Specifically, it employs a "repainting schedule"—a combination of reverse and forward diffusion steps—to iteratively refine the content generated within masked regions. This process allows RePaint to more effectively reconstruct complex or irregular missing areas compared to traditional single-pass diffusion-based approaches.

In the initial phase of this project, the official RePaint repository was cloned for experimental use. As RePaint only provides an inference framework, a pretrained Stable Diffusion model—trained on a dataset referred to here as DATASET—was integrated for performing the inpainting tasks. The primary goal was to assess how well a general-purpose diffusion model could handle restoration on artificially degraded film frames.

To this end, a set of degraded images was generated, along with corresponding binary masks indicating the corrupted regions. These masks were used directly in the inpainting process, rather than relying on a predicted mask, in this case from the U-Net, to isolate the performance of the inpainting model itself.

Initial experiments revealed that the model struggled to convincingly restore the masked regions, especially when these were large or structurally complex. To address this, a simple morphological operation—dilation—was applied to the binary masks, expanding the masked areas by a few pixels. This modification helped reduce edge artifacts and led to more coherent results around the boundaries. While this significantly improved performance and yielded results that exceeded initial expectations, the model still faced limitations with larger degraded regions, where the reconstructions often lacked structural and semantic consistency.

These observations have shaped the direction for the second phase of the project. The next step involves training a custom diffusion model tailored specifically for film restoration tasks. A key enhancement will be the inclusion of temporal conditioning: the model will be conditioned on adjacent frames from the same film sequence. By incorporating visual context from neighboring frames, the model will be better equipped to produce temporally consistent and contextually accurate inpainting results, particularly in scenes with extensive damage or missing data.

8 Preliminar Results

As previously stated in this report, a suitable metric for evaluating both the segmentation and inpainting models has yet to be identified. I anticipate completing a comprehensive evaluation study by the time the next report is due. In the meantime, since the models have already been trained, I have compiled a set of inference results to enable a qualitative comparison.

8.1 Segmentation experiments

In order to qualitatively assess the performance of the segmentation model, I present several inference examples below. Each result is visualized by overlaying a color-coded mask onto the original frame. The colors represent the following cases:

- **Yellow:** True Positive (Correct detection of an artifact)
- **Green:** False Positive (Incorrect detection where no artifact is present)
- **Red:** False Negative (Missed detection of an existing artifact)

For clarity, Table 4 summarizes the color legend used in the visualizations:

Color	Meaning	Interpretation
Yellow	True Positive	Correctly detected artifact
Green	False Positive	Detected artifact where there is none
Red	False Negative	Missed artifact

Table 4: Color legend for segmentation results

The following subsections detail the results according to the different model settings:

Reference Image

For the purposes of this report, I will focus the qualitative comparison on two representative images. These two have been selected due to their problematic nature prone to false positives. While this report highlights only these two images for clarity and space considerations, it is important to note that a more extensive internal evaluation has been conducted across a wider range of examples to verify the generality of the observations presented here.

The two selected images serve to illustrate how the models behave across varying difficulty levels and will be consistently used throughout the following comparisons.

Figure 7: Test image used in inference. Extracted from the movie A Fistful of Dollars, Sergio Leone 1964, with added synthetic damage.

Figure 8: Test image used in inference. Also extracted from the movie A Fistful of Dollars, Sergio Leone 1964, with added synthetic damage.

Baseline Model

First, I present the results obtained using the baseline U-Net model trained on the original dataset: As it can be seened in the image, the model predicts half of the pixels of the image as artefacts. This was attributed to a bad dataset, since the artifacts are just black spots, so the model learns to detect contrast differences and select that.

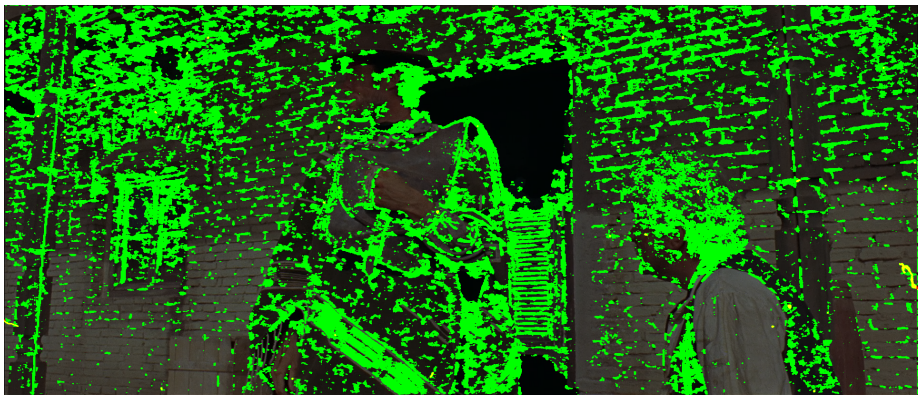

Figure 9: Inference result with the Fig. 8 image, using the baseline U-Net model.

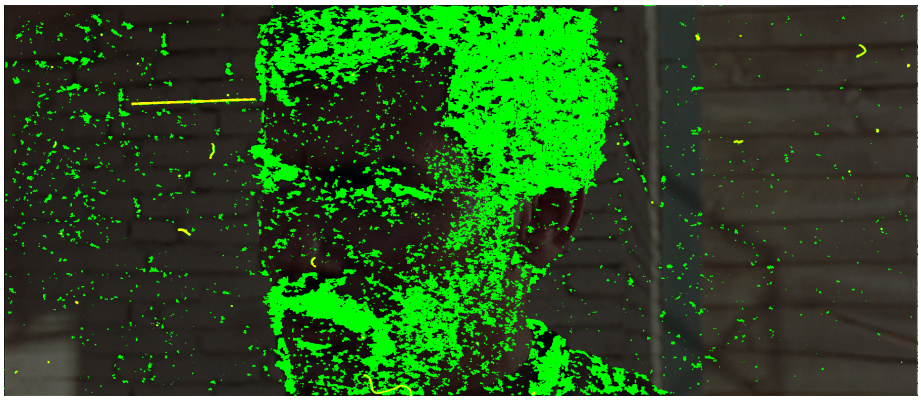

Figure 10: Inference result with the Fig. 7 image, using the baseline U-Net model.

Dataset improvements and new models

After the poor results of the baseline model+baseline dataset, a new dataset was created. Then it was applied not only to a U-Net, but also an AttU-Net and R2AttU-Net. The following images are the result of the first image tested with the U-Net model to compare.

As it can be seen in Fig. 12, the dataset improvements have improved performance by a huge

Figure 11: Inference result with the Fig. 8 image, using the baseline U-Net model with the new perfected dataset.

margin, practically removing the false positives detected in the first iteration.

Figure 12: Inference result with the Fig. 8 image, using the baseline AttU-Net model with the new perfected dataset.

Figure 13: Inference result with the Fig. 8 image, using the baseline R2AttU-Net model with the new perfected dataset.

On the next image, the one trained with the AttU-Net model we can see even better performance than the U-Net model, mostly on the dark regions of the image. Finally we can see that the worst model in this iteration of training is the R2AttU-Net model, missing a lot of the spots other models detect.

Pretraining

Finally the pretrained model was trained. It was only tested the pretrained AttU-Net model, as the AttU-Net was found to be the best performing one yet. As it can be observed mostly in

Figure 14: Inference result with the Fig. 7 image, using the baseline AttU-Net model with the perfected dataset.

Figure 15: Inference result with the Fig. 7 image, using the pretrained AttU-Net model with the perfected dataset.

the hair of the character, the baseline AttU-Net detects more false positives than the pretrained one. this can be observed across the test dataset.

8.2 Inpainting

Unfortunately the current results for the RePaint experiments are poor in number. I've only generated a few examples, even though they are end to end (the mask has been detected with the segmentation model), and the diffusion model is still not a custom model but an already trained one, however the results are already acceptable in this stage of the project. Firstly I tried to run the RePaint inference just with the original mask generated by the segmentation. As it can be seen in the following side-by-side images, it did not change at all, the specs are still there. To fix this, I tried to dilate the mask generated for a few pixels, as the results would not be much affected if it generated a bit more image than it should. As it can be seen in the next comparaisons, it worked perfectly.

Figure 16: Original dirty image.

Figure 17: Result of the RePaint generation

Figure 18: Dirty image with dilated mask.

Figure 19: Result of the RePaint generation

Figure 20: Dirty image with dilated mask.

Figure 21: Result of the RePaint generation

Figure 22: Dirty image with dilated mask.

Figure 23: Result of the RePaint generation

9 Evaluation metrics for segmentation

In order to assess the performance of the binary segmentation model, several standard evaluation metrics were computed using different and significant images. These metrics include: **Precision**, **Recall**, **Accuracy**, **Dice Coefficient** and **Intersection over Union (IoU)**. All these metrics are commonly used in image segmentation tasks, particularly when evaluating models trained to separate two distinct classes (positive class and negative class).

9.1 Metrics

- **Precision**: Quantifies the proportion of correctly predicted positive pixels among all pizels predicted as positive. Low precision indicates a large number of false positives.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall**: Also known as sensitivity, measures the proportion of true positive pixels that were correctly identified by the model. High recall with low precision suggests oversegmentation. Although it is important to have a high rate of true positive pixels, for the current task it is extremely important to not oversegmentate, since it would mean that it would inpaint correct pixels.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **Accuracy**: Overall proportion of correctly classified pixels (both positive and negative). This metric becomes misleading in segmentation tasks with highly imbalance distributions, as is the case in this project. Nevertheless it will be studied and compared anyway.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Dice Coefficient**: Harmonic mean of precision and recall, placing equal emphasis on false positives and false negatives. It is particularly useful for tasks like the one at hand.

$$\text{Dice Coefficient} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

- **Intersection over Union (IoU)**: Ratio of overlap between between the predicted and ground truth masks to their union. It is a more conservative metric than the DICE and penalizes both tipes of error. On paper this seems like the most useful for the task in this project.

$$\text{Intersection over Union (IoU)} = \frac{TP}{TP + FP + FN}$$

9.2 Results

	Precision	Recall	Accuracy	Dice Coefficient	IoU
Perfect Image	1.000	1.000	1.000	1.000	1.000
Good Image	0.790	0.969	**1.000**	0.870	0.770
Poor Image	0.004	**0.990**	**0.824**	0.008	0.004
Image 0 (all negative)	—	0.000	**0.999**	0.000	0.000
Image 1 (all positive)	0.001	**1.000**	0.001	0.001	0.001

Table 5: Evaluation metrics for representative segmentation outputs.

9.3 Result analysis

Perfect Image

This case demonstrates ideal segmentation performance. All predicted pixels exactly match the truth. All evaluation metrics return a value of 1.0 (perfect), as expected.

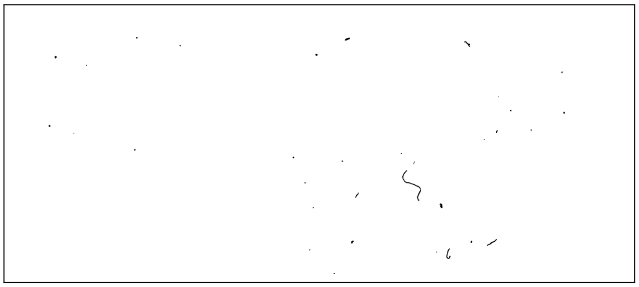

Figure 24: Ground truth image

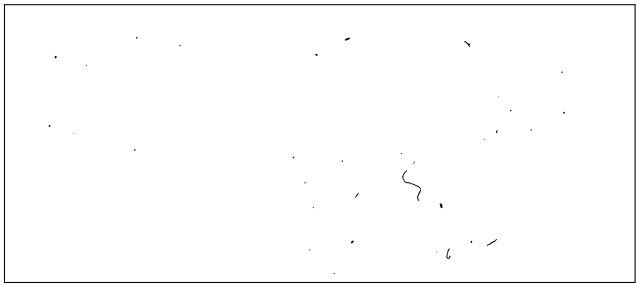

Figure 25: Perfect prediction

Good Image

The prediction closely aligns with the ground truth but includes some false positives. This results in a slightly reduced precision (0.79), while recall remains high (0.969). The Dice coefficient (0.87) and IoU (0.77) reflect good overall overlap. Notably, accuracy remains at 1.0, a result of the dominance of true negative pixels, which can inflate this metric even when positive segmentation is imperfect.

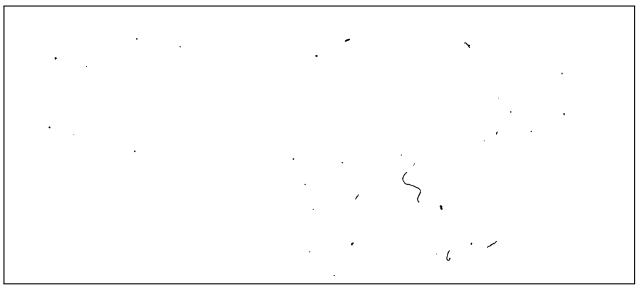

Figure 26: Ground truth image

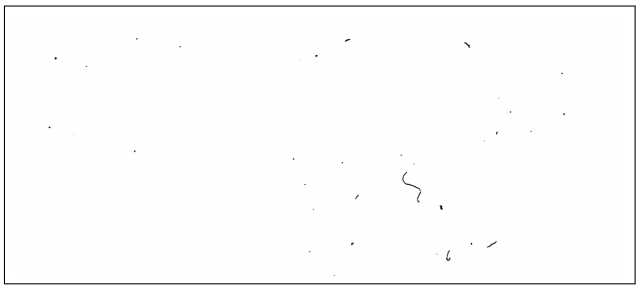

Figure 27: Good prediction

Poor Image

In this example, the model overpredicts, labeling a large portion of the negative mask, as positive. This leads to extremely low precision (0.004) while recall remains high (0.990), as the positive mask is still mostly detected. Despite the very poor overlap between prediction and ground truth, the accuracy remains relatively high. Dice and IoU values correctly evaluate the image, being close to zero, reflecting correctly the segmentation failure.

Figure 28: Ground truth image

Figure 29: Poor prediction

Image 0 (All negative)

This experiment tries to simulate the model failing to detect any positive pixels. Recall drops to 0.0, and Dice/IoU are also 0.0, indicating total undersegmentation. Precision is undefined (NaN) due to a zero denominator (TP + FP = 0). Despite this, accuracy remains high (0.999), as the background is correctly identified.

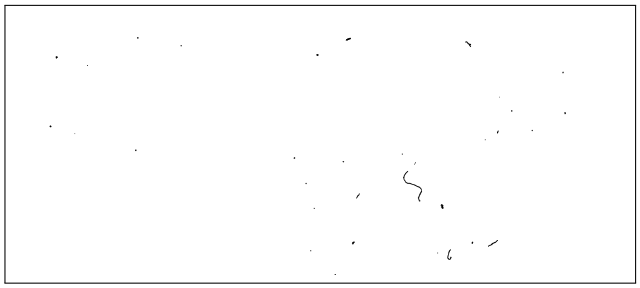

Figure 30: Ground truth image

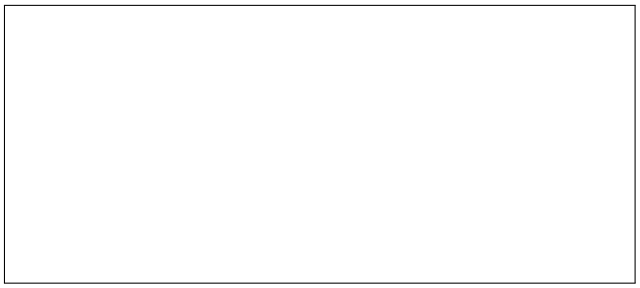

Figure 31: No prediction

Image 1 (All positives)

This case shows the model predicting every pixel as positive. Recall is maximized (1.0), since no actual positive is missed. However, precision drops to 0.001 due to a massive number of false positives. Dice and IoU, being sensitive to both types of errors, remain extremely low (0.001), and accuracy falls to 0.001 due to the incorrect classification of nearly all negative pixels.

9.4 Conclusion

For a comprehensive evaluation of segmentation models, given the results of these experiments, the selected metrics to evaluate all segmentation models trained in this project will be the **Dice Coefficient** and **Intersection over Union**.

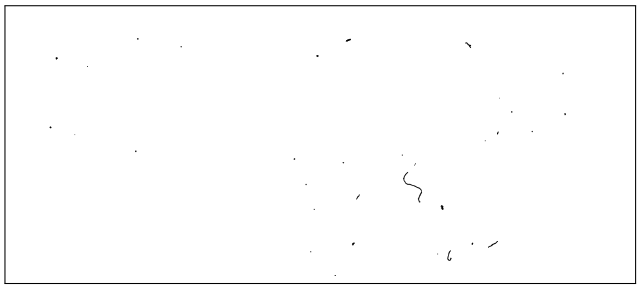

Figure 32: Ground truth image

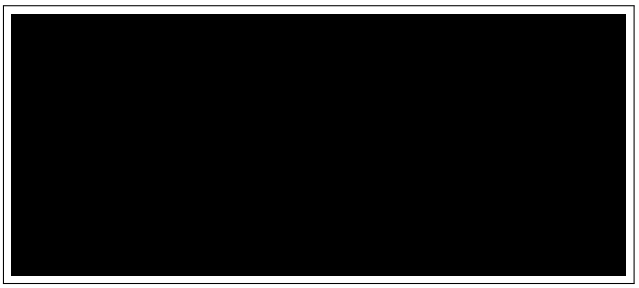

Figure 33: All positive prediction

10 Final goal of the project

Following the completion of this report, I have deemed it appropriate to revisit and refine the initial objectives of the project. The original aim centered on developing a viable solution to the posed problem. While the current pipeline represents significant progress toward this objective, it has yet to achieve optimal performance. Consequently, I have found it pertinent to explore an alternative methodological approach—one that diverges substantially from the present framework.

Specifically, I intend to investigate a unified diffusion-based architecture capable of addressing both the "segmentation" and inpainting tasks simultaneously. Although this novel approach does not explicitly perform segmentation in the conventional sense, it holds promise as a holistic solution that may streamline and potentially outperform traditional multi-stage pipelines. As this direction remains largely unexplored in the context of this project, it offers a compelling opportunity for comparative analysis and innovation.

Moreover, a critical milestone for the next phase of the work will involve conducting a thorough evaluation of the metrics used to assess model performance. A comprehensive study of both quantitative and qualitative evaluation criteria will be essential in ensuring the reliability and consistency of the comparisons drawn between the different models. This step will not only strengthen the empirical foundation of the project but also provide a clearer perspective on the efficacy and limitations of each proposed approach.

References

[1]

[2] Md Zahangir Alom, Mahmudul Hasan, Chris Yakopcic, Tarek M. Taha, and Vijayan K. Asari. Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net) for Medical Image Segmentation, May 2018. arXiv:1802.06955 [cs].

[3] Emmanuel Asiedu Brempong, Simon Kornblith, Ting Chen, Niki Parmar, Matthias Minderer, and Mohammad Norouzi. Denoising pretraining for semantic segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4174–4185, 2022.

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

[5] Shizuo Kaji and Satoshi Kida. Overview of image-to-image translation by use of deep neural networks: denoising, super-resolution, modality conversion, and reconstruction in medical imaging, June 2019. arXiv:1905.08603.

[6] leejunhyun. LeeJunHyun/Image_segmentation, February 2025. original-date: 2018-06-18T08:27:27Z.

[7] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models, 2022.

[8] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention U-Net: Learning Where to Look for the Pancreas, May 2018. arXiv:1804.03999 [cs].

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[10] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. DiffIR: Efficient Diffusion Model for Image Restoration. pages 13095–13105, 2023.

[11] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers, October 2021. arXiv:2105.15203 [cs].

[12] Rui Xie, Yinhong Liu, Penghao Zhou, Chen Zhao, Jun Zhou, Kai Zhang, Zhenyu Zhang, Jian Yang, Zhenheng Yang, and Ying Tai. STAR: Spatial-Temporal Augmentation with Text-to-Video Models for Real-World Video Super-Resolution, January 2025. arXiv:2501.02976 [cs].

[13] Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool. Denoising Diffusion Models for Plug-and-Play Image Restoration, May 2023. arXiv:2305.08995 [cs].