Master in Computational Colour and Spectral Imaging (COSI)

DSTT-MARB: Multi-scale Attention Based Spatio-Temporal Transformers for Old Film Restoration

Master Thesis Report

Presented by

Tawsin Uddin Ahmed

and defended at the

Norwegian University of Science and Technology

September 2022

Academic Supervisor(s): Dr. Marius Pedersen
Host Supervisor: Waaseth Kjartan Sebastian
Jury Committee:

1. Damien Muselet, University of Jean Monnet, Saint-Etienne, FRANCE
2. Pauli Fält, University of Eastern Finland, Joensuu, Finland

Submission of the thesis: 10$^{\text{th}}$ August, 2022
Day of the oral defense: 2$^{\text{nd}}$ September, 2022

Abstract

Within the scope of this effort, our goal is to integrate all aspects of cinema restoration into a unified conceptual framework, within which we will carry out the Spatio-temporal restoration. The most important realization is that the majority of degradations in older films, particularly structured defects, are temporally variant. This means that structured artifacts that covered up information in a region in one frame may disclose their content in subsequent frames even if they were covered up in the previous frame. Therefore, rather than depending on the illusion, we suggest repairing the damage by utilizing the Spatio-temporal environment instead of relying on it. We present a Multi-scale Attention Residual Block based Decoupled Spatio-Temporal Transformer (DSTT-MARB) for the purpose of restoring old film while resolving typical old film artifacts. The DSTT-MARB is split up into two sub-task: The first sub-task involves attending to temporal content flow on multiple frames at the same spatial locations. This is accomplished by using a temporal transformer block that is responsible for temporal information analysis. The second sub-task involves the flow of content on the frame at all spatial positions. This is accomplished by using another spatial transformer block that analyses spatial features. The layering of the two transformer blocks interacting with one another enables our proposed model to pay closer attention to both moving content and texture information, and as a result, the convincing, as well as temporarily plausible content that is attended can be conveyed to reconstruct the defect region. Also, a hierarchical encoder is used prior to the spatial and temporal transformer blocks to discover hierarchical features that preserve multi-level spatial patterns, which leads to more delegate tokens for the transformers. When these two new designs are put together, they do a prominent spatial-temporal attention module. The encoder part also includes a Multi-scale Attention Residual Block (MARB) that extracts features in multiple scales and combines them to obtain more robust features that are fed into the hierarchical encoder.

The proposed DSTT-MARB model is trained in two different manners: video inpainting and video denoising. The main objective is to figure out which one is a more reliable approach to follow in the old film restoration task. Both video inpainting and denoising approaches are evaluated in both reference and non-reference image quality assessment metrics PSNR, SSIM, LPIPS and BRISQUE. Along with the quantitative analysis, a qualitative evaluation is also conducted on old film, which validates the effectiveness of the proposed methodology in the real-world scenario. Current state-of-the-art models are also subject to the evaluation process for a quantitative and visual comparison with the proposed DSTT-MARB model.

Acknowledgment

I would want to use this opportunity to express my gratitude to the almighty Allah, who has planned my life in the best way and made my life easier (Alhumdulillah). You are the one who made it possible for me to complete my goals. I will continue to put my faith in you for the sake of my future and the life hereafter. I would also like to take this opportunity to extend heartfelt gratitude to my parents and the rest of my family for their unwavering support and understanding as I worked on doing research and composing my project. Your prayers for me have been the important thing that has kept me going up to this point.

In closing, I would like to show my appreciation to my academic and industrial supervisors, Dr. Marius Pedersen and Waaseth Kjartan Sebastian, who made this work possible. Their guidance and advice carried me through all the stages of writing my project. I would also like to thank Bernt Erik Baltzersen and my friend Ehsan Ullah for their thoughtful comments and suggestions throughout my masters thesis period.

Contents

CONTENTS

1 Introduction

Classic movies from decades ago still have the ability to evoke strong emotions and stir up audience members' imaginations. Unfortunately, we have experienced a reduction in popularity due to the fact that viewers are no longer accustomed to the poor resolution and unpleasant artifacts that are generated by the aging of photographic film. Despite the fact that film restoration methods have been created, the process of bringing such older movies back to life is still laborious and time-consuming because of the analog nature of the format. The initial task is to restore the film's physical integrity. This is followed by a digital remaking step in which distortion and artifacts are removed. Because such operations need a considerable investment of time and resources that might have a range of millions of dollars, they are presently done manually by professionals. As a result, large sectors like publishing, television, and print have a pressing need for high-quality restoring procedures to use on their massive amounts of old, damaged video content. However, modern restoration is done digitally, which also requires the professionals to inspect individual frames carefully, painstakingly retouch the imperfections, cure the flickering, and so the costs associated with fixing an entire vintage film are daunting. As a result, people have a strong need for a technology that is capable of doing all of these laborious processes mechanically, making it possible to resuscitate old films while giving them a contemporary appearance.

In general, older films are plagued by a variety of degradations, which, as far as we are aware, there is not much research made that attempts to remedy. While it is possible to use specialized frameworks for restoration in a step-by-step process, one might not expect specialized systems to be generalized to the degradations that occur in the actual world. In recent times, several degradations approach (61)(73) have already been recommended to analyze real-world degradations. However, these works primarily take into account the photometric deteriorations, including graininess, haziness or blurriness, and global noise/distortions, instead of the structured flaws like scratches, cracks and so on that disrupt the most in older films. (60) seeks to define complicated deteriorations in historical photographs; however, its frame-wise operation on ancient movies does not maintain temporally stable

(a) (b)

Figure 1.1: *Two scenes from a Norwegian film '**Den-forsvundne-polsemaker (4)**' where yellow circles point to the typical old film artifacts*

outputs.

Dust, scratches, blotches, flicks, and so on are just some of the elements that can degrade an old film over time. Scratches and blotches are the most typical examples of these types of factors. Scratches are typically caused by mechanical rubbings that occur during the process of film duplication. Since scratches show up in the position of the film strip on subsequent images over the film, The inclusion of granules on the surface of the film, such as small hairs, dust, fingerprints, and so on, is what causes blotches to appear. They are easily identifiable as bright or dark spots of arbitrary form and structure, and they only show up in a single frame of the movie film. This is their defining characteristic. Some examples of damaged areas caused by scratches and blotches can be seen in Fig 1.1.

The content of the article is organized as follows: In chapter 1, the research questions and contributions and some background knowledge related to the research will be discussed. Following that, chapter 2 discusses some research works related to video inpainting, video denoising, and old film restoration (both traditional and deep learning approaches). After that, the proposed methodology of this research will be explained along with some details about several components of the proposed model. Then, the experiment part (chapter 4) comes into discussion, where dataset collection, preprocessing, ablation study and a brief description of the training setup are included. The representation of both quantitative and qualitative result analysis and discussion take place in chapter 5. And finally, the article ends with a discussion on the limitation and conclusion.

1.1 Research questions and Contribution

According to the Figure 1.1, old film degradation is not similar at all to the traditional global image/video degradation like gaussian noise (37), blocky artifacts due to jpeg compression (38), blur (43), and so on. This kind of dust or scratches can be visible in any region of the image or video in an inconsistent manner. The occurrence of these kinds of artifacts can not be predefined with an exact region throughout the frames. It is highly possible that old film artifacts that are seen in a region of a frame would not be visible at the exact location in the successive frames of a video. And also, if we analyze the nature of these kinds of artifacts, the degradation of the frame comes with missing valid or contextual pixels in the affected region. Observing this nature, the solving approach may be formulated as a video inpainting task. Inpainting is the process of recreating replacement contents in missing portions of a frame in such a way that the alteration is both aesthetically pleasing and semantically appropriate (29). However, this problem might also be solved by a denoising approach where the main objective is to restore the noisy image by removing physical degradation, keeping as many details as possible compared to the clean true image (56).

In addition, video inpainting or denoising is a tough subject to solve because of the challenges posed by complicated movements and the high emphasis placed on temporal consistency. The inpainting and denoising techniques may be applied to each frame separately for a basic method of doing video inpainting and denoising. However, this disregards the motion constancy resulting from the video mechanics and is thus unable to estimate changes in appearance that are not just minor throughout the course of time in frame space. In addition to this, the implementation of this approach will always result in temporal discrepancies and significant flickering artifacts. Most of the recent approaches use an optical flow estimation sub-network that is used to maintain the temporal consistency in the output frames (68) (63). The purpose of optical flow prediction is to derive an estimation of the motion space based on the changing visual intensity over a time period. This additional flow estimation network results in more computational latency in the old film restoration process.

Considering the discussion above, the research questions can be formulated as follows:

- Which approach is more feasible to follow for old film restoration tasks in between Video inpainting or video denoising?
- How to maintain temporal consistency of the restored video sequences without

involving an optical flow estimation network?

These are the main research questions that are targeted to be resolved in this research. However, apart from that, this research is aimed to achieve the following objectives, which can be addressed as the research contribution:

- Developing a model been able to remove artifacts that the old films mostly suffered from and produce a plausible restored video.
- Preserving temporal consistency of the restored video without involving a separate network for optical flow estimation.

1.2 Background Study

1.2.1 Inpainting

The act of reconstructing a missing portion of an image or video in a manner that is designed to be undetectable to the human eye is referred to as "inpainting." There are two distinct aspects of inpainting, which are image and video inpainting. Image inpainting is a technique that is used on older, more expensive images in order to repair damaged areas of the image, remove scratches, and eliminate objects from the image. The image inpainting technique has a wide range of applications the defect restoration of images/videos, multimedia editing, the replacement of regions in an image for the purpose of protecting a person's privacy. Since the beginning of image processing, scientists have been searching for a method to automatically guide the process of inpainting. This notion of image inpainting dates back a very long time ago. Image Inpainting is a process that restructures damaged areas or pieces that have been misplaced in an image by exploiting the spatial information of the adjacent regions (51). It is a procedure that involves performing operations on an image with the purpose of either improving the image's quality or removing an item from the image. In other words, it is the process of attempting, on the basis of the background information, to fill in any damaged or missing data in an image. Some techniques focus on the application of structural inpainting, while others concentrate on the application of textural inpainting. Consequently, it is essential to pick the appropriate method in accordance with the requirements. In general, methodologies for inpainting may be grouped into two distinct approaches: those based on diffusion and those based on Exemplar. A method based on diffusion works very well for images that lack texture (76). The origins of the Exemplar-based technique may be traced back to the Exemplar-based texture synthesis (11). In comparison to the Diffusion-based inpainting method, the Exemplar-based technique yields

superior results even in the circumstance of a significant amount of missing territory.

Videos communicate more effectively than voice, written words, or still pictures. Several educational and industrial sectors are now using video as a medium for learning as well as communication. Anyway, a video is a succession of frames, and the end product of video inpainting should keep the spatial and temporal link intact. While compared to image inpainting, the number of pixels that need to be filled in when doing video inpainting is quite high. Surveillance video, cinematography, the production of entertainment, and a variety of other instruments all contribute to the creation of many videos that are used in everyday life. The vast majority of the time, these videos are edited by hand, which requires both time and financial investment. Therefore, it is important to provide an automated restoration system for the old film, as well as automatic removal of objects from a series of frames while maintaining a visually believable result. Therefore, the Video Inpainting technique could make it practicable.

1.2.2 Image Noise

The term "noise" refers to an unexpected change in the content's brightness or color, and it is commonly caused by the technical limitations of the photograph collecting sensor or by inappropriate environmental conditions. These challenges are usually unavoidable in real-world situations, which makes picture noise a widespread problem that has to be dealt with proper denoising strategies.

It is a challenging process to remove noise from a picture since the noise is linked to the high-frequency information of the image, also known as the details. As a consequence of this, the objective is to find a solution that minimizes the amount of background noise while minimizing the amount of information that is lost.

The existence of noise in an image may have an additive, multiplicative effect on the overall quality of the image. The production of a distorted signal using an additive noise follows the following rule

$$C(x, y) = I(x, y) + N(x, y) \tag{1.1}$$

Here, $I(x, y)$ is the intensity of the source image, and $N(x, y)$ is the noise that is added to form the distorted signal $C(x, y)$ at the (x, y) pixel location. In a similar manner, the Multiplicative Noise Model works by multiplying the initial signal by the noise signal. It's possible for noise to be introduced into the image at any point throughout the process of acquiring or transmitting the photograph. There is a possibility that noise is introduced into the picture as a result of a number of different circumstances. The quantity of pixels in an image that is incorrectly

colored is one of the factors that is used to quantify the amount of noise.

The most common contributors of noise in digital photographs:

- There is a possibility that the optical sensor will be affected by environmental conditions.
- Image noise may be caused by low light levels and high sensor temperatures.
- Noise in the digital image might be caused by dust particles that are present in the scanner.

The form of the noise and the statistical qualities it has are what set it apart from other types of noise. There is a diverse assortment of different kinds of noise. Gaussian noise, salt-and-pepper noise, poison noise etc. are by far the most significant types.

1.2.3 Old Film Noise

The widespread use of digital imaging is inspiring many people who are passionate about photography to convert their picture collection into digital format. The scanning of negatives and positives, both transmissive media, are favored over the printing of reflected images because these media have a wider variety of tonalities. However, both the negatives and the positives are somewhat tiny, so they need to be scanned at a high resolution in order to be watched on display or printed. The picture scanned at a high quality also made it easier to see dust grains and scratches on the surface of the photograph. In contrast, consumers are becoming more concerned about these aesthetically objectionable flaws. These kinds of flaws also exist in the film industry; the damage is one of the most recognizable aspects of film footage. It's become something of a visual cliché that for a video to resemble actual film, particularly antique film, it must be covered with dust, glitter, and scratches. Cinema, even contemporary film, does, in fact, suffer from these forms of artifacts, although to a lesser extent. Film pictures are literally extremely small bits of material, but when they are projected, they are magnified many times over, which causes degradation. Because of the significant extent of magnification that takes place, even a tiny piece of hair may appear as a noticeable length of rope when watched in a projector. After the film has been scanned and converted into a digital format, any degradation of this sort has already been a part of the picture and, as a result, been converted into pixels, so there is no longer any sign of the image content that is underlying.

1.2.3.1 Categories of old film artifacts

- A major problem with vintage films is the accumulation of dust and filth. They show as either dark or light patches in the image, depending on their color. Every sort of old photograph has some amount of dust which is most often seen on photographs that have been digitized from slides or negatives. Image artifacts such as dust may be seen as either bright or dark areas. Dust may show up in an image as relatively small proportion particles, and they can occasionally seem to be rather thin and lengthy. Any particles of dust that occur on the surface of each frame of film before it is scanned will be permanently incorporated into the picture. Each individual particle of dust is represented in the digital picture by a size that is just a few pixels wide, and they are scattered in a manner that is completely random. Inadequate storage conditions or contamination during the duplicating process are the most common causes of dust (32).

- In more than one frame of the picture, scratches seems to be moving in the direction of the film. If the vendor's system has particles, this is the most common source of the problem. According to the authors of the 1999 study (52), Film line scratches, according to Kim and Kim (2009) (30), may be seen as bright or dark vertical lines. In order to identify a scratch, there are three things to look for: a lower or higher brightness than the surrounding pixels, a vertically long thin line, and the ability to occur in consecutive frames. Vertical line scratches on preserved film are a regular occurrence, according to Kuiper and Sigmund (2005) (32).

- A humid storage environment is what leads to the growth of mold, mildew, and fungus. In most cases, these biological organisms begin their assault on the film roll from the outer edge and work their way inside. The emulsion is susceptible to suffering severe harm at the hands of mold, mildew, and fungus. At first, the growth seems like a series of matte-white patches, but as it continues to spread, it transforms into a lacy pattern that resembles a web (14).

- Image vibrations are caused by the film-transportation mechanisms of movie cameras and duplicating equipment, which have a limited degree of mechanical precision. It's also possible that the shaky camera connection during the shooting was the source of the problem. Regular movements of the camera, such as tilting, panning, zooming, or rotating, might have image vibrations layered on top of them. Image vibrations, often known as jitter, may be attributed to the frequent loading, unloading, winding, and rewinding of the film strips, as stated by Chambah (8). This causes damage to the film holes,

which compromises their ability to maintain a steady and repeated alignment of the pictures while the film is being projected.

- Fluctuation in hue or global brightness that is noticeable from one frame to the next is referred to as flicker. When we talk about variations being "global," we imply that they are consistent within a certain frame. According to Zhang et al. (75), flicker is defined as changes in perceived picture intensity that are not natural and do not come from the source scene. The old black-and-white film is characterized by a flickering effect, which is said to have been generated by the early cinema cameras' erratic exposure times. Flicker is a phenomenon that may occur in contemporary cameras as a result of interferences between the exposure and the illumination when the footage is being recorded. Ohuchi, Seto et al. (44) state that visual flickers are caused by a number of different things, including unstable chemical synthesis, duplicating, the natural aging phase of the film, and inaccurate post-processing. Dust and aliasing have been added to the list of potential causes by Van Roosmalen et al. (58).

1.2.4 Old Film Restoration

There are several choices available to a user for cleaning up dust and scratch marks on a digital photograph. Adobe Photoshop (1) provides users with a variety of choices to choose from. A dust and scratch filter is one of the choices that may be made, and it can be found in the "Filter/Noise" menu. The whole picture or a specific section may be blurred with this method. There is no detection of any kind carried out by it. It produces a hazy image when applied to the whole picture all at once. The user may apply the filter on a regional scale by picking areas that have flaws, although doing so involves a significant amount of human labor on their part. In the event that a flaw manifests itself in a textured region, the filter will smooth out the texture. The use of Photoshop's stamp tool is yet another solution to the problem of removing imperfections. Using this tool, faults in textured regions may be repaired. However, using this program properly takes a significant amount of effort and requires some prior experience.

In old film restoration, all available source content is analyzed and patched together into the sequence recommended by exhibition history and production documents in film restoration efforts to produce a particular version of the movie. In order to make up for previous harm, picture and sound quality must often be improved. Usually, the major goal is to recreate how the movie seemed when it first came out. According to Haas et al. (52), the essential factors why film restoration is necessary are as follows: Since the beginning of the 20th century, the whole content that can be found in cinema and television archives serves as a

record of the historical progression of cultural and creative expression throughout all spheres of life. It is essential that it be protected. A restoration step is required in order to maintain a copy in a state that is as similar to the original as is humanly feasible. The economic benefit of restoring old films is the second motivation for doing so. This increase paves the way for the formation of a new sector for film and television archives as a result of the rapid development of communication mediums such as multi-media, satellite, and cable TV. The fact that these records are only partially usable without the restoration step, however, makes it vital for film and television archives to undergo both maintenance and restoration.

2 Literature Review

This chapter comprises of section-wise analysis of some previous research works that are conducted on video inpainting, video denoising and old film restoration field. It should be mentioned that both traditional and deep learning approaches are presented in the old film restoration section.

2.1 Image denoising

2.1.1 Comprehensive Solutions for Removal of Dust and Scratches from Images

According to Bergman et al. (39) image reconstruction can be thought of as a challenge that has to be overcome when damaged pixels need to be fixed. Because of this issue, the pixel that is being fixed, along with a significant number of the pixels that are immediately next to it, may be damaged. This means that the values of these pixels are only tangentially connected to their respective initial values. Because of this, the reconstruction process should not take into account the values found at these pixels. The faulty zones are simply smoothed out using simple reconstruction techniques, such as the one described in (48). For instance, a median filter or another kind of averaging may be used. The median reconstruction is quite effective in terms of the amount of time it takes to do computations, but it does not provide particularly high-quality images. An example of median correction is shown in Figure 2.1 3(d). The flawed pixels have been corrected, but the huge scratch may still be seen since the repair is too smooth and does not correspond to the texture in that region.

One of the techniques that are used for local reconstruction is an extension of the bilateral filter (57), which is a filter that reduces noise. When the nature of noise follows a Gaussian distribution, the bilateral filter is able to differentiate between the noise and the image features; in this case, it will get rid of the noise while

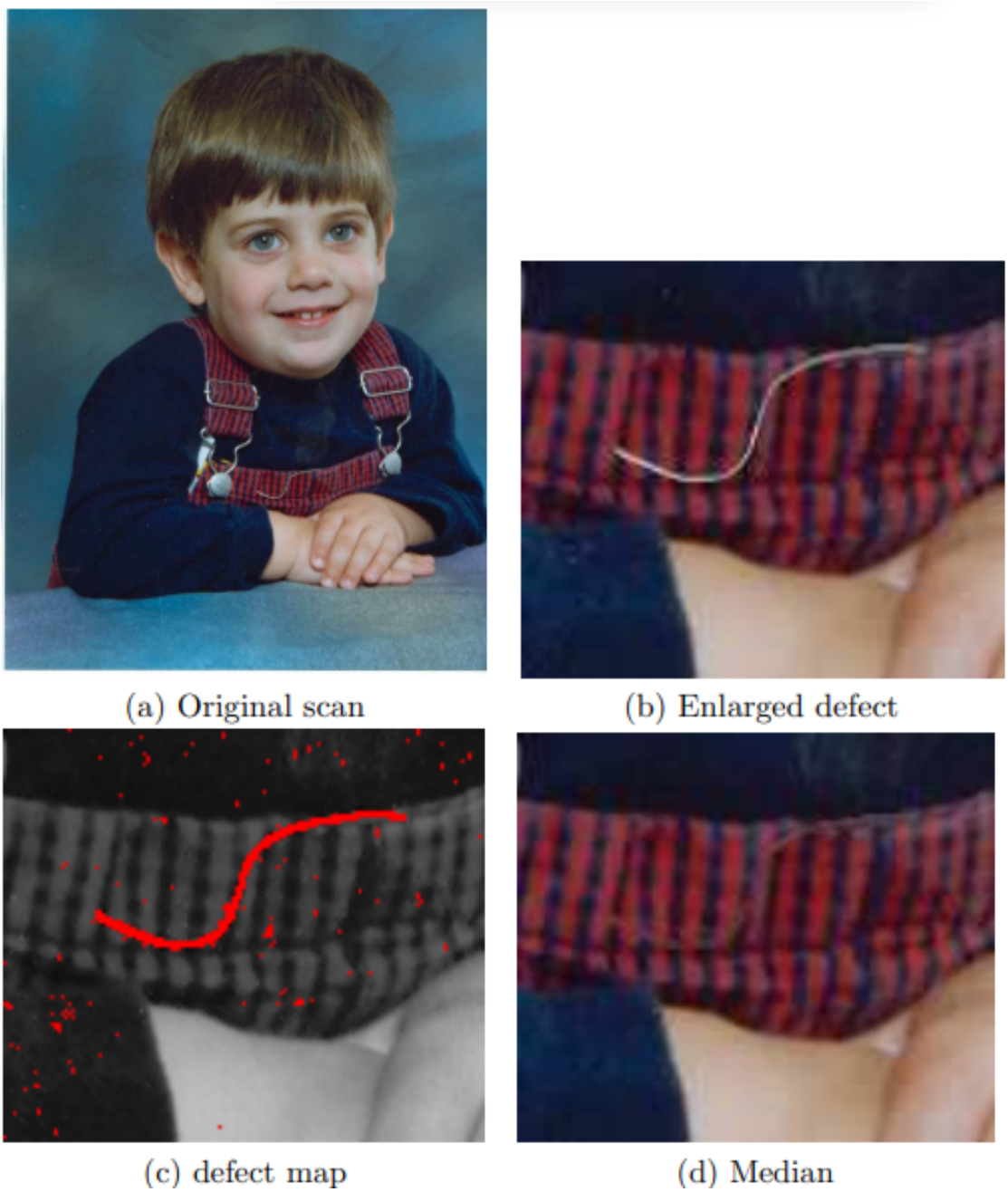

Figure 2.1: *The results of the comparison of the reconstructions. (a) displays the original photograph, and (b) provides a more comprehensive review of the dust (b). (c) illustrates the defect location (d) median reconstruction. The figure is taken from (39)*

keeping the features. The bilateral filter distinguishes noise, which holds a local contrast that is comparatively lower, from features that have a local contrast that is considerably large. This occurs when the noise follows a Gaussian distribution having a known variance. When an image's noise does not match a Gaussian distribution, the bilateral filter may perceive substantial local differences as image characteristics rather than flaws that need to be smoothed out. This occurs when the noise does not have a Gaussian distribution. Consequently, the dust and scratch issues would be exacerbated by the use of the bilateral filter.

For the purpose of cleaning dust and scratches from digital photos, (39) offered a complete collection of algorithms. The work of removal was broken down into two stages: the first stage was detection, and the second stage was rebuilding. Their algorithms have been updated to include a number of new and improved features. The software-only local detection stage makes use of differences in local contrast, which allows the algorithm to differentiate between picture defects and preserve image edges better than the more traditional gray level differences could. Their research has led them to the conclusion that, in the absence of surrounding information, regional features are required in order to distinguish between dust and scratch defects and picture features that share comparable characteristics. Through the use of simple yet perceptive heuristics, the phase of regional categorization successfully eliminates any erroneous detections. To illustrate, the heuristic for identifying texture is absolutely necessary in order to prevent blurry results. Glint removal should be avoided at all costs since faces appear in the vast majority of photographs.

As part of the reconstruction process, the authors also take into account regional factors in order to fix the texture. The contextual reconstruction is successful in producing a restoration of the faulty pixels that is extremely convincing. The repair performs better than local algorithms when applied to textures. The directed reconstruction, which is their local method for mending from a defect map, is effective at fixing damaged pixels, and it can often compensate for incorrect detections by recreating the features. Due to the fact that it is a local technique, it works best for situations in which performance is of the utmost importance. They presented a four-tiered system of dust and scratch removal techniques that are adaptable to a variety of contexts. They demonstrate that improved picture quality may be achieved by either increasing the amount of side information or relaxing the performance restrictions. The outcomes of the solution at each tier match well with those of similar solutions offered by competitors.

2.1.2 A Generative Adversarial Approach with Residual Learning for Dust and Scratches Artifacts Removal

Following Mironică (40), it has been demonstrated that Generative Adversarial Networks, more commonly referred to as GANs, achieve superior performance in a wide range of automated image restoration tasks when compared to other methods (33)(22). Because of this, the author decides to look into how GANs can be used in the post-production system of movies. In this work, the author presents a GAN-based approach to removing dust and scratches from film. The training process is sped up with the help of residual learning, and the potency of denoising can also be significantly improved. In this research, the authors present a method that has been developed specifically for the purpose of repairing dust and scratch defects that are present in older film pictures. It is possible to make use of the multi-scale redundancy of natural image artifacts with the help of a novel approach that is based on adversarial learning and is developed. The method can be broken down into two components: a generator that is analogous to a feed-forward CNN network and a discriminator that is analogous to the PatchGAN framework. The author uses pixel loss, gradient loss, and perceptual loss in order to get the predicted image closer to the reference image so that there is less of a difference between the two. The author also aims to increase the perceived quality of the picture that has been repaired, which is why we have added the adversarial loss in our algorithm.

According to the author, the state-of-the-art pix2pix algorithm (25) that is integrated with Photoshop is also capable of removing the majority of artifacts; however, this results in the resulting picture being fuzzy, and many crucial features are lost in the process. This issue is solved in this research as the proposed architecture is capable of capturing all the image details. However, the model evaluation is conducted on the basis of the Peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM), which does not take into account the human visual system or the perceptual aspect the image.

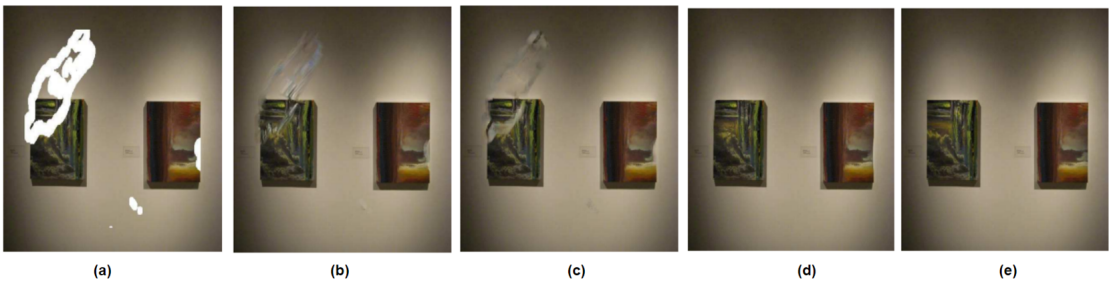

Figure 2.2: *(a) image with missing region of arbitrary shapes (b) inpainted result of Iizuka et al. (24) (c) Yu et al. (69) (d) Partial Convolution (34)(e) Ground truth. The figure is taken from (34)*

2.2 Inpainting

2.2.1 Image Inpainting for Irregular Holes Using Partial Convolutions

Recent image inpainting attempts have made use of deep neural networks, which train contextual and relevant latent representations in an end-to-end way. This has been done in order to improve accuracy. Images are processed using convolutional filters by these networks, which replace any material that is filtered out with a constant value. As a consequence of this, these methods are subject to the drawback of being dependent on the initial hole values. This frequently takes the form of a loss of texture in the hole areas, evident color contrasts, or false edge reactions encircling the hole. When the output is conditioned on the hole values, it finally produces a variety of visual artifacts, the removal of which requires costly post-processing. These refinements, however, are not capable of resolving all of the artifacts that are depicted as 2.2(b) and 2.2(c). Liu et al. (34) attempt to create well-incorporated hole estimations that are irrespective of the hole initialization parameters and do not require any further post-processing.

Another shortcoming of many contemporary methods is that they concentrate on producing holes of the rectangular shape, which are typically presumed to be centered in the picture. According to their findings, these restrictions may result in overfitting the rectangle holes, which, in the end, reduces the value of these models when they are applied. The authors collect a large baseline of images with irregular masks of varied shapes so that we may concentrate on the more realistic irregular hole use case. Within the scope of our study, they investigate not only the impact of the hole's shape but also the question of whether or not the holes make contact with the image border.

To sum up, the typical approaches to image inpainting relying on deep learning employ a conventional convolutional network over the damaged image. These techniques employ convolutional filter outputs that are convoluted on valid pixels and the replaced values in the holes. In most cases, the substitution is done with mean values. This frequently results in artifacts like color inconsistency and blurriness in the image. Post-processing is typically utilized to lessen the appearance of such artifacts. However, this method is costly and there is a chance that it may fail. The authors recommend making use of partial convolutions, in which the feature extraction is masked and renormalized such that it is conditioned on just pixels that are valid. As an additional component of the forward pass, they have included a system that will automatically create an updated mask for the subsequent layer. When it comes to irregular masks, their model performs better than previous techniques. In order to validate their technique, they provide both qualitative and quantitative assessments with many different methodologies.

2.2.2 Decoupled Spatial-Temporal Transformer for Video Inpainting

Even with successful deep learning algorithms, video inpainting is still a difficult challenge since it seeks to fill the supplied Spatio-temporal holes with a natural look. Recent efforts have been successful in achieving greater performance and have introduced the prospective transformer framework into deep video inpainting. However, it continues to have the issue of generating a fuzzy texture in addition to having a very high computational burden. To this aim, Liu et al. present a unique Decoupled Spatial-Temporal Transformer (DSTT) (35) for the purpose of optimizing the video inpainting process while maintaining an exceptionally high level of effectiveness. The task of learning spatial-temporal attention is separated into two parts. The first sub-task involves attention to temporal object movement patterns on multiple frames at the same spatial locations. This is accomplished by using a transformer block that is temporally decoupled. The second sub-task involves attention to similar background textures in the same frame at all spatial positions. This is accomplished by using a transformer block that is spatially decoupled. The pile of such two blocks interacting with one another enables the proposed model to pay closer attention to background textures and moving objects, and as a result, the convincing and temporally consistent appearance that is attended may be transmitted to fill in the gaps. In addition, a hierarchical encoder is used before the stacking of transformer units. This is done for the purpose of learning resilient and hierarchical features that retain multi-level local spatial structure, which ultimately leads to the production of more representative token vectors. Their suggested model delivers superior performance to state-of-the-art

video inpainting techniques with considerably increased efficiency, and the flawless merging of these two innovative designs produces a superior spatial-temporal attention framework.

2.3 Old Film Restoration

2.3.1 Digital Image Processing Techniques for Restoring Old Damaged Films and Their Applications to Korean Film Restoration

Kim et al. (28) chose frames that were temporarily nearby in order to clean up any dust or blemishes. A motion estimator analyzes difference maps between chronologically consecutive frames in order to account for the fact that motion pictures feature dynamic frames. Estimating a motion vector requires reducing the mean absolute error (MAE) between each pair of frames, and a specified region-of-interest (ROI) is used in this process. In order to minimize the appearance of block artifacts, the difference maps are then improved in the area around an object's border. In the end, defects are found by contrasting the motion-compensated areas of two different frames and imposing topological connections. Figure 2.3 depicts the algorithm that follows five steps: (i) motion map adjustment on the existence of a defect, (ii) the assessment of inter-frame intensity correspondence, (iii) the spatial restoration of the missing pixels, (iv) the frame restoration with respect to the time domain, and (v) restoration in the remaining regions is done by the distribution of texture.

The inaccuracy of the disparity map is due to the fact that flaws, such as Dust and scratches, distort the image's geometrical structure. In order to guarantee the accuracy of the map, they first repeatedly propagate the area orientation toward the border of the defects, and then they replace each displaced component with the one that is considered to be the most indicative of its surrounding area, that is still valid. In the third stage, they spatially restore an estimated picture pattern using the same frame by iterative propagating from valid defect boundaries. This allows them to use this structure as a starting point for further processing. After taking advantage of the temporal neighborhood, the authors mix them together using a filter with three taps. The last step in the process involves extracting a textured pattern from a spatial region and inserting it into areas that have inadequate temporal restoration.

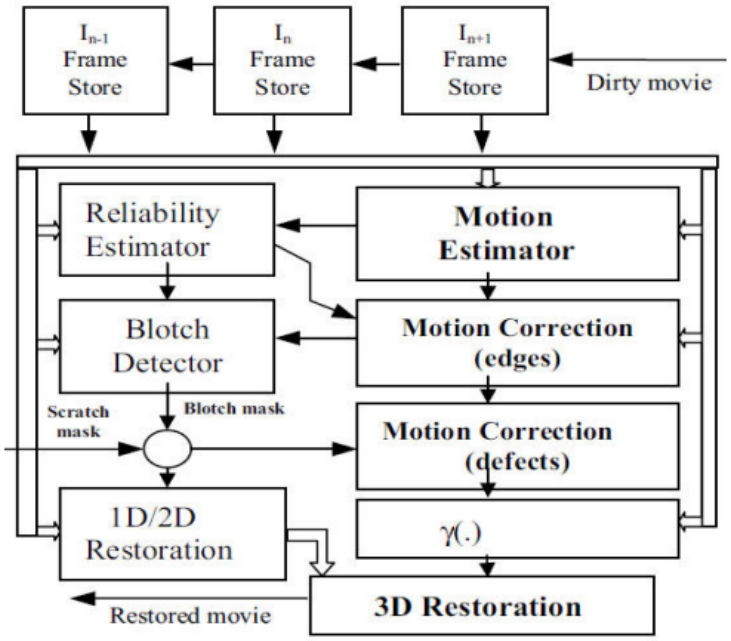

Figure 2.3: *Dust and scratch removing algorithm proposed by Kim et al. (28).*

2.3.2 Image Processing for Restoration of Heavily-Corrupted Old Film Sequences

Saito et al. (50) describe several effective image processing algorithms that are used for the flicker repair and scratch and blotch elimination activities. These activities are also involved in the process of restoring ancient films. They provide a restoration approach that, in order to correct flickering, first assesses the characteristics of the restoration model based on an input of an older film sequence and then adjusts the flickering based on the estimated model. Scratch and blotches are corrected by filters that are of the blending kind. They make use of morphological image processing for the purpose of scratch detecting, while for blotch identification, they make use of the spatial-temporal continuity evaluation. The simulations have shown that their methods are virtually flawless in removing flickers, scratches, and blotches from images.

2.3.3 DeepRemaster

DeepRemaster (23) is one of the state-of-the-art technologies in the domain of old film restoration. However, this restoration framework includes both artifact removal as well as colorization steps. According to the authors (Satoshi Iizuka

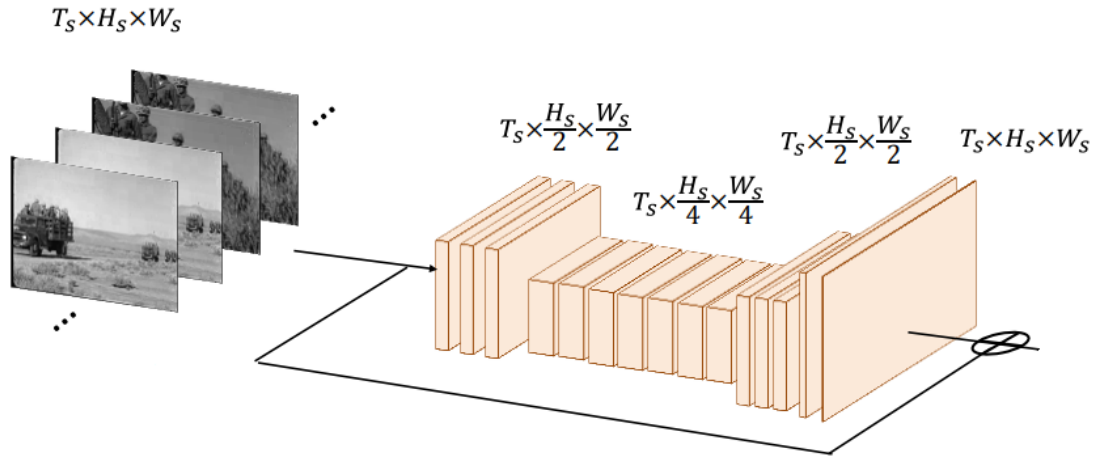

Figure 2.4: *A series of b/w photos are fed into the model, and after being restored using a preprocessing network, the authors utilize those as the luminance channels for the colorization network. Figure is collected from (23)*

and Edgar Simo-Serra), it is not as easy as applying a noise reduction technique accompanied by a colorization technique in a pipeline form to remaster an ancient video. The noise filtering operation and the colorization method are interconnected and impact each other. In addition, the majority of older films are blurry and have poor resolution; hence, improving the sharpness is essential in order to compensate for these issues. The authors propose a comprehensive pipeline for remastering black-and-white videos. This pipeline comprised numerous trainable components, all of which are trained within a single end-to-end framework. The authors are able to train the model to remaster videos with noise removal and introduce color, as well as increase the resolution and sharpness and enhance the contrast with temporal consistency. This is made possible by employing a data creation and augmentation strategy that is both meticulous and comprehensive. The initial part of their work is more relevant to our work, which they address as the preprocessing network shown in Fig 3.2. This preprocessing network is responsible for dust and scratch removal.

The preprocessing network is composed entirely of layers of 3D convolution addressed as temporal convolution. A long skip connection is utilized between the input and output of the network. Half reduces the frame dimension in the model's encoder-decoder framework before being restored to its original dimension using trilinear upsampling at the very end of the process. The majority of the refining is carried out at a low resolution to lessen the burden placed on the workstation. The end result of the preprocessing network is considered the frame's luminance

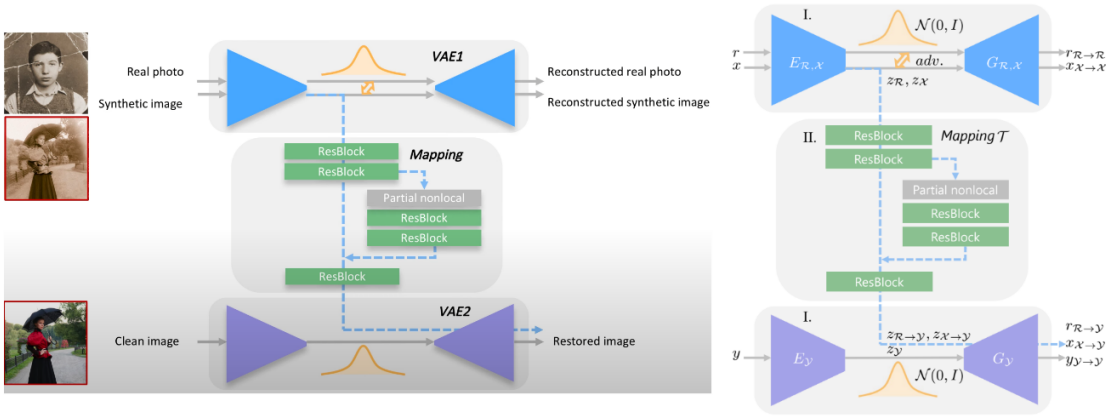

Figure 2.5: *The design of the network proposed in (60) for restoration. (I.) For pictures in real photographs, the authors train VAE1 and VAE2 using an adversarial discriminator; VAE1 is trained for both real and synthetic images, while VAE2 is trained for clean images. Images are compressed into a smaller latent space using VAEs. (II.) Afterward, they discover the mapping that returns corrupted pictures in the latent space to their original form. Figure is taken from (60)*

channel for further colorization.

The limitation of the work is that the suggested method is based on learning that is completely supervised, and as a result, it is unable to fill in missing frames or deal with high deterioration, which causes a major portion of the image to be missing for many frames. In some instances, there is an insufficient amount of information present, which renders the process of remastering difficult. Remaking new plausible portions of the video would need picture completion-based methodologies, which are beyond the purview of this particular piece of work.

2.3.4 Bringing old photos back to life

Wan et al. (60) propose a deep learning technique to rebuild old photographs that have suffered from significant deterioration over the years. The deterioration in photographs is intricate, and the domain disparity between synthetic images and real-world old photos makes it impossible for the network to generalize its results. Traditional reconstruction tasks, on the other hand, can be fixed through supervised learning. They propose an innovative triplet domain translation network that makes use of both real images and a massive number of artificial image pairs. In particular, they train two variational autoencoders (VAEs) to translate old and clean images into two distinct latent spaces, one for each type of image. In addition,

the transformation between two latent spaces is something that can be learned with the help of synthetic paired data. The disparity between the compact latent space has been closed, which allows this transformation to generalize to real photographs.

In addition, branches with partial non-local blocks are designed to pay attention to the structured and unstructured artifacts, including scratches and dust, noises, blurriness, and so on. When it comes to restoring severely aged and deteriorated photographs, their approach shows promising results. However, their method is incapable of dealing with intricate shading. This is due to the fact that their dataset only contains a small number of older photos that have such flaws.

2.3.5 Bringing Old Films Back to Life

The goal of Wan et al. (59) is to integrate all aspects of cinema restoration into a unified conceptual framework, within which they carry out the spatial-temporal restoration. They recommend a bi-directional recurrent neural network that does the following: it accumulates the information from the picture over subsequent frames, which successfully reduces the intensity fluctuation (flickering) of the video. The description of the scene's information is embedded deep inside the concealed state of the recurrent component. After the alignment is complete, the restoration process for a particular frame fuses the hidden representation, providing valuable insight into the content information behind the faults. A strategy like this one that repeats itself has three distinct advantages. To begin with, it is possible to completely recover the film even if it has been damaged, regardless of how serious the damage can be, provided that the information is still legible in other frames. Second, the preservation of the hidden information in an explicit form guarantees that the restoration of the frames will be chronologically stable over an extended duration. A significant disparity can be seen between the content of the frame and the hidden state in certain places, which makes it possible for the structural faults to be targeted in an unsupervised approach. This is the most crucial point. In contrast to (60), which necessitates the use of a defect segmentation network, such defect localization is more generally applicable to the old film degradations that occur in the actual world.

During the frame alignment, the authors require a component that can take into consideration the tiny discrepancy that exists spatially. As a result, they suggest making use of the Swin Transformer (36); this means that even when the hidden depiction is not properly aligned, the association of the relevant pixels can still be replicated via the use of self-attention. In point of fact, they witness more stable

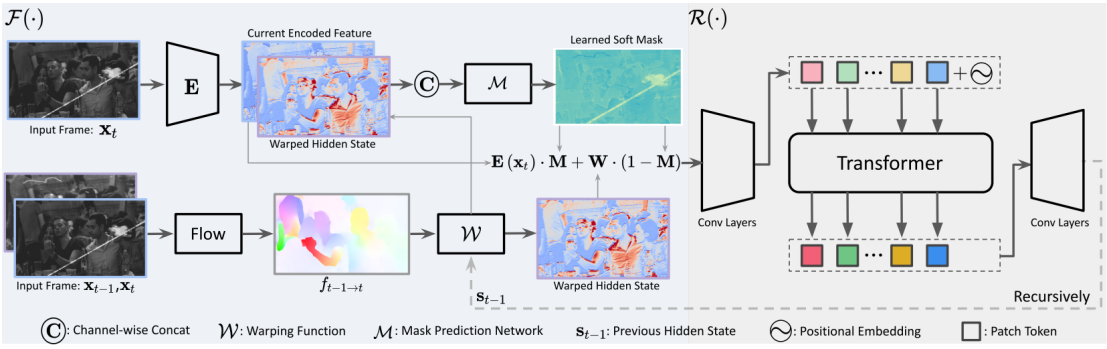

Figure 2.6: *The architecture of the once-recurrent forward propagation, which consists of the temporal aggregation component F and the spatial restoration transformer R. The same model applies to the process of backward propagation. Figure is taken from (59)*

training of the recurrent component as a result of the utilization of attention. In addition, the transformer blocks provide greater restoration capacity for combined degradations, which would be difficult to deal with using a dedicated ConvNet (54). This is made possible by the higher expressivity of the transformer blocks. Therefore, the proposed framework allows the most of both the recurrent component and the transformers, which are the essence of preserving past information. The recurrent components stand to benefit from the temporal consistency. Meanwhile, the protracted modeling potential of transformers enables spatial reconstruction. This surpasses state-of-the-art frameworks on synthetic datasets, and it generates extraordinary efficiency when rebuilding old movies.

3 Methodology

This chapter is dedicated for explaining the proposed approach for the old film restoration task. However, before diving into the actual methodology description, some insight about the several components of the proposed Methodology is presented at the beginning of the chapter.

The old film restoration section of Chapter 02 discusses both the traditional and deep learning approaches. However, similar to other problem domains, deep learning approaches are dominating the old film restoration field nowadays. There are a handful of reasons for adopting data-driven approaches rather than algorithmic strategies.

- Traditional or algorithmic approaches are developed to target predefined types of old film artifacts because (e.g. 2.3.1 deals with dust, scratches; 2.3.2 resolves flickering, scratch and blotch) So, it is not feasible to expect an algorithmic approach to present a generalized model that is able to take into account all sorts of old film artifacts without prior knowledge. However, on the other hand, the restoration capabilities of the data-driven approach deep learning are not restricted to artifact categories. The model is normally well generalized to any sort of old film artifact until the training dataset includes sufficient artifact samples.

- In the case of human intervention, the question that has to be answered is how well the experts can replicate the original. There is a possibility that they may discover certain inconsistencies when viewing the sequence of the corrected film. Two potential pitfalls: first, it takes a significant amount of time, and we all know that time equals money; second, it may result in an inconsistent appearance when applied to many sequences.

Due to the reasons mentioned above, this research is highly motivated to work with deep learning approach for old film restoration. The individual components of the proposed model are explained as follow:

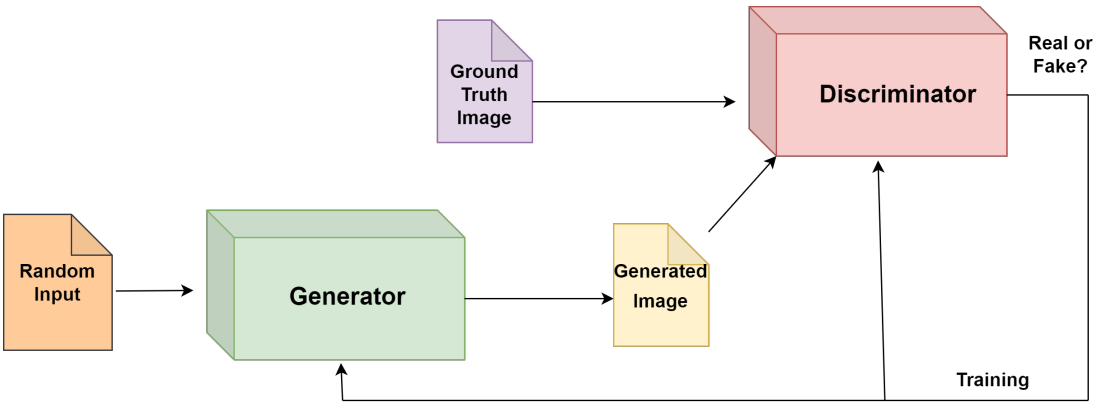

Figure 3.1: *A typical Generative Adversarial Network (GAN) consisting of a Generator and a Discriminator*

3.1 Generative Adversarial Network (GAN)

The term "Generative Adversarial Networks," more commonly abbreviated as "GANs," refers to a framework that makes use of two convolution neural networks that are opposed to each other (hence the name "adversarial") in order to produce synthesized examples of data that are convincing enough to be mistaken for real data. It is first introduced by Goodfellow et al. (17). The field of image restoration, video and audio formation all make extensive applications of GANs. It has been said that adversarial training and GANs are two of the most exciting concepts to come out of the field of machine learning in recent years. Because GANs may learn to imitate any distribution based on and derived from the data, they have a significant potential to be both a benefit and a scourge in the future. GANs have the potential to be trained to dynamically produce a wide variety of things, including pictures, audio, voice, and even written text (20)(65).

GANs are comprised of two components: a generator, which may be conceptualized as a convolutional neural network and contributes to the generation of new data samples; and a discriminator, which examines the newly generated data samples to determine whether or not they are real. The discriminator evaluates each example of data that it looks at and chooses whether or not that data sample belongs to the real training dataset. The discriminator will also provide a penalty to the generator if it produces results that are improbable. It is also possible to think of it as an adversarial process in which the generator attempts to fool the discriminator by producing data that is comparable to that which is found in the training set. They both operate concurrently to learn and train complicated data like audio, video, or picture files.

The following is a condensed version of the stages that a GAN takes:

- An image is produced by a generator after it has been fed a series of arbitrary integers.
- After that, the generated image is included in the sequence of images that are put into the discriminator, which also includes images collected from the real dataset.
- The discriminator examines both the genuine and the false images and then delivers probability in the form of an integer between zero and one. A probability of one indicates that the item is legitimate, while a probability of zero indicates that the image is fake.

Generative models have emerged as a dynamic and fast significant reform, bringing on their pledge of modeling and analysis instances across many application domains, most prominently in the image-to-image translation tasks.

3.2 Encoder-decoder

In this part, we are going to discuss an encoder-decoder network, in general, which is used in the generator part of a GAN architecture. The encoder's job is to reduce the amount of original data to a form that the decoder can interpret while still retaining the most important characteristics. The best way to remember a long sequence is to break it down into smaller chunks that are easier to recall, such as whole numbers, and then reconstruct the entire sequence from those smaller chunks. So, the encoder is responsible for downsampling the data into a lower dimension, preserving as many significant features as possible that represent most of the image information for reconstruction.

However, the decoder operates in a manner that is similar to that of the encoder but in the opposite direction. It is taught to read these compressed data representations rather than produce them, and then it generates images depending on the information it has learned. Obviously, it tries to reduce the amount of damage done during reconstruction. The generator loss is used to assess the outputs by evaluating the restored image against the reference one; the lower the difference is, the more identical the restored image is to the original.

At this point, all of the parameters in the encoder are going to be updated by propagating backward via the decoder. As a result, the decoder and the encoder both have their performance in their respective duties reviewed and their parameter settings updated depending on the discrepancies that are found between the reference and the predicted image.

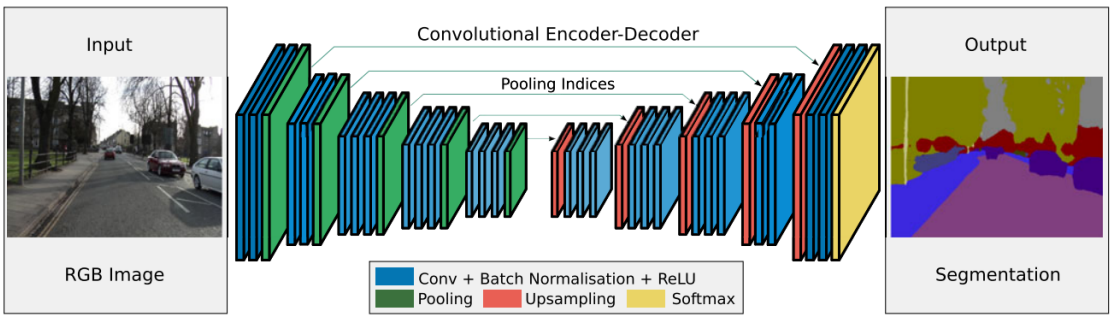

Figure 3.2: *An example of an Encoder-Decoder architecture where the input RGB image is downsampled by the encoder while the decoder brings it back to its initial dimension through upsampling. Figure is collected from (7)*

3.3 Multi-scale Attention Residual Block (MARB)

In order to improve a model's understanding of its surroundings and the context, multi-scale feature extraction, proposed by Chen et al. (10), has become an increasingly popular technique. This multi-scale feature extraction technique combines features of varying scales in an efficient manner. Mathematical expressions can be used to provide a comprehensive description of the Multi-scale Attention Residual Block (MARB). With reference to Figure 3.5, the MSARB is configured to receive the feature F_0 as its input. This feature then travels through the two convolutional layers with kernel dimensions of 3 x 3 and 5 x 5 separately and concatenated after feature extraction.

$$F_1^{3x3} = k_{3x3}(F_0); F_1^{5x5} = k_{5x5}(F_0) \tag{3.1}$$

In equation 3.1, F_1^{3x3} and F_1^{5x5} refers to the extracted feature obtained from convolution filters with kernel (k) shape $3x3$ and $5x5$ respectively. The visual characteristics are retrieved even further using convolution filters with a kernel size of either 3 x 3 or 5 x 5.

$$F_2^{3x3} = k_{3x3}(F_1^{3x3} + F_1^{5x5}); F_2^{5x5} = k_{5x5}(F_1^{3x3} + F_1^{5x5}) \tag{3.2}$$

Equation 3.3 shows the concatenation of multi-scale features exacted using a kernel with dimensions 3 x 3 and 5 x 5. So, in order to increase the robustness of the model, the authors propose multi-scale feature fusion within the layers. This kind of feature fusion is capable of integrating multi-scale information with the characteristics of various levels. This framework ensures that the feature information can be transmitted across all of the layers. As a result, the MARB is

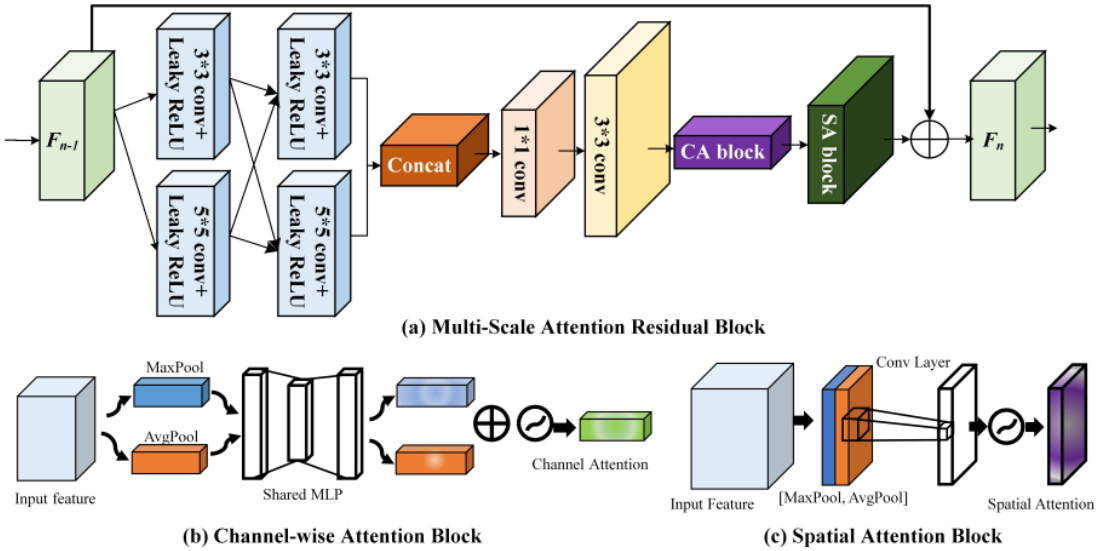

Figure 3.3: *(a) The multi-scale residual blocks and the feature attention module make up the architecture of the Multi-scale Attention Residual Block (MARB). The channel-wise attention (CA) block (b) and the spatial attention (SA) block are both consecutive sub-modules that are included inside the feature module (c). Figure is collected from (10)*

able to learn the key image characteristics by using a variety of scales. In order to hasten the training process, a global skip connection has been implemented between the various residual block schemes. This connection makes it possible to back-propagate gradients and change parameters. Because this skip link is able to additionally immediately disseminate uncompressed information across the whole network, it is helpful for predicting the final output.

The study presented by Yingjun Du et al. (16) reveals that using a channel-wise attention technique may assist the network in better preserving the pixel intensity information than earlier methods, which handle various channels in an equivalent fashion. As a result, the channel attention may be able to detect the artifact zone and contribute to the feature extraction process. Until then, the patterns of dust are unequal and vary greatly throughout a wide range of geographical regions. As a result, paying attention to space could also be necessary while dealing with artifact regions. Convolution layers with sizes of 1x1 and 3x3 are used to produce multi-scale feature fusion. Channel and spatial attention sub-units are incorporated to elevate feature fusion. The final equation can be formulated as follows:

$$F_{out} = SA(CA(k_{3x3}(k_{1x1}(Concat(F_2^{3x3}, F_2^{5x5}))))))) + F_0 \tag{3.3}$$

In equation 3.3, SA() and CA refer to the spatial and channel attention blocks, respectively and F_{out} is the output refined features after applying the multi-scale attention residual block.

3.4 Hierarchical Encoder and Transformers

The model that we have presented contains a stack of transformer blocks that have been built expressly for the purpose of transmitting spatial-temporal information from known pixels that are located outside the dust or artifact region at each frame. However, owing to the need to preserve the picture's spatial arrangement, an ineffective method for dealing with low-level vision problems is to crudely split the image into a large number of patches and then compress those patches into token sets like Vision Transformer (ViT) (15). The self-attention mechanism that is applied to the patches provides an overall spatial arrangement relying upon this relative position among the patches. However, the local spatial structure within a patch is not preserved, which is undesirable for frame formation. This is due to the fact that the self-attention technique on the patches depends on the relative location among the patches. Consequently, in order to obtain accurate and credible token sets, we need an encoder that is both sophisticated and in-depth. The feature maps obtained from the earlier stages are passed through a hierarchical encoder (HE), proposed by Rui Liu et al. (35) which thus combines extracted features across multiple phases to generate tokens that have more representable information. If we pass a frame f of h (height), w (width), and c (channel) through our model, until the hierarchical encoder part, the frame is downsampled and we obtain a feature map F with the dimension of h/4 * w/4 * c. We formulate a hierarchical framework for multi-level channels by combining generated feature maps.

Initially, let's consider the feature map as the first stage feature F_1. In the beginning, it goes into a convolution layer that has a kernel size of 3x3 in order to get the second stage feature map. The spatial form of this map is the same as the shape of the first stage feature map, but it has a bigger receptive field. Then the first stage and second stage convoluted features are concatenated together, preserving the spatial features of both stages. As the features are concatenated along the channel dimension, we get the feature dimension of h/4 * w/4 * 2c. After that, features from several stages are passed to the next stage of the hierarchical encoder. This time, the receptive field becomes wider and the first stage feature is always combined with the feature map of the current stage. This procedure is outlined in detail:

$$\hat{F}_{i+1} = ReLU(Conv(F_k^i))), \hat{F}_{k+1}^i \in \mathbb{R}^{h/4 \times w/4 \times c} \tag{3.4}$$

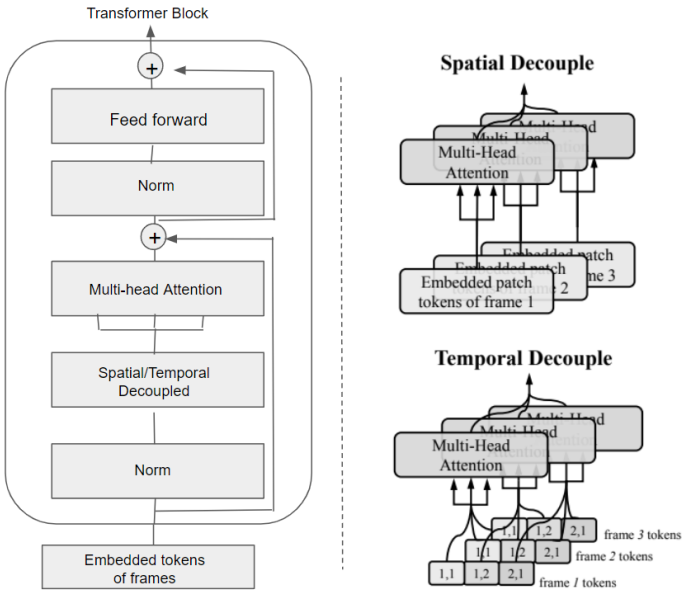

Figure 3.4: *The depiction of components of the Transformer block. The parts of a transformer block are multi-headed attention module and an feed-forward net. Figure is inspired by (35)*

$$\hat{F}_{i+1} = Concat(F^i_{k+1}, \hat{F}^i_k), k = 1, ..., N-1 \tag{3.5}$$

In equation 3.4, Conv refers to the convolution layer and ReLU indicates the non-linear activation function. Channel-wise concatenation and hierarchical layers are denoted by Concat and N respectively in equation . A resilient and informative feature map will be produced, which is useful for further patch-level spatial and temporal accumulation since the final feature map will comprise multi-stage local spatial features from the input frame. We use a group convolution in every layer to handle feature refining at different stages in order to preserve the local spatial information. In this research, five encoder layers with a group number of 1, 2, 4, 8, and 1 are included in the hierarchical encoder. Setting up groups in such a way prevents the intervening layers from fusing features and so preserves the actual local spatial information of each layer's features.

Both the spatial and temporal transformer blocks follow the architecture of Vision Transformer (15) where each block that makes up the transformer encoder is made up of three key processing units: the Layer Norm, the Multi-head Attention Network, and the Multi-Layer Perceptrons. The encoder is composed of many blocks (MLP). Layer Norm ensures that the training process stays on track and

enables the model to adjust to the differences present in the training pictures. Multi-head Attention Network is a network that is accountable for the development of attention maps based on the embedded visual tokens that are provided. These attention maps assist networks in concentrating their attention on the most vital parts of a picture, such as artifacts.

3.5 Discriminator: Temporal PatchGan

Artifacts might appear in arbitrary locations within a frame in an old film. Thus, It is important to take into account both global and local features within the same frame, along with the temporal stability of the features. Introducing a loss function to each of the three factors separately would not be a proper approach to follow. Observationally, Chang et al. (9) discovered that it is challenging to regulate the weights of various loss functions, particularly when they involve GAN losses. Including GAN loss is a highly popular technique to improve the realism of image restoration results. In substituting the usage of global and local GAN, Yu et al. suggested an effective SN-PatchGAN (70). This architecture applies GAN loss on extracted features of the discriminator. Their method addresses the loss weight adjustment problem and resolves the unrestricted image artifacts issue. However, it does not take into account temporal stability, which is essential for high-quality video restoration. They provided inspiration for (9) to further incorporate the temporal aspect and create a unique Temporal PathGAN discriminator that emphasizes various spatial-time aspects in order to analyze both the global and local spatial as well as temporal features.

The temporal patchGan discriminator has six three-dimensional convolutional layers, each with a filter shape of 3x5x5 and a stride of 1x2 pixels. To improve training constancy, spectral normalization (42), newly suggested, is used in GAN (generator, discriminator) training. Additionally, we employ the hinge loss as the loss function to determine whether the frame is real or not. Because each 3D convolution layer in the discriminator has a filter dimension of 3x5x5, each extracted feature's receptive field spans the entirety of the input video, negating the requirement for a global discriminator like that found in (69). The Temporal PatchGAN is limited to focusing on high-frequency feature information since it only discourages the size of patches after learning to recognize each spatial-temporal patch as real or fake. T-PatchGAN might effectively enhance the output frame quality since the l1 loss function already concentrates on low-frequency characteristics.

3.6 Activation Functions

3.6.1 LeakyReLU

The ReLU activation function has been considerably upgraded because of the introduction of the LeakyReLU function. In the case of ReLU, the gradient is zero for all input parameters that have a value less than zero, the neurons in that area will become inactive, which may lead to a decaying ReLU issue. In order to solve this issue, Leaky ReLU has been proposed (66). Instead of specifying the ReLU activation function as a value of 0 when inputs(x) have a negative value, it is defined as a linear element of x that has a very minimal fraction of amplitude. The equation for the leaky ReLU activation function can be found as follows:

$$f(x) = max(0.01 * x, x) \tag{3.6}$$

This operation will yield x if it is given any positive input. However, if x is provided with a negative value, it will instead return an extremely small number that is equal to 0.01 times x. As a result, it provides an output even for numbers that are negative. After making this one simple adjustment, the gradient that runs down the left side of the graph is transformed into a value that is not zero. As a result, we would no longer come across neurons that had died in that place.

3.6.2 Sigmoid

There is another name for the sigmoid activation function, and that name is the logistic function. It is the same function that is involved in the method for the categorization of logistic regression data. Any real number can be used as an input for the function, and it will return output in the range from 0 to 1. If the value of the input is greater, then the value of the output will be nearer to 1. On the other hand, if the value of the input is lower, then the value of the output will be nearer to 0.0.

The following formula is used to derive the sigmoid activation function:

$$s(x) = \frac{1}{1 + \exp -x} \tag{3.7}$$

Where exp stands for the theoretical constant that serves as the foundation of the natural logarithm.

3.7 Proposed Methodology

So far, we have discussed the components of the proposed methodology. Figure 3.5 demonstrates how these components are organized in a single framework for old film restoration tasks. Now, the generator part of our proposed GAN architecture is going to be explained. A frame sequence consisting of 5 frames is collectively passed as the input for the model. Each frame is grayscale and so has a single channel and the input dimension is (120, 216). The frame restoration process is described in the following steps:

- Initially, the model has a 2D convolution layer with a kernel size of 3x3. The number of input channels is one since grayscale input frames are fed into the model. Since this layer has 64 convolution filters, each filter generates individual feature maps and so a total of 64 feature maps are produced in this layer. The layer has a stride of 2 which means the input data is downsampled with a kernel sliding of 2 steps at a time. As a result, the output features have a dimension of 60x108 instead of 120x216. Anyway, the overall input shape is (8, 5, 1, 120, 216), where each of the values represents the batch size, number of frames per batch, number of input channels, frame height and width, respectively.

- We have another 2D convolution layer with the same kernel size 3x3 as the second layer of the proposed encoder-decoder. However, twice the number of previous layers' filters are included in this layer which generates 128 output feature maps. The stride value is also 2 in this layer to have more downsampled features.

- After two consecutive convolution layers, a multi-scale attention residual block is added to the framework. The objective of adding a multi-scale attention block is to apply spatial and channel attention mechanisms on the feature map extracted using multiple kernel size 3x3 and 5x5. In this layer, the input and output feature dimensions and channel are the same.

- A 2D convolution layer is involved again with a kernel size of 3x3 and in this case, the stride is set to 1. So, no downsampling is taken place in these layers. However, the feature maps got increased from 128 to 256 in this layer.

- After a potential amount of feature extraction, the hierarchical encoder is placed to facilitate the more meaningful token generation for the temporal and spatial transformer blocks. A set of five convolution layers with identical kernel size (3x3) and stride (1) is formed to extract features and combine them hierarchically.

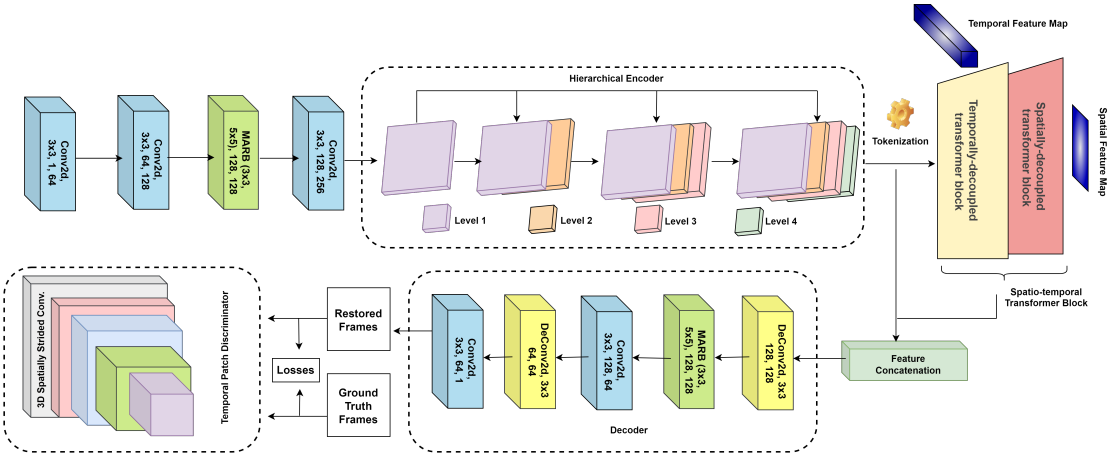

Figure 3.5: *The architecture of the proposed DSTT-MARB model that is initiated with two traditional 2D convolution layers followed by a Multi-scale Residual Block (MARB). Then just after another Conv2D block, a Hierarchical Encoder is placed to obtain hierarchical features. Temporal and spatial feature attention is applied by inclusion of temporal and spatially decouple transformer blocks. The decoder part is comprised of convolution, deconvolution and MARB block. Finally, a temporal patchgan is included that works as a discriminator in the GAN architecture.*

- Then, temporal and spatial transformer blocks with individual objectives are stacked one after another to analyze the extracted hierarchy features in both temporal and spatial domains. The refined features obtained from the transformer blocks are concatenated with features generated from step 4 to preserve enough feature information in the following image reconstruction part performed by the decoder.

- The downsampled concatenated features are upsampled with a deconvolution layer having kernel size 3x3 and padding equal to 1. The output feature dimensions are upsampled from 30x54 to 60x108.

The Deconvolution unit represents a layer that executes a function that is diametrically opposed to the convolution operation. In addition to this name, this process is also known as transposed convolution, which more accurately describes the underlying mathematical activity. Because convolution is not an invertible technique in and of itself, we are unable to directly reverse the process and move from the result to the input. Alternatively, the deconvolution units need to learn their weights in the same manner that the convolution layers do. The functionality of deconvolutional blocks is comparable to that of convolution blocks, but the impact of stride is the reverse.

- Again, for the second time in the architecture, we have a Multi-scale Attention Residual Block (MARB) and the input and output feature maps are the same, which is 128. This layer is followed by a 2D convolution layer that is used to have more selective feature maps for the second upsample operation.
- The final upsampling is done with a deconvolution layer that has the same parameter as the previous deconvolution layer. In this layer, the input features with 60x108 dimensions are upsampled to the output features dimension of 120x216. The proposed model ends with a 2D convolution layer to get the final output grayscale images.

4 Experiments and Ablation Studies

At the beginning of the chapter, the whole data collection and data preprocessing process will be demonstrated. Following that, a brief introduction on the several candidate components in terms of model architecture that are included during experiments is provided. This chapter also explains the loss functions and hyperparameters selection for the model training. Finally, the overall experiment process is discussed in a chronological way to sum up the chapter.

4.1 Dataset Collection and Preprocessing

4.1.1 Frame collection

The acquisition of long-term spatial-temporal features is essential for the completion of several video processing jobs. The investigation of spatial-temporal information for video segmentation is typically confined by the range of accessible video segmentation data sources; for example, the previous largest video segmentation dataset just includes ninety short videos. This means that end-to-end sequential learning to analyze spatial information for video segmentation is extremely difficult. In order to find a solution to this issue, Xu et al. (67) construct a massive video object segmentation dataset that is addressed as the YouTube Video Object Segmentation dataset (YouTube-VOS).

In order to generate the dataset, the authors first identify a set of video categories that include animals (including bears, dogs, camels, and people), vehicles (such as buses, bicycles, trucks, and sharks), items (such as eyeglasses, hats, and bags), and common objects (such as knives, signs, umbrellas), as well as humans engaging in a range of activities (e.g., tennis, skateboarding). They also gather films about humans using a list of action tags to improve the variety of human movements and habits. Since the videos depicting human activities have a diverse range of

Figure 4.1: *A subset of the large-scale dataset YouTube-VOS (67) that includes images with diverse objects, environment and lighting conditions*

appearances and motions, they use this method. The majority of these videos show engagements between a human and a related item. The whole group consists of seventy-eight different categories that span a wide range of different things and actions, and they are intended to be typical of daily situations. After that, they acquire a wide variety of high-resolution clips from the large-scale video classification dataset YouTube-8M proposed by Abu-El-Haija et al.(6) and categorize them using the categories that they have chosen.

This dataset includes millions of videos hosted on YouTube that are connected to over 4,700 different types of visual elements. They make use of the category annotations that it provides in order to get potential clips that catch our attention. Utilizing videos from YouTube to build the dataset has a number of beneficial effects. To begin with, objects in YouTube videos may take a wide variety of forms and move in various ways. YouTube videos often include a number of challenging scenarios, including occlusions, rapid object movements and changing appearances. These scenarios may be particularly difficult to analyze. Second, due to the fact that videos on YouTube may be uploaded by both experts and hobbyists, varying degrees of camera movement can be seen across the crawling clips. It is possible that techniques that are trained on such data would be better able to manage camera movement and, therefore, be much more effective. The last point to make, but certainly not the least, is that modern smartphones capture many videos on YouTube, and there is a large market for the ability to categorize items in such videos for use in multimedia applications.

For this research, we are not taking into account the whole YouTube-VOS due to the time constraints to conduct training on the whole dataset. Instead, a subset of 103845 images is selected for this research while still maintaining the diversity in the dataset in terms of image contents and environment.

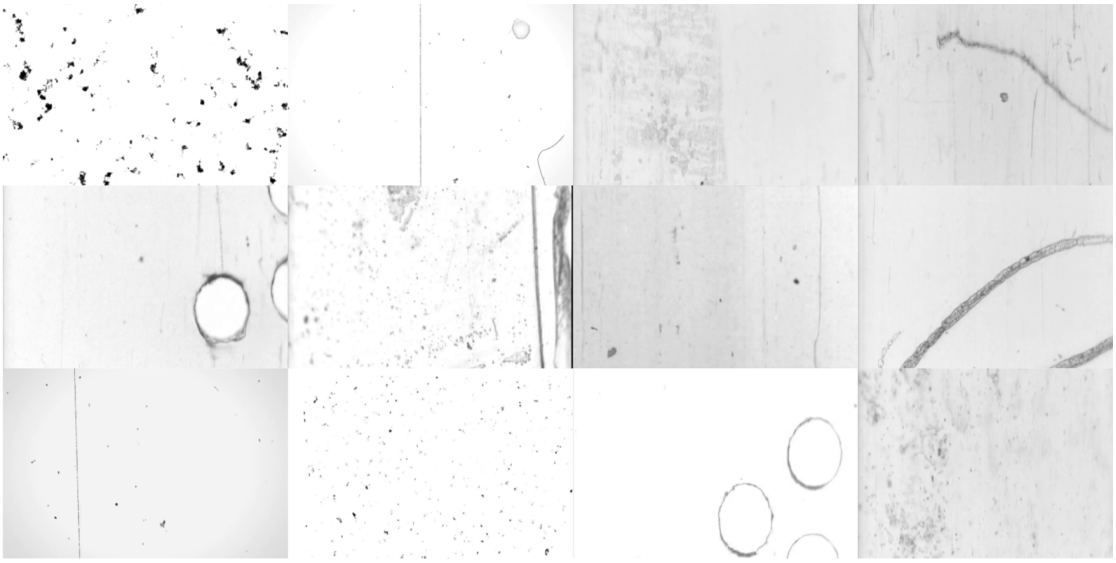

Figure 4.2: *Samples from the noise dataset representing a subset of typical old film artifacts.*

4.1.2 Noise collection

Iizuka et al. (23) propose a old film noise dataset which we include in this research. The authors model synthetic old film degradation to simulate film artifacts, including varieties of old film artifacts. According to the authors, these instances of degradation were painstakingly compiled by searching the web for "film noise" and also created using programs like Adobe After Effects. The foundation noise pattern for produced noise is created using fractal noise, which is then enhanced by adjusting the contrast, sharpness, as well as brightness to produce film artifacts such as scratches and dust.

In this research, we include the noise dataset and expand the dataset with more noise images. These additional noise images are collected online and mostly from the online streaming site YouTube (5). The keyword for noise video searching is “old film dust and scratches." These noise frames are overlaid on the clean reference frames to generate synthetic noisy frames. There are several blending techniques that will be discussed later. After the data collection process, we gathered 5770 images of artifact/noise images. It should be mentioned that the noise dataset can be divided into subcategories: black old film artifacts and white old film artifacts. The reason is that in a typical old film, artifacts appear in both black and white forms. During the noise collection process, noise images that produce black artifacts are collected and a simple operation of those noise images contributes to the white

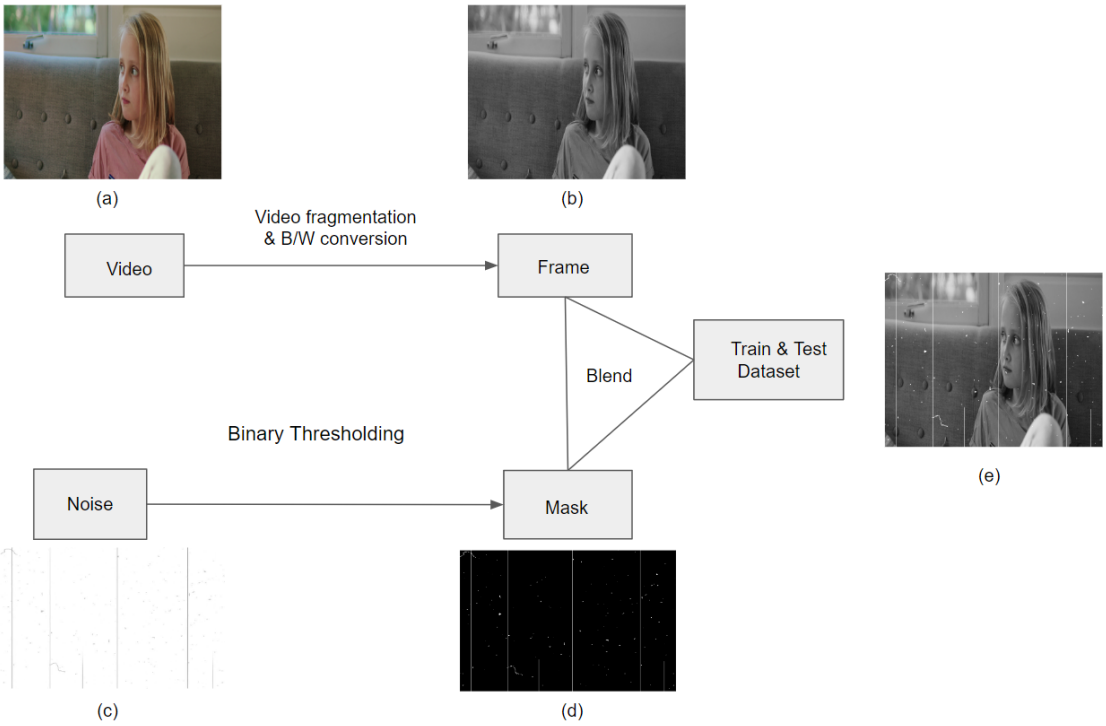

Figure 4.3: *An example of simulating white artifacts on the frame using white masks*

artifacts. The noise image samples that are shown in Figure 4.2 are responsible for simulating black artifacts on the movie frames. The same amount of images are produced by applying inverse transformation on the noise images that are used to introduce white artifacts on the input frames. As a result, a total of 11540 images are considered as the initial dataset for the old film restoration task. This number is raised more with the augmentation and noise fusion processes, which are explained in the latter part of the section.

4.1.3 Data processing for video inpainting

The dataset processing for video inpainting tasks is slightly different than video denoising. As the task of video inpainting seeks to patch up missing video segments, creating artificially generated missing regions on the clean frames is the main objective of this preprocessing phase. Figure 4.3 depicts the overall data preprocessing scheme for video inpainting tasks. Initially, all the collected frames with RGB channels are converted to grayscale frames. In the case of noise images, the intention is to generate masks from those noise images that help to mask out arbitrary regions from the clean image.

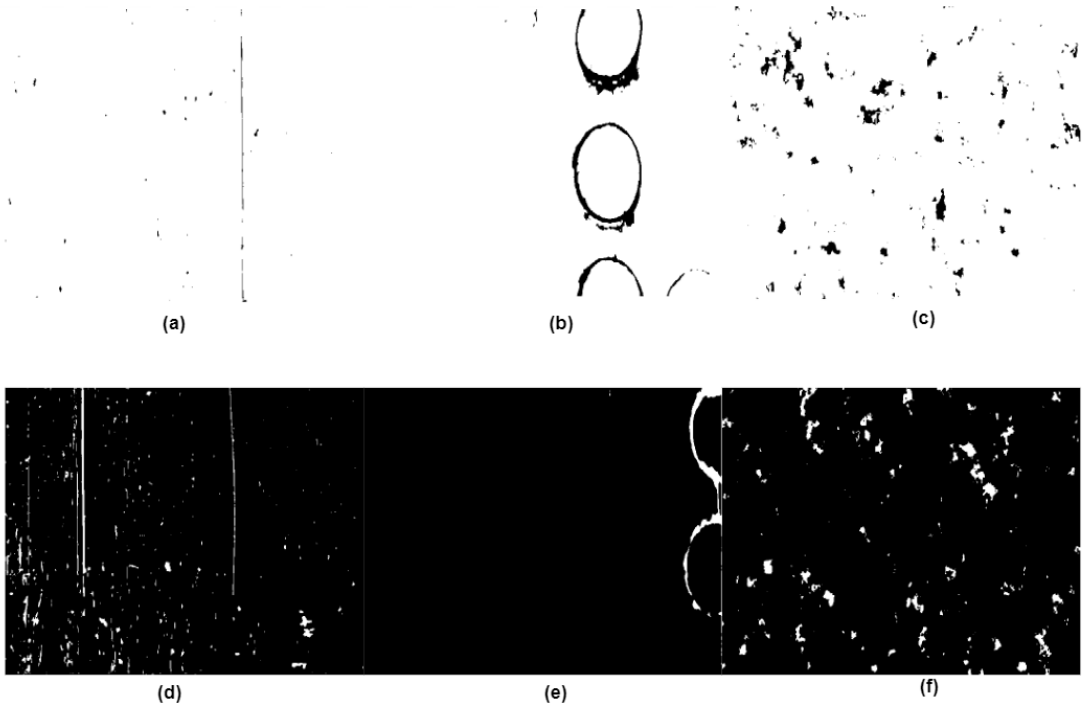

Figure 4.4: *Black (first row) and white(second row) binary masks used for masking out pixels from the clean images to simulate black and white artifacts, respectively*

The binary thresholding technique is employed as a tool for producing masks. The condition for the thresholding process is that all pixels in the noise image whose intensity values are greater than a certain threshold value are activated or set to 255, while all other pixels are kept inactive or set to 0 (Figure 4.4 (a), (b), (c). These binary images are used to simulate black artifacts on the clean images. On the other hand, Figure 4.4 (d), (e), (f) that are the inverse transformation of the Figure 4.4 (a), (b), (c) are applied to simulate white artifacts on the frames. The threshold selection is also kind of tricky in this case because not all the noise images have the same range of pixel intensity. So, after some trials, the threshold value is set to 190. So, for each noise image, if the pixel value is greater than or equal to 190, then it is set to 0 (for white mask) or 255 (for black mask) and vice versa.

So, the white mask (Figure 4.4 (d), (e), (f)) is responsible for generating white artifacts whereas the black masks (Figure 4.4 (a), (b), (c)) are created to produce black artifacts on the frames.

The blending of the black and white masks is conditioned on the clean frame in the following way:

- If the pixel value in the white mask image is 0, the corresponding pixel in the clean frame will not be affected elsewise it will be set to 255.
- And if the pixel value in the black mask image is 255, the corresponding pixel in the clean frame will not be affected elsewise it will be set to 0.

Performing the blending technique, images with black and white holes (shown in Figure 4.3 are developed for model training. Finally, we have a total of 11540 black and white masked images before augmentation.

4.1.4 Data processing for video denoising

The data preprocessing phase for video denoising is straightforward compared to video inpainting. As the aim is to remove the old movie artifacts, the dataset must contain all sorts of typical old film noise patterns. It helps the model to be robust towards a wide range of old film degradation. A subset of the noise patterns, as well as their corresponding blending techniques, are demonstrated in Figure 4.5. According to Figure 4.5, Initially, all the clean frames are converted into grayscale images before the blending operations, as it is reasonable because of the monochromatic nature of the typical old film. This research incorporates three different blending modes: multiply, screen, and merge-grain.

Blend modes are a component of digital image processing and computer vision that are used to decide the manner in which two images are blended together. There are many more methods that exist to blend two images together. The standard blend mode, in most cases, basically obscures the bottom layer by concealing it with whatever is present in the upper layer. However, there are several diverse approaches to combining two images. Several blending modes are considered that are applied to the collected 103,845 images based on the noise categories. Three categories of noise images are included in this research: noisy images with black noise, noisy images with white noise (the inverse of black noise), and noisy images with fused noise (noise fusion will be discussed later). For these three individual artifact types, Multiply, Screen and Grain-merge blending modes are employed, respectively.

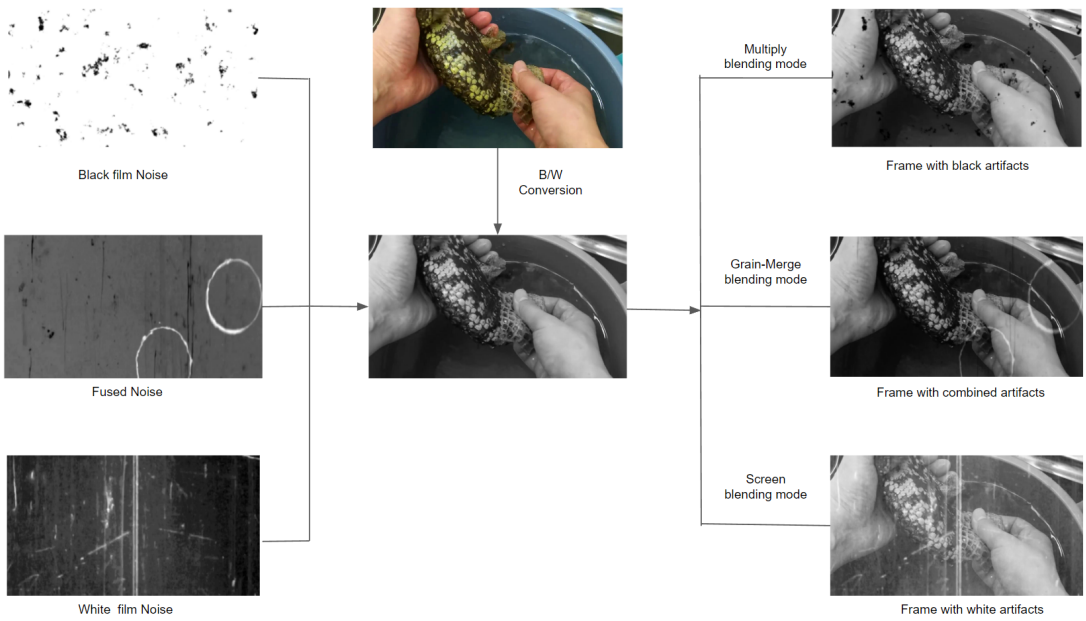

Figure 4.5: *The data preprocessing steps for training the model in video denoising approach. The process includes grayscale conversion and three separate blending techniques for three old film noise categories*

4.1.5 Data Augmentation

Deep learning methods typically require a significant amount of training data that is not always readily accessible. This prevents the models from making accurate predictions. As a result, the already available data is extended in order to provide a more generalized model. The technique of producing additional data points from previously collected data is known as data augmentation (46). This procedure is used to artificially increase the quantity of data that is gathered. This may include making insignificant adjustments to the data or employing machine learning models in order to produce additional data points inside the latent space of the existing data in order to expand the dataset.

Although data augmentation may be used in a number of other fields, the most prevalent application for it is in computer vision. We apply a considerable number of data augmentation parameters, which are provided in Table 5.1 to the synthetic noisy/masked frames as well as the ground truth frames with the intention of accomplishing the following goals: Firstly, we would like to significantly boost the generalization of the proposed framework to a wider variety of video formats; second, we want the model to be capable of restoring a variety of artifacts that are

Table 4.1: *A set of noise-level and frame-level augmentation parameters and their corresponding values and probabilities. The 'Target' columns indicate on which image set the corresponding augmentations are applied (f and c refer to the clean frames and blended noisy/masked frames, respectively.*

	Name	Target	Probability	Range	Additional Information
Noise Augmentation	Horizontal and Vertical Flip	n	0.5	-	-
	Zoom	n	0.5	10-30 (%)	-
Frame Augmentation	Horizontal and Vertical Flip	f, c	0.5	-	-
	Scaling	f, c	1	0.7-0.9	resizing images with a scaling factor range
	Rotation	f, c	1	10-25	rotated randomly from 10-25 degree clock or anti-clockwise
	Random Brightness	f, c	0.2	0.8-1.2	increasing as well as decreasing the intensity level of the image
	Random Contrast	f, c	0.2	0.9-1.0	-
	Image Compression (JPEG)	c	0.9	15-40	the value range represents the image quality range where 15 is the lower and 40 is the upper bound
	Gaussian Blur	c	0.5	3-5	filter size for the Gaussian distribution to use is in 3-5 range
	Gaussian Noise	c	0.1	0-0.04	-

typically present in the video content, including blur, brightness fluctuation, etc.; and finally, we want to ensure that the model is able to produce accurate results.

4.1.5.1 Noise-level Augmentation

Some data augmentation parameters in terms of the geometric transformation of the images are considered to apply to the artifact images. The selection of these parameters is decided carefully, taking into account the realistic aspect of the film artifacts (e.g., old film scratches appear vertically with the variable lengths within the frame). So, for example, applying rotation transformation on scratch images is not feasible to consider as a noise-level data augmentation parameter. For artifact or noise image augmentation, only geometrical transformation parameters are applied. In the following Table 5.1, zoom with a range of 10%-30% and horizontal and vertical flip are included as the augmentation parameters. These kinds of geometric transformations help to augment the noise dataset while maintaining the realism of the old film artifact pattern. So, five augmented noise images are generated out of a single noise image, resulting in a total of 34,615 noise images for simulating white old film artifacts on the clean frame. The same number of noise images are generated to simulate black old film artifacts on the training frames. Finally, at this point, we have 69,230 black and white old film artifacts after applying noise augmentation.

4.1.5.2 Frame-level augmentation

These kinds of augmentation parameters are applied to the clean and blended noisy/masked frames. The blended noisy frames (addressed as 'c' in Table 5.1) and the corresponding clean frames (denoted as 'f' in Table 5.1) are subject to the frame-level augmentation techniques. Most of the image augmentation parameters are encountered in this phase. Most of the image augmentation parameters are considered based on the state-of-the-art old film restoration research (23). Some typical geometrical transformation parameters, like horizontal and vertical flip, scaling, and rotation, are applied to both clean and noisy frames. Most of these parameters cause spatial pixel location changes in the images. So that is why, to maintain the pixel correspondence on both the clean and noisy frame, it is required to apply these kinds of augmentation parameters in parallel on both sets. However, as mentioned earlier, to be responsive to other pixel level frame degradation like blurriness, gaussian noise, intensity and contrast variance, and lower image quality, we have included a number of pixel level data augmentation parameters. Jpeg compression, gaussian noise, and random changes in brightness and contrast, gaussian blur are the pixel level augmentations considered in this

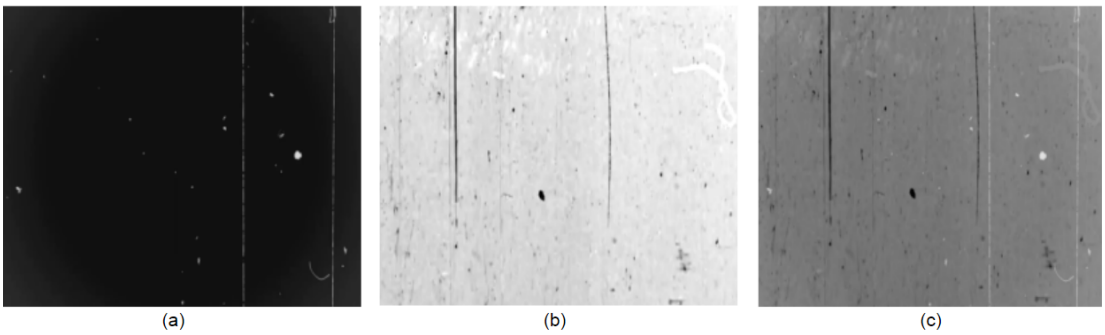

Figure 4.6: *(a) and (b) are the sample noise images to blend on clean frames that simulate noisy frame with white and black old film artifacts. (c) is another noise image that is produced by the fusion of (a) and (b) images. (c) includes both black and white old film artifacts in it*

research. Most of these augmentation parameters are applied only to the noisy image so that during training, the model is able to learn to resolve these kinds of distortions in the restored outputs.

4.1.6 Noise Fusion

We have discussed the collection and augmentation of the black and white artifact image dataset. However, in addition to considering black and white artifacts separately, the noise fusion technique is adopted to simulate combined black and white artifacts in a single frame. In order to achieve it, two black and white noise images are blended on each other to produce a combined noisy image like Figure 4.6. This image blending is applied to each of the 34,615 black and white noise pairs that construct a fused noise dataset of another 34,615 images. Therefore, the final noise dataset consists of 103,845 black, white, and combined noise images for the proposed old film restoration model training.

4.2 Deformable Convolution

Traditional convolutions have a conventional fixed kernel size or predefined grid sampling for feature extraction. However, deformable convolutions introduce a two-dimensional offset/deviation to the kernel. It makes it possible to alter the kernel dimensionality in a free-form manner. Additional convolutional layers are used in order to learn the offsets based on the feature information that came before them. Consequently, the deformation is constrained by the input characteristics in a way that is both locally dense and adaptable. Traditional convolution and a

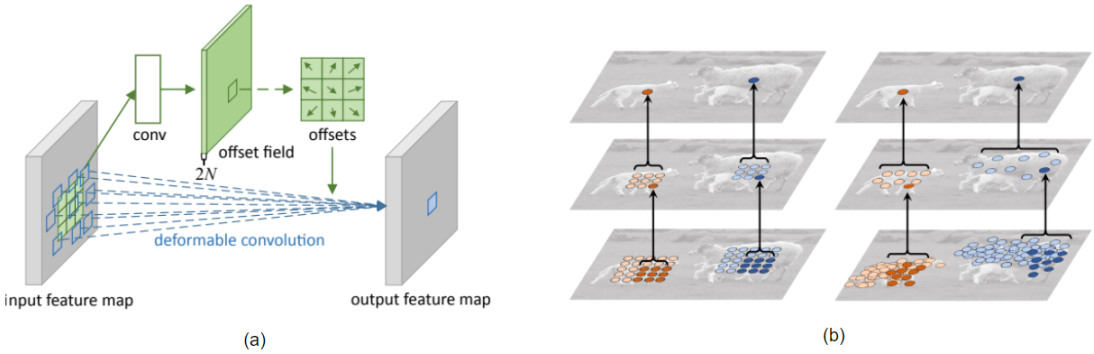

Figure 4.7: *(a) Architecture of Deformable Convolution layer with a learnable offset (b) visualization of the fixed and variable receptive field of the tradition convolution layer and deformable convolution layer*

second convolution layer that learns two-dimensional offsets for each input are the two elements of deformable convolution. So, the two-dimensional convolution is broken down into two stages:

- Subdivide the input feature space F into the small region (sampling) with the help of a grid k.
- Summing up the values of the samples using a weighting factor w. The area and dilatation of the receptive field are both defined by the grid K.

So, if we form that in an equation, the regular convolution can be expressed as follows:

$$y(m_0) = \sum_{m_n \in \mathbb{R}} w(m_n).k(m_0 + m_n) \tag{4.1}$$

whereas the equation of the deformable convolution can be:

$$y(m_0) = \sum_{m_n \in \mathbb{R}} w(m_n).k(m_0 + m_n + \Delta m_n) \tag{4.2}$$

In equation 4.1 and 4.2 m is a variable that iterates across the positions in R. In the process of deformable convolution, the kernel k is modified by adding offsets $m_n|n = 1, ..., N, whereN = |R|$.

At this point, the sampling is being done at $m_n + \Delta m_n$ positions, which are erratic and off-center. The process of obtaining the offsets Δm_n is shown in Fig 4.7(a) and involves putting a convolutional layer on top of the original input feature map. The spatial area, as well as dilation of the receptive field, are identical to

those of the currently active convolutional layer (for example, likewise 3 x 3 kernel dimension having dilation of 1 in Fig 4.7). It should be noted that the spatial area of the resulting offset fields is the same as that of the input feature space. In training, both the kernels for producing the output features as well as the offsets are concurrently learned. Backpropagation of the gradients via bilinear processes is done so that the offsets can be learned.

4.3 Dual Attention Block

In the field of computer vision, the phrases "spatial" and "channel" are likely the ones that are used the most, particularly in reference to the convolutional layers. Tensors are used as the input to all convolutional layers, and tensors are also used as the output of these layers. This tensor is distinguished by the three-dimensional notion, which is as follows: h, which stands for the height of each feature map; w, which stands for the width of each feature map; and c, which stands for the total number of channels. As a result of this, the typical notation for the dimensions of the input is (cxhxw). The feature maps can be thought of as individual slices of the cube-shaped tensor; alternatively, the feature maps can be thought of as layers on top of one another. The number of convolution filters determines the value of the channel dimension. Let's become familiar with the meanings of these phrases before delving into the particulars of spatial attention and channel attention, respectively, so that we can understand why it's important to keep them in mind.

Zamir et al.(72) introduced the Dual Attention Unit (DAU) in order to obtain significant features from convolutional layers. Figure 4.8 presents the DAU architecture in its schematic form. The DAU eliminates elements that are not as helpful and only focuses on the more relevant ones that go to the next stage. This feature adjustment is accomplished via the use of processes such as channel attention (21) and spatial attention (64).

The Channel Attention (CA) branch makes use of the squeeze and excitation procedures in order to take advantage of the inter-channel linkages that are present in the convolutional feature maps. The squeeze operation employs "Global Average Pooling" (GAP) over spatial dimensions to capture a global perspective. As a result, it produces a feature descriptor that has a feature dimension of 1x1xC when assigned to a feature map with the input dimensions M (HxWxC). The activations are produced by the excitation operator after it processes features by way of two convolutional layers, and then the sigmoid function comes into play. When all is said and done, the output of the channel attention block can be derived by the activation. Rescaling the input features with the activation outputs provides the final refined features.

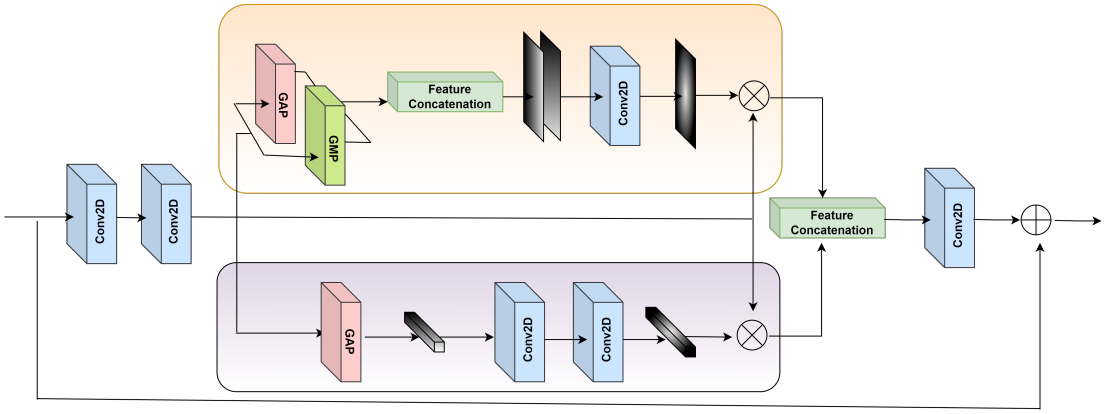

Figure 4.8: *Architecture of the channel and spatial combined attention unit proposed by Zamir et al.(72)*

The Spatial Attention (SA) branch is built with the intention of making use of the spatial relationships that are present in the extracted features. The purpose of spatial attention is to produce a map of spatial attention and make use of it to refine the features. The SA branch must first separately apply two pooling techniques, which are the global average pooling (GAP) operation and the global max pooling (GMP) operation on the input feature maps along the channel dimensions, before concatenating the refined features to construct a feature map with the dimensions HxWx2. This allows the SA block to produce the spatial attention map. After performing a convolution followed by a sigmoid activation function, the spatial attention map is obtained, which can be utilized to rescale the input feature.

4.4 Fused MBConv

An Inverted Residual Block is a sort of residual block that is used for image models that use an inverted structure for the sake of efficiency. This type of residual block is also known as an MBConv Block. The MobileNetV2 CNN architecture was the target of the first proposal for it. Since then, it has been recycled for use in a few other CNNs that are tailored for mobile devices.

The number of channels in a conventional residual block is organized in a framework that goes from broad to narrow to wide. The input has a large number of channels, each of which is subjected to a 1x1 convolution in order to achieve compression. After that, the number of channels is raised once again using a 1x1 convolution in order to accommodate the addition of input and output.

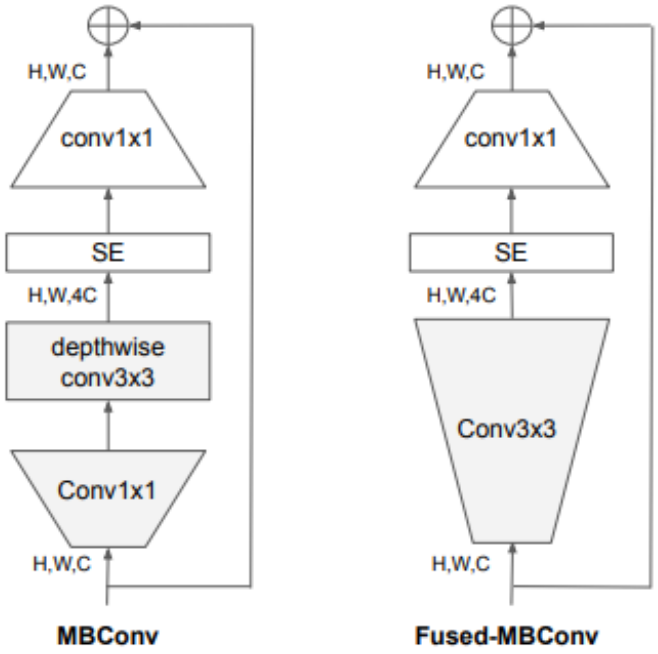

Figure 4.9: *Structure of MBConv and Fused-MBConv. The Figure is collected from (55)*

In comparison, an Inverted Residual Block has a technique that goes from narrow to broad to narrow, which is where the term "inversion" comes from. We begin by widening the image with a 1x1 convolution, then a 3x3 depthwise convolution is included (which significantly decreases the number of parameters), and last, we have a 1x1 convolution to cut down the number of channels such that input and output can be combined.

However, due to the depthwise convolution, the model suffers from high computational costs. However, combining the depthwise convolution along with the expansion convolution into a single typical 3x3 convolution is the answer to the slowing that is caused by depthwise convolutions. Fused MBConv is introduced to substitute the early layers of the EfficientNetV1 designs, which had previously been composed of the MBConv layers. In (55) research, the authors discovered that changing layers 1-3 in the EfficientNet design improved training time with just a modest rise in the size of parameters. Both layers make use of "squeeze and excite" (SE) blocks as a means of dynamically re-calibrating channel-wise feature outputs by modeling interrelations across channels.

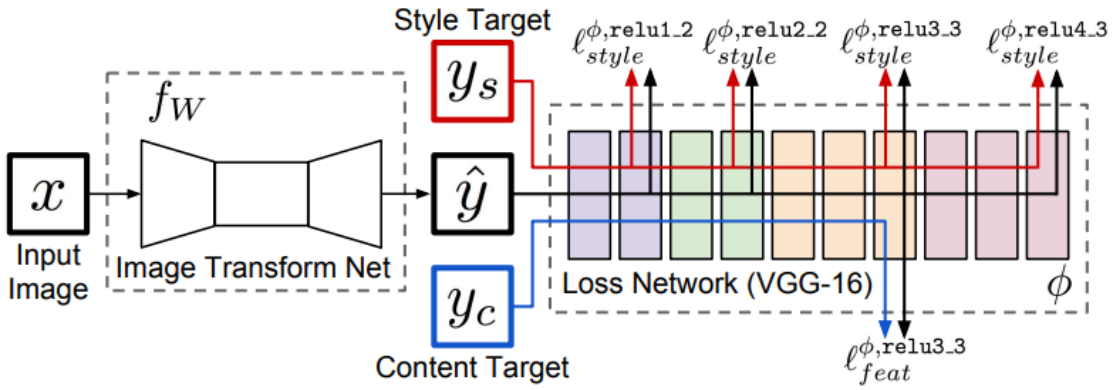

Figure 4.10: *An overview of the perceptual loss calculation technique. To convert input pictures into output images, Johnson et al. (26) build an image transformation network. Perceptual loss functions are constructed using a loss network that has been trained to classify pictures based on their content and style. During training, the loss network does not change.*

4.5 Loss Functions

4.5.1 Perceptual Loss

Calculating the difference between two similar images based on each pixel's value is the working principle of the traditional per-pixel loss function. However, if the images are conceptually similar yet differ from one another by even one pixel, they will be highly dissimilar from one another according to per-pixel loss functions like L1, L2. Per-pixel loss does not take into account the perceptual similarity between the reconstructed and target image. That's because the viewpoint of the image could be identical, but the per-pixel difference might be distinct. This is the reason why it is not always feasible to consider a per-pixel loss function where perceptually refined output is desirable. Recent research has demonstrated that perceptually refined images may be produced by utilizing perceptual loss functions that are modeled by assessing the differences among the feature information (high-level) extracted from pre-trained CNN rather than disparities between pixels. Compared to per-pixel losses, perceptual losses more accurately measure image similarity. Johnson et al. (26) introduced the concept of perceptual loss utilizing two use cases: style transfer and image super-resolution.

The proposed approach consists of two networks where each network accounts for a separate responsibility. According to Figure 4.10, the image transformation network is trainable and trained to predict restored or transformed output images.

This trainable network learns its weight by calculating and backpropagating the loss between the predicted and ground truth image. This disparity is assessed by the loss network, which is basically a pretrained vgg-16 network (53) that is trained on the ImageNet dataset (12). Convolutional neural networks that are trained to classify images have already mastered the art of encoding the perceptual and contextual data that the loss function will assess. That will, of course, rely on the loss network. This pre-trained VGG-16 network is incorporated to formulate feature reconstruction loss by calculating the euclidean distance between the intermediate high-level feature representations. Let us assume $\phi_k(\hat{y})$ and $\phi_k(y)$ are the feature maps with ch*H*W dimension in the k^{th} convolution layer. So, the euclidean distance between the feature maps formalized the image reconstruction loss:

$$loss_{perp}(\hat{y}, y) = \frac{1}{ch_k H_k W_k} \|\phi_k(\hat{y}) - \phi_k(y)\|_2^2 \tag{4.3}$$

Reconstruct an image $\hat{y}$ that reduces the feature reconstruction loss frequently results in images that are identical to the ground truth image y in appearance.

4.5.2 Deep Image Structure and Texture Similarity (DISTS) Loss

Ding et al. (13) propose a perceptual image quality metrics titled "Deep Image Structure and Texture Similarity (DISTS)". The objective is to create a noble full-reference image quality assessment system that includes responsiveness to structural distortions/artifacts caused by, e.g., blur, noise, and so on, with the distribution of the texture. Using a CNN, as is standard practice for perceptual Image Quality Assessment (IQA) approaches, the authors begin by first transforming both the reference and the damaged image into a representation in the feature space. In this feature representation, they build a set of measures that are enough to describe the style of a range of discrete visual patterns. This allows us to more accurately describe the data. Finally, they create an IQA measure by combining the aforementioned texture properties with global structural measures. Unlike the perceptual loss calculation, the DISTS loss calculation is also conducted with the VGG-16 network that is responsible for feature extraction. However, in order to improve texture resampling, the authors bring about some modifications in the vgg net: The authors intend for the preliminary conversion of the images to be aliasing-free so that it could offer a strong platform for the invariances that are required for texture resampling. According to the Nyquist theorem, in order to prevent aliasing when sampling with a factor of 2, it is necessary to blur the signal using a filter whose cutoff frequency is lower than $\frac{\pi}{2}$ radians/sample. In accordance with this guiding concept, the author modified the VGG net such that

all max-pooling layers are now weighted l2 pooling.

In addition, the injective property is a quality that the authors want their transformation to have. This means that different inputs should correspond to different outputs. This step is essential for ensuring that the final quality measure is a valid metric (in the context that the term is used in mathematics) because, if the interpretation of an image is not distinctive, then equality of the output interpretations will not indicate equality of the corresponding input images, even if those representations are identical. This characteristic has been shown to be helpful in perceptual optimization, despite the fact that it is absent from many of the more current approaches. At each level of transformation, the VGG algorithm, like other CNNs, throws away information. In order to guarantee an injective mapping, the authors just involve the input image into the vgg net as an extra feature map at the "zeroth" level of the structure. After that, the representation is made up of the input image, which has been merged with the convolution outputs of five different VGG layers.

4.5.3 Hole and Valid Loss

In the case of the old video inpainting task, mask-based loss functions: hole loss and valid loss are introduced by Liu et al. (35) are applied for this research. Although these hole and valid losses are nothing but L1 losses, the method involved in the frame reconstruction task is kind of tricky. A typical pixel-wise L1 loss function is applied to the whole predicted and ground truth images. So, in that case, the loss function calculates the pixel-wise Mean Absolute Error (MAE) over all the pixel values in the image. However, in both hole and valid loss calculations, the whole predicted and ground truth image are not considered at the same time. As we mentioned in the data preprocessing step for video inpainting, artificial holes have been generated in the frames using mask based approach. The objective of the video inpainting technique is to fill these holes with the contextual/relevant pixel information that maintains the perceptual coherence of the overall frame. Instead of considering the whole region, it is more feasible for an inpainting task to pay attention only to the hole regions which are needed to be filled. Figure 4.11 (a), (b) represent a frame and mask randomly selected from the dataset. Using the mask m, we can derive the hole region by multiplying the mask m with the frame y. Figure 4.11 (c) demonstrates only the missing pixels, which are subtracted from the clean frame in order to obtain the masked frame shown in Figure 4.11 (d).

Now, in order to calculate the hole loss, the mask is applied to both the restored and ground truth images. The formula for calculating hole loss is given below:

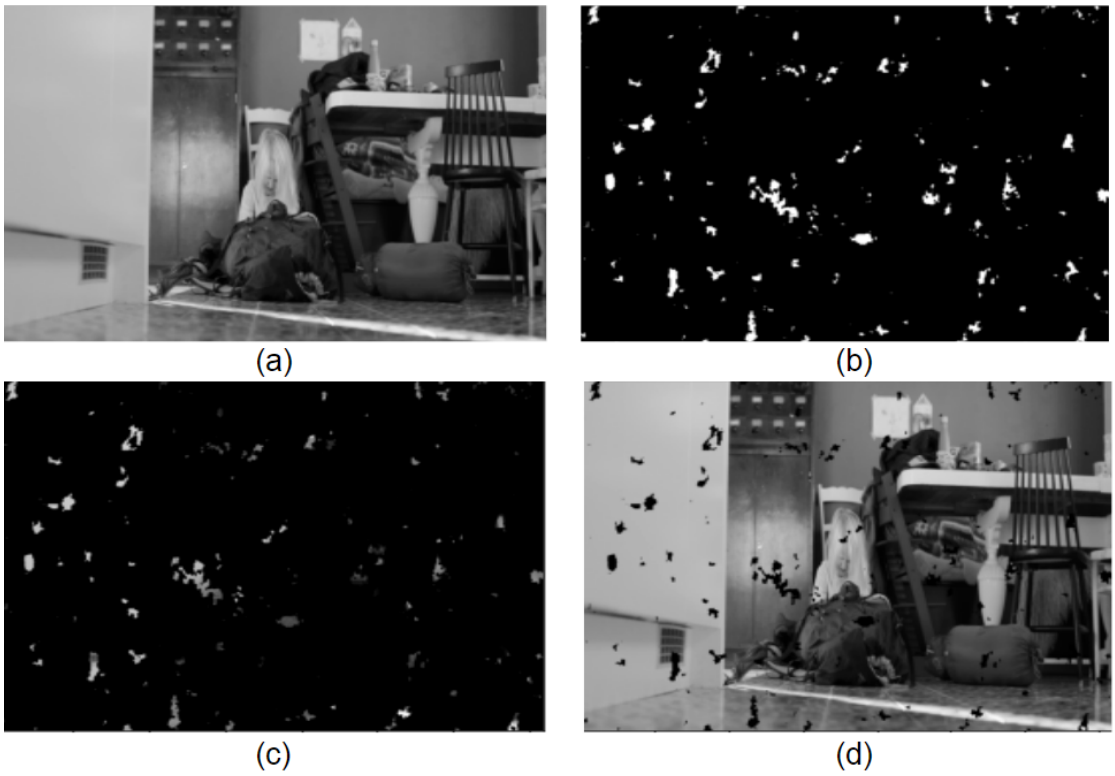

Figure 4.11: *(a) and (b) represent the randomly selected frame and mask from the dataset. (c) and (d) refer to the hole and valid regions of the frame respectively*

$$L_{hole} = \frac{\|m^t \odot (\hat{y}^t - y^t)\|_1}{\|m^t\|_1} \tag{4.4}$$

In equation 4.4 L_{hole}, m, $\hat{y}$ and y refers to the hole loss, mask, restored image by the model and ground truth image respectively.

On the other hand, the equation for estimating valid loss (L_{valid}) is as follows:

$$L_{valid} = \frac{\|(1 - m^t) \odot (\hat{y}^t - y^t)\|_1}{\|(1 - m)^t\|_1} \tag{4.5}$$

From both equations, we can observe that the loss values are divided by m and (1-m), respectively. The reason behind this division is quite tricky because, in the reconstruction process for video inpainting, more focus should be paid to reconstructing the hole region. And most of the pixel values of m are zero, so its mean is also close to 0. That is why the division with m results in adding more weight to the hole loss. The reverse technique goes for valid loss, where valid loss ends up being divided by a higher value close to 1, which ultimately scales down the weight of the valid loss.

4.6 Training Hyperparameters

In machine learning, hyperparameters are variables whose values have a great impact on the learning process. "hyper" signifies that these parameters are at the "highest level" in the process of learning and as a consequence, govern the model parameters which are generated by it. As machine learning researchers who are tasked with developing a model, we select and configure the settings for the hyperparameters that our proposed model will utilize before even beginning the process of training the model. In this context, hyperparameters are considered to exist outside of the model, which is attributed to the reason that the model's properties cannot be altered during the training process (71).

The training process makes use of hyperparameters while learning. However, these hyperparameters are not included in the model that is produced as a consequence of this learning. At the termination of the model training, we will have the trained weights of the model, which are, in a practical sense, what we mean when we talk about the model. The model does not include the hyperparameters that were adjusted throughout the training process. We are only aware of the model parameters that are learned; we are unable to determine, for example, the hyperparameter values that are adapted to train a model based on the model properties. The examples of some common hyperparameters that will be discussed in the following section are optimizer, activation function, loss function, learning rate, batch size, number of training iterations, and model training platform description.

4.6.1 Optimizer: Adam

One way to think of Adam is as a mix of RMSprop and Stochastic Gradient Descent (SGD) that also takes momentum into account (31). It scales the learning rate based on the squared gradients, much as RMSprop does, and it makes use of momentum by calculating the moving average of the gradient rather than the gradient itself, just like SGD with momentum does. As an adaptive learning rate approach, Adam evaluates independent learning rates based on a variety of variables. Adam employs estimation of the first and second moments of the gradient to change the learning rate for every weight in the neural network, thus its name, "adaptive moment estimation." Adam makes use of exponentially moving averages, which are produced based on the gradient that is assessed on a current mini-batch, in order to calculate the moments:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \tag{4.6}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \tag{4.7}$$

Where m and v represent the moving averages while g indicates the gradient on the current mini-batch and betas are the hyper-parameters that are incorporated into the method. They come with incredibly helpful default values, which are 0.9 and 0.999, respectively. These principles are almost never subject to revision. At the beginning of the first iteration, the values of the moving average vectors are all set to zero.

Some significant properties of the Adam optimizer are given below:

- The step size hyper-parameter provides an approximation of the actual value taken by the Adam algorithm throughout each iteration of the process. This characteristic gives previously unintuitive learning rate hyper-parameters a more intuitive grasp than they had before.
- The fact that the step size of the Adam updating algorithm is not dependent on the degree of the gradient is very helpful when traveling through regions with very little gradients (such as saddle points or ravines). In these regions, it is difficult for SGD to traverse them swiftly.

Adam was developed in order to combine the beneficial aspects of Adagrad and RMSprop (49), both of which perform very well in on-line contexts and with sparse gradients, respectively. Adam is also comparable to the combo of RMSprop and SGD when momentum is added to the mix.

4.6.2 Loss Function

As mentioned before, two approaches are adopted in old film restoration tasks, which are video inpainting and video denoising. The loss functions used in both approaches are discussed down below:

4.6.2.1 Video Inpainting

As discussed in section 4.5.3, two loss functions named hole and valid loss is incorporated for training the proposed architecture (section 3.7). These hole and valid loss functions are responsible for reconstructing hole and valid regions of the input frames during video restoration. So, the total loss function for the video inpainting approach can be formulated as follows:

$$L_{total} = \lambda_{hole} \cdot L_{hole} + \lambda_{valid} \cdot L_{valid} + \lambda_{perp} \cdot L_{perp} + \lambda_{adv} \cdot L_{adv} \tag{4.8}$$

In equation 4.8, L_{hole}, L_{valid},L_{perp}, L_{adv} refers to the hole loss, valid loss, perceptual loss and adversarial loss respectively. λ indicates the weight value that regulates the relative importance of the loss function during model training. We empirically set λ_{hole} to 1, λ_{valid} to 1, λ_{perp} to 1 and λ_{adv} to 0.01 for experiments.

4.6.2.2 Video Denoising

In the video denoising part, the L1 loss function is used for pixel reconstruction and the perceptual loss function is included to improve the perceptual quality of the video reconstruction during the experiment. The total loss for old film restoration tasks using the video denoising approach can be formulated as follows:

$$L_{total} = \lambda_{L_1} \cdot L_1 + \lambda_{perp} \cdot L_{perp} + \lambda_{adv} \cdot L_{adv} \tag{4.9}$$

In equation 4.9, L_1, L_{perp}, L_{adv} refers to the L_1 loss, Perceptual loss and adversarial loss respectively. Again, λ indicates the weight value that regulates the relative importance of the loss function during model training. The weight λ_{L_1} and λ_{perp} are set to 1 and λ_{adv} to 0.01 during experiment process.

4.7 Evaluation Metrics

Image Quality Assessment (also known as IQA) is a trait that is regarded to be distinctive of an image. Image quality evaluation is used to quantify the deterioration of images as they are viewed. In most cases, deterioration is measured in relation to a "reference image," which is an image that is considered to be pristine. The quality of an image can be characterized technically and also subjectively in order to demonstrate how much it deviates from a reference. In addition to this, it refers to a person's own interpretation or anticipation of a picture, such as an image depicting a human face. The presence of noise in an image might result in a degradation of the image's quality. The significance of this noise is determined by how well it coincides with the content that the observer is trying to extract from the image.

The processing of visual data can be broken down into several different stages, such as the capture, augmentation, compression, or transmission of the data. After the processing is done, there is a possibility that some of the information that was supplied by the characteristics of an image may be altered. For this reason, human vision perception needs to be used in the evaluation of quality. In everyday life, there are essentially two categories of image quality assessment: subjective and objective assessment. Implementing a subjective review takes a lot of effort and often

results in additional costs. After that, the objective image quality measurements are generated by taking into account a variety of factors.

The objective evaluation of an image's quality may be accomplished via the use of a number of different methods and measurements. These methods are divided into three distinct groups according to whether or not a reference picture is available. These are the types that are in image quality assessment:

- **Full-Reference methods:** The primary objective of this technique is to evaluate the quality of an input image in relation to a reference/ground truth image. This image serves as a point of reference and is regarded to be of the highest possible quality. Take, for instance, a comparison of the original image and the noisy version of the image (47).
- **No-Reference method:** The metrics place their primary emphasis on the evaluation of the quality of a single test image. In this procedure, a reference picture is not utilized at any point (47).
- **Reduced-reference method:** This image quality metric is developed in order to analyze the perceptual quality of a distorted picture using just limited information from the ground truth image (47).

MSE (Mean Square Error), PSNR (Peak Signal to Noise Ratio), SSIM (Structured Similarity Index Method), LPIPS (Perceptual Image quality metric), etc. are a few of the image quality techniques that are widely applied to measure and analyze the quality of images. In this study, SSIM, PSNR and LPIPS methodologies are used to determine the reference-based assessment of the restored test images.

4.7.1 Peak Signal-to-Noise Ratio (PSNR)

PSNR is used as a quality measuring tool in a significant amount of research that is connected to image and signal processing (19). PSNR is determined by taking the logarithm of the mean square error (MSE). MSE conventionally employs the summing-up process as the primary focus of its analysis. Assessment of grayscale images is performed in MSE according to the HxW dimensions.

$$MSE = \frac{1}{HW} \sum_{i=0}^{W-1} \sum_{j=0}^{H-1} [E(j,i)]^2 \tag{4.10}$$

In equation 4.10 H and W refers to the height and width of an image. The Mean Squared Error formula is derived from the aggregation of the square of the reference and distorted image subtraction. This can be understood by examining

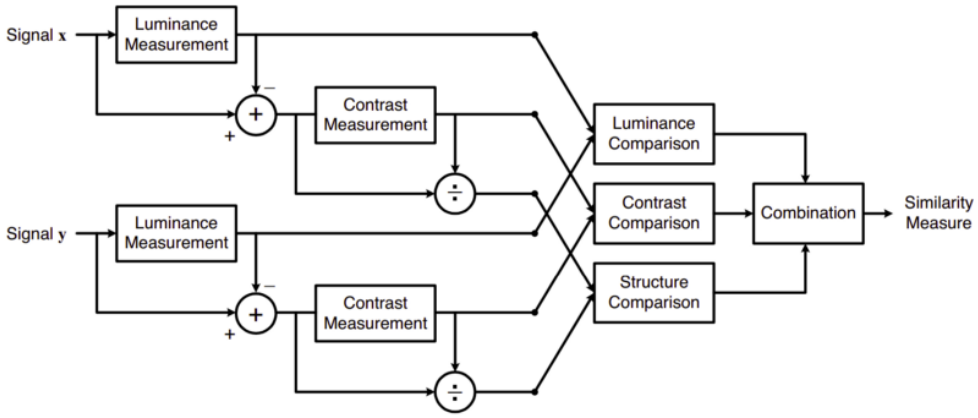

Figure 4.12: *Diagram of the structural similarity (SSIM) measurement system. Figure is collected from (62)*

the equation. PSNR and MSE have a close association with one another due to the fact that the MSE value is utilized in the process of calculating the PSNR value (27).

$$PSNR = 20\log_{10}(\frac{max}{\sqrt{MSE}}) \tag{4.11}$$

In this context, "max" refers to the highest possible scale value in the 8-bit grayscale. It is clear, based on the equations used to compute the PSNR (equation 4.10 and 4.11), that the MSE is the primary constituent of the PSNR, and that the PSNR is made of this value. The disparity in the gray levels that are measured at the same locations is what gives rise to the error value. PSNR will provide a value of infinity and will also not modify the pixel value; on the other hand, if there are more variances in per location gray-level values between the two pictures, PSNR will generate a value that is lower.

4.7.2 Structural Similarity Index Measure (SSIM)

The vast majority of methods for evaluating picture quality include assessing the disparity between a reference and a test image. Quantifying the degree to which the test image and the reference image vary in terms of the values of each of the associated pixels is a typical kind of measurement (MSE, PSNR etc.). However, the human visual sensory system is extremely able to distinguish spatial features from a scene, and as a result, it is capable of identifying the variations in the information derived from a reference and test image. As a result, a metric that successfully imitates this pattern will do much better on tasks that require distinguishing between the reference and test image. During image quality assessment, SSIM

takes into account the three key features of an image: Luminance, Contrast, and Structure (62). The structure and operation of the SSIM system are shown in the following figure (Fig. 4.12): Both the reference image and the test image are referred to as "Signal x" and "Signal x," respectively.

This method determines the Structural Similarity Index (SSI) between two images that are provided, which is a number that ranges from minus one to plus one. A score of +1 suggests that the two images provided are very identical to one another, while a score of -1 denotes that the two images provided are highly unlike one another. These values are often modified to fall somewhere in the range [0, 1], in which both extremes have identical meanings.

4.7.3 Learned Perceptual Image Patch Similarity (LPIPS)

PSNR and SSIM are two quality measures that are often employed in the quality assessment of image reconstruction tasks; nevertheless, they frequently fail to capture changes that are perceptually connected. It is possible for a picture to have a high PSNR or SSIM score but a blurrier appearance than an image created by a generative network but with a lower score. We have also utilized a framework called "Learned Perceptual Image Patch Similarity" (LPIPS), which was offered by Zhang et al. (74). This allowed us to analyze the performance of our technique statistically and numerically evaluate it with other methods on the basis of perceptual attributes. LPIPS makes use of extracted feature maps obtained from the convolutional layers of deep neural networks that have been pre-trained on classification tasks. These networks were then trained on a large dataset that comprised image patches along with human opinion ratings for the purpose of perceptual distance measurement. After that, it determines a perceptual distance between a reference picture and a changed version of it by evaluating them separately in tiny areas and comparing the results. The pre-trained AlexNet model and the default input parameters were used in the computations, which were carried out using version 0.1 of the LPIPS repository. The lower the LPIPS value, the better the test image is similar to the reference image.

4.7.4 Image Spatial Quality Evaluator (BRISQUE)

In the following paragraphs, we will discuss the necessary procedures for the BRISQUE method, which is used for No-Reference IQA. The computation of BRISQUE is broken down into its component parts in Figure 4.13.

Natural and deformed pictures have different pixel intensity distributions. When

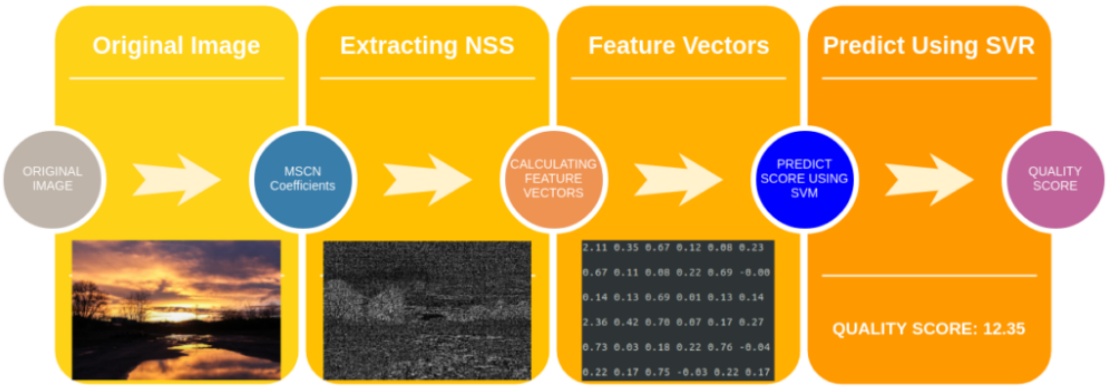

Figure 4.13: *The five steps involved in determining an image's quality using the BRISQUE model. (3)*

we standardize intensity values and determine the distribution across these intensities, the discrepancy is more evident. After standardization, natural image intensity values resemble a Gaussian distribution, whereas damaged image intensities do not. Departure from an ideal bell curve measures visual distortion. Image normalization may be accomplished in a number of distinct methods. One example of this kind of normalization is known as Mean Subtracted Contrast Normalization (MSCN). To extract local relationships, Mittal et al. (41) employ pairwise MSCN image and MSCN image shift products. The pairwise product is determined using the following four viewpoints: horizontal, vertical, left-diagonal and right-diagonal (RD).

Up to this point, a total of 5 pictures from the initial image are produced. These include one MSCN image and four pairwise product images that represent neighbor connections. Afterward, a feature vector with a dimension of 36 by 1 by using these five images is computed. The width and height of the source picture might be arbitrary, but the feature vector will always be 36 by 1. After fitting the MSCN images to a Simplified Gaussian Distribution, the first two components of the 36-by-1 feature vector are determined. After that, every one of the four pairwise product pictures is given a new shape by having an Asymmetric Generalized Gaussian Distribution (AGGD) fitted to it. An asymmetric version of Generalized Gaussian Fitting is referred to as AGGD (GGD).

As a result, a feature vector that has 18 elements is found. The picture is then shrunk to one-half its initial dimensions, and the same steps are replicated in order to generate 18 more numbers, giving a total quantity of 36 values.

An image is initially transformed into a feature vector in the majority of applications that make use of machine learning. After that, an educational program

such as Support Vector Machine is provided with the feature vectors and outputs of each picture that is included in the training dataset.

4.8 Experiments

In this research, several loss functions are taken into account during experiments. The loss functions are chosen based on their individual aspect of contributing to the overall video restoration process. For instance, L1 loss focuses on the pixel-wise difference between the restored and ground truth image, while perceptual or DISTS loss is responsible for maintaining the perceptual consistency of the restored image. All the experiments in this research are conducted on the same dataset and training settings for fare comparison. The whole experiment process is explained in the following steps:

- The experiment starts with retraining of the model proposed by Liu et al. (35) which is addressed as the DSTT in the following discussion. This is the current state-of-the-art model for video inpainting tasks. It is also interesting to explore how well it performs in the case of old film restoration.

- Then, a single-scale attention block combined with channel and spatial attention modules, which is denoted as the Dual Attention Unit (DAU) discussed in section 4.3, are included in both the encoder and decoder structures of DSTT. The reason is to observe if the inclusion of an attention scheme improves the performance of DSTT.

- In addition to applying a single-scale attention module, this research incorporates a Multi-scale Attentive Residual Block (MARB) discussed in section 3.3. In the multi-scale attention block, the attention is applied to the features that are extracted in multiple scales (using multiple kernel sizes to be specific), which helps to obtain better feature representation (able to consider even tiny objects in the image). On the other hand, the features that are gleaned through filters of a single size are inadequate for the construction of discriminative representations of videos.

- Another kind of residual block named fused inverted residuals, explained in section 4.4 is included in both encoder-decoder parts of the DSTT model during the experiments.

- Convolutional neural networks have the intrinsic limitation of only being able to mimic geometric changes since their construction modules consist of fixed geometric shapes. Section 4.2 demonstrates a new mechanism to boost

the transformation modeling capabilities of CNNs. This module is called deformable convolution, which replaces all the convolution layers in DSTT.

The Adam optimizer (31) with the parameters $\beta 1 = 0.9$ and $\beta 2 = 0.999$ is incorporated in this research. The learning rate of this optimizer begins at 0.0001 and decays once with the factor of 0.1. As mentioned before, a total of 103,845 images are used in the model training. The batch size is set to 8, and in each batch, the training module processes five frames at a time. So, each iteration consists of 40 frame-processing. The proposed old film restoration model is trained for 200,000 iterations. In addition, the learning rate is decayed with a factor of 0.1 after 120,000 iterations by observing the oscillation state of the model. A learning rate that is too high will produce weight updates that are excessively high, and the performance of the model (such as its loss) will fluctuate over the course of the training iteration if the learning rate is set high. It is believed that weights that are diverging are the reason for oscillating performance.

4.9 Implementation platform and Hardware Requirements

The implementation of the model training scheme is done using PyTorch as the backend. PyTorch is a scientific computing tool written in Python that takes advantage of the processing power provided by graphics processing units (GPU) (45). It is also one of the most popular platforms for deep learning research, and it was developed to give the highest possible level of flexibility and speed. It is well-known for delivering two of the most significant advantages, namely, developing neural networks on a tape-based autograd mechanism and tensor operations with significant GPU acceleration support. This library is one of the many current Python libraries that have the ability to transform the way machine learning model training is carried out. There are numerous more existing Python libraries with similar potential as well. PyTorch's popularity may be attributed, in large part, to the fact that it is written entirely in Python and allows users to easily construct models of neural networks. When compared to its other rivals, it is still a relatively new player, yet it is making significant headway in a very short amount of time.

Part of the model training is conducted on the Amazon cloud service platform using an Amazon EC2 P3 instance (2). Amazon EC2 P3 instances provide high-performance computation in the cloud for machine learning applications. These instances are equipped with up to 8 NVIDIA® V100 Tensor Core GPUs and up to 100 Gbps of networking speed. It has been shown that using Amazon EC2

P3 instances may reduce the amount of time needed for model training from days to minutes and also increase the number of simulations that can be run for high-performance computing by a factor of three to four. The specific instance type that is employed for model training is p3.2xlarge which has one GPU, eight virtual CPUs, 61 gigabytes of memory, and a network capability of up to 10 gigabits per second.

Another GPU machine from NTNU ColorLab named GPU-6-A014 is employed in the model training. The GPU configuration of GPU-6-A014 is: Alienware Aurora r8 Intel(R) Core(TM) i7-9700K CPU @ 3.60GHz 32 GB RAM GPU : GeForce GTX 2080 , 11GB. In addition to that, for parallel model training, the NTNU GPU cluster is also incorporated. The cluster GPU configuration is NVIDIA Tesla V100 16Gb.

5 Result and Discussion

In this section, we explain and analyze the obtained results from several experiments from both quantitative and visual point of view. In addition, limitation of the proposed model is also discussed in this chapter.

5.1 Quantitative Analysis

5.1.1 Video Denoising Approach

The quantitative analysis part is broken into sub-parts where we will discuss the performance of the proposed methodology in comparison with the current state-of-the-art methods, several network modifications, and experiments with loss functions. Initially, the proposed model is compared to the recent works on video inpainting and old film restoration. The evaluation is done based on the 200 test images each for black and white artifacts. Figure 5.1 represents a subset of the test images that are used during the evaluation. It is worth mentioning that these images are not used during the model training and so are completely unseen images for the models. Anyway, we have a separate evaluation process for black and white artifacts so that it can be demonstrated how these models perform in both black and white artifacts removal.

In this research, progressive modifications of Decoupled Spatial-Temporal Transformer (DSTT) (35) are made in this research in the task of old film restoration. As a result of that, Table 5.1 demonstrates the restoration performance of the several additions to DSTT model in terms of three image quality evaluation metrics: PSNR, SSIM, LPIPS

As an initial state of the upgradation of the DSTT model, a single-scale Dual Attention (channel and spatial) Unit is included. The inclusion of the single scale attention improves the restoring performance by a mentionable margin in all metrics according to table 5.1 for both artifact categories (black and white). In the case of both black and white artifacts removal, the restored images by the single scale dual attention-based DSTT model have higher PSNR values which are 34.362 (black)

Figure 5.1: *A set of sample images from the test image dataset depicting both black and white artifacts*

Table 5.1: *A performance comparison among several versions of DSTT (35) model in terms of three reference-based image quality metrics PSNR, SSIM, LPIPS*

	PSNR		SSIM		LPIPS	
	Black	White	Black	White	Black	White
DSTT	31.631	29.908	0.944	0.940	0.0316	0.0352
DSTT + DAU	34.362	33.778	0.967	0.969	0.0166	0.0166
DSTT + Fused MBConv	33.281	33.564	0.972	0.971	0.0115	0.0142
DSTT + MARB	**35.279**	**34.285**	**0.982**	**0.975**	**0.0072**	**0.0115**

and 33.778 (white). For the DSTT model without the attention module the values are 31.631 (black) and 29.908 (white), respectively. If we also observe the other two evaluation metrics SSIM and LPIPS, the DSTT+DAU network achieves significant improvements. It obtains higher SSIM values 0.967 (black), 0.969 (white) and lower LPIPS values which are 0.0166 for both black and white artifacts removal whereas the SSIM and LPIPS values for DSTT model are in 0.944 (black), 0.940 (white) and 0.0316 (black), 0.0352 (white) accordingly. However, the integration of the Fuse Inverted Residual (Fused MBConv) block achieves a slightly lesser

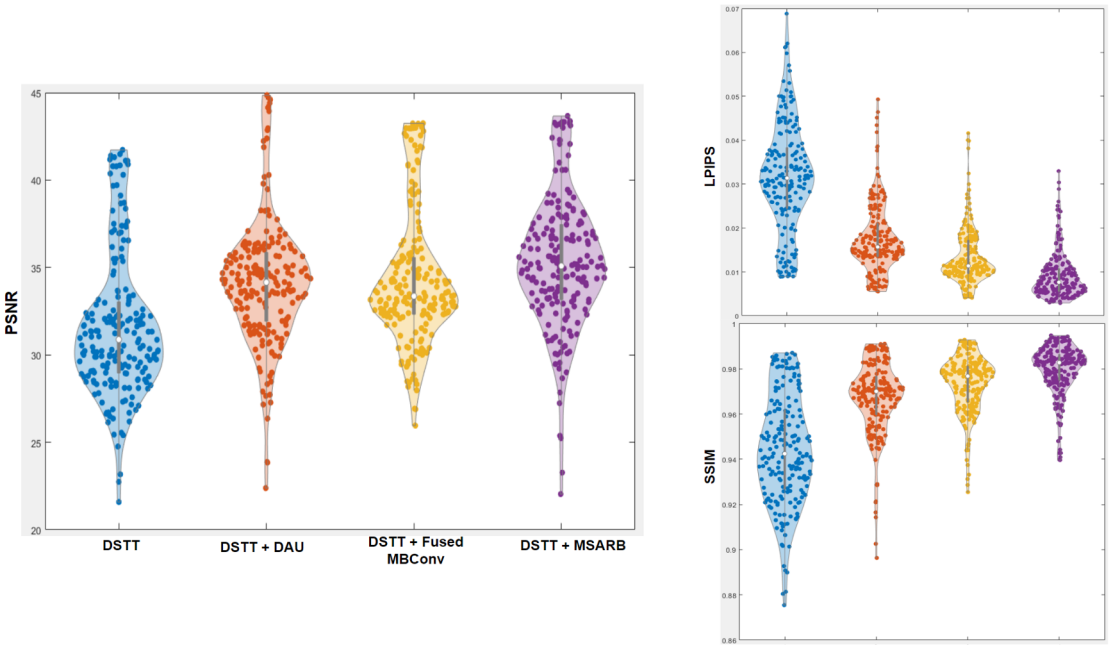

Figure 5.2: *Violin plots depicting the performance of the candidate models on the 200 test images with black artifacts with respect to PSNR, SSIM and LPIPS Image Quality Assessment (IQA) metrics. X-axis indicates the models' name and Y-axis refers to the IQA metrics*

performance in terms of PSNR, which are 33.281 (black) and 33.564 (white). It is not the case for SSIM and LPIPS performance metrics, as the DSTT+MBFuse model overcomes the DSTT+DAU model with a narrow margin by SSIM values of 0.972 (black), 0.971 (white) and LPIPS values of 0.0115 (black) and 0.0142 (white).

The Multi-Scale Attention Residual Block (MARB) continues improvement in the old film artifacts restoration task by achieving superior performance compared to the DSTT, DSTT + DAU and DSTT + Fused MBConv. Following table 5.1, the DSTT + MARB block performs the restoration process on the 400 test images and restored with the PSNR, SSIM and LPIPS values of 35.279 (black) and 34.285 (white), 0.982 (black) and 0.975 (white), 0.0072 (black) and 0.0115 (white) respectively.

To have a clear understanding of restoration performance of the candidate models mentioned in table 5.1, a violin plot is shown in Figure 5.2. A violin plot is used to illustrate the distribution of statistical information for one or more categories. The breadth of each curve is approximately proportional to the number of data points that are located in each area (18). So, Fig 5.2 represents the models' performance comparison mentioned in Table 5.1 in terms of PSNR, SSIM and

Table 5.2: *A performance comparison between ConV2D and Deformable ConV2d layers in terms of three reference-based image quality metrics PSNR, SSIM, LPIPS*

	PSNR		SSIM		LPIPS	
	Black	White	Black	White	Black	White
DSTT+MARB	35.279	34.285	0.982	0.975	0.0072	0.0115
DSTT+Deform. Conv.+MARB	**38.087**	**36.599**	**0.988**	**0.984**	**0.0048**	**0.0081**

LPIPS values in the restoration of the 200 test images with black artifacts (violin plots for white artifacts in provided in the appendix). It should be mentioned that the Violin Plot in Figure 5.2 also provides a brief insight of the standard deviation of restored images' PSNR, SSIM and LPIPS. So, we can observe that the DSTT+MARB model has higher PSNR, SSIM and LPIPS values compared to the other models.

Now, the DSTT+MARB model is reformed by replacing all the 2D convolution layers with the 2D Deformable convolution layer, followed by the LeakyReLU activation function. As mentioned before, deformable convolution layers extract features using a kernel with variable dimensions, whereas a traditional convolution layer uses a kernel with a fixed shape. Quantitative analysis of the test images restored by DSTT+DeformConv+MARB model demonstrated in Table 5.2, reveals that the model achieves higher psnr (38.087, 36.599), ssim (0.988, 0.984) and lpips (0.0048, 0.0081) values compared to DSTT+MARB model which gains (35.279, 34.285), (0.982, 0.975) and (0.0072, 0.0115) values in psnr, ssim and lpips evaluation metrics in (black, white) artifacts removal task. As we can also observe from the violin plot in Figure 5.3 that the presence of the deformable convolution layer boosts the performance of the proposed approach in terms of quantitative analysis. However, the visual results during the real-time evaluation are not as promising as the matric values. The DSTT+DeformConv+MARB model produces some unexpected texture distribution inefficiency in the restored frames which will be discussed in the following section. So, we proceed with the experiments with the DSTT+MARB model.

A loss function-based evaluation of the proposed model is drawn in Table 5.3. As mentioned before L1 loss function is experimented with combined with two

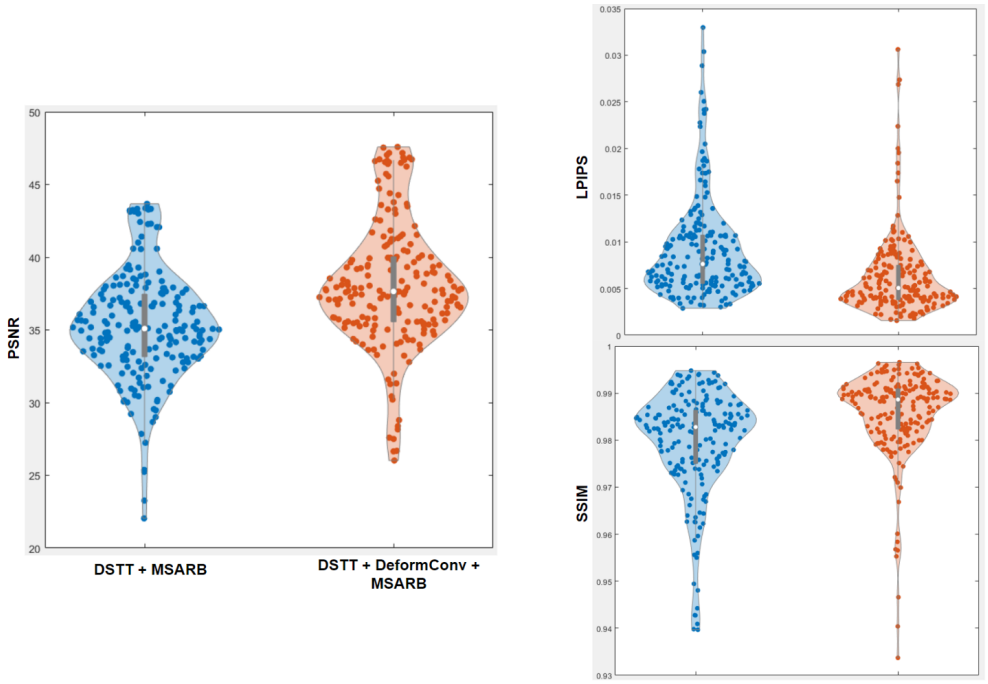

Figure 5.3: *Violin plots depicting the performance of the ConV2D and DeformConV2D on the 200 test images with black artifacts with respect to PSNR, SSIM and LPIPS Image Quality Assessment (IQA) metrics. The X-axis indicates the models' name and Y-axis refers to the IQA metrics*

other loss functions, which are perceptual and DISTS loss functions. Initially, the proposed model is trained with only L1 loss function and then it is trained with L1+Perceptual loss and L1+DISTS loss. Table 5.3 represents the values of the performance evaluation metrics psnr, ssim and lpips for corresponding loss functions. It can be visible that the involvement of the DISTS loss function with L1 loss in the video reconstruction reduces the restoration efficiency of the model. L1+DISTS loss function has the mean psnr, ssim and lpips value of (33.043, 32.645), (0.958, 0.956) and (0.0248, 0.0263) as (black, white) artifacts removal. While the proposed model (DSTT+MARB) involving only L1 loss function during model training results has better psnr (35.279, 34.285), ssim (0.982, 0.975) and lpips (0.007, 0.0115) values in (black, white) artifacts removal. However, a significant improvement in the result is obtained by replacing dists loss function with perceptual loss where the mean psnr, ssim and lpips values reach (37.728, 36.456), (0.984, 0.982) and (0.0067, 0.0085) for (black, white).

From the violin plot in Figure 5.4, a clear performance difference among the proposed model trained with several training losses is visible. The combined loss function of L1 and perceptual loss has superior results compared to the L1 and

Table 5.3: *A performance comparison of the proposed model trained with L1, L1+DISTS and L1+Perceptual loss functions in terms of three reference-based image quality metrics PSNR, SSIM, LPIPS*

	PSNR		SSIM		LPIPS	
	Black	White	Black	White	Black	White
DSTT+MARB+ L1+DISTS loss	33.043	32.645	0.958	0.956	0.0248	0.0263
DSTT+MARB+L1	35.279	34.285	0.982	0.975	0.007	0.0115
DSTT+MARB+ L1+Perceptual loss	**37.728**	36.456	**0.984**	**0.982**	**0.0067**	**0.0085**

Table 5.4: *A performance comparison among the DeepRemaster (23), DSTT (35) and proposed DSTT+MARB model in terms of three reference based image quality metrics PSNR, SSIM, LPIPS*

	PSNR		SSIM		LPIPS	
	Black	White	Black	White	Black	White
Deep Remaster	28.114	27.497	0.912	0.910	0.0578	0.0698
DSTT	31.631	29.908	0.944	0.940	0.0316	0.0352
DSTT + MARB	**37.728**	**36.456**	**0.984**	**0.982**	**0.0067**	**0.0085**

L1+DISTS loss.

And finally, as mentioned in the literature part, DeepRemaster 2.3.3 proposed two networks responsible for two separate tasks: artifact removal and colorization. For old film artifacts removal tasks, they employed a preprocessing network and for colorization tasks, they introduced a source-reference attention module. Due to the fact that in this research, we aim for old film noise removal, we only consider the preprocessing network for training from scratch on the YouTube-VOS dataset.

According to Table 5.4 the DeepRemaster (23) model achieves (28.114, 27.497),

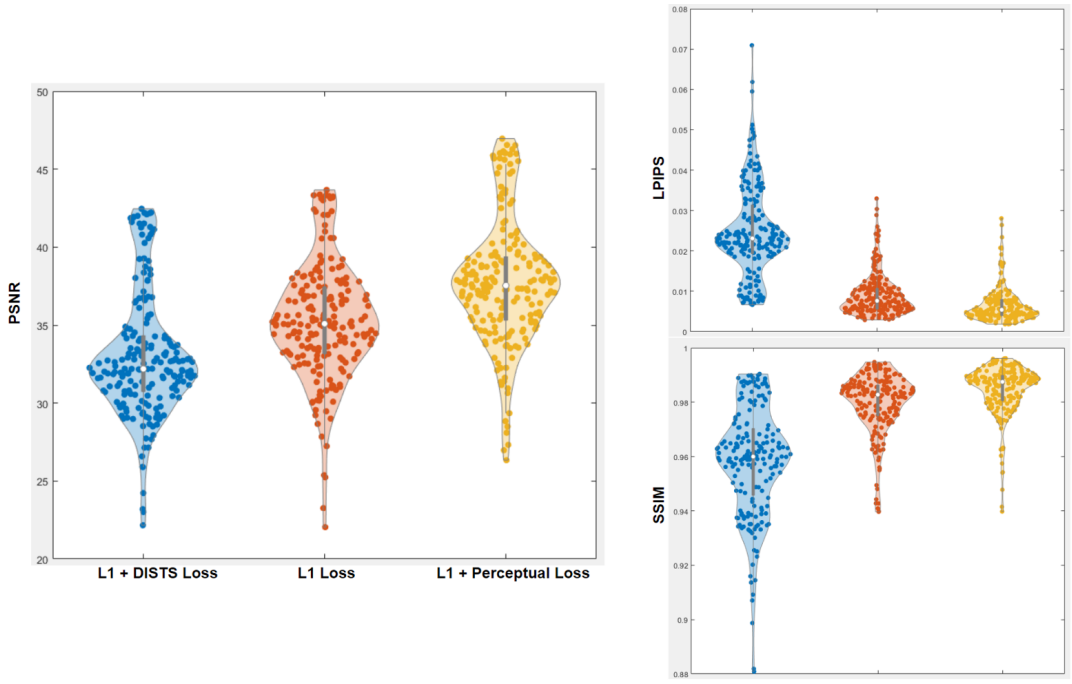

Figure 5.4: *Violin plots depicting the performance of the proposed model trained with L1, L1+DISTS and L1+Perceptual loss functions on the 200 test images with black artifacts with respect to PSNR, SSIM and LPIPS Image Quality Assessment (IQA) metrics. X-axis indicates to the models' name and Y-axis refers to the IQA metrics*

(0.912, 0.910) and (0.0578, 0.0698) in psnr (black, white), ssim (black, white) and lpips (black, white) metrics respectively in the task of artifacts removal. These values are the lowest among the candidate models considered. For DSTT (35) which performs way better than the DeepRemaster model by achieving psnr (31.631), ssim (0.944), lpips (0.0316) in black artifact removal and psnr (29.908), ssim (0.940), lpips (0.0352) in white artifact handling. However, it should be mentioned that DeepRemaster itself is a simple encoder-decoder architecture with 3D convolution involved, and DSTT is a more complex architecture form that is an encoder-transform-decoder-like architecture. On the other hand, the proposed model, which can be considered the improved version of DSTT, outplayed all the models in terms of the three evaluation metrics psnr, ssim and lpips by gaining 37.728, 0.984, 0.0067 and 36.456, 0.982, 0.0085 in the restoration task of test images with black and white artifacts respectively.

Again, the violin plot in Fig 5.5 demonstrates the distribution of psnr, ssim and lpips values of the restored test images by the models in Table 5.4. The restoration performance of the DSTT+MARB has higher performance over the state-of-the-art

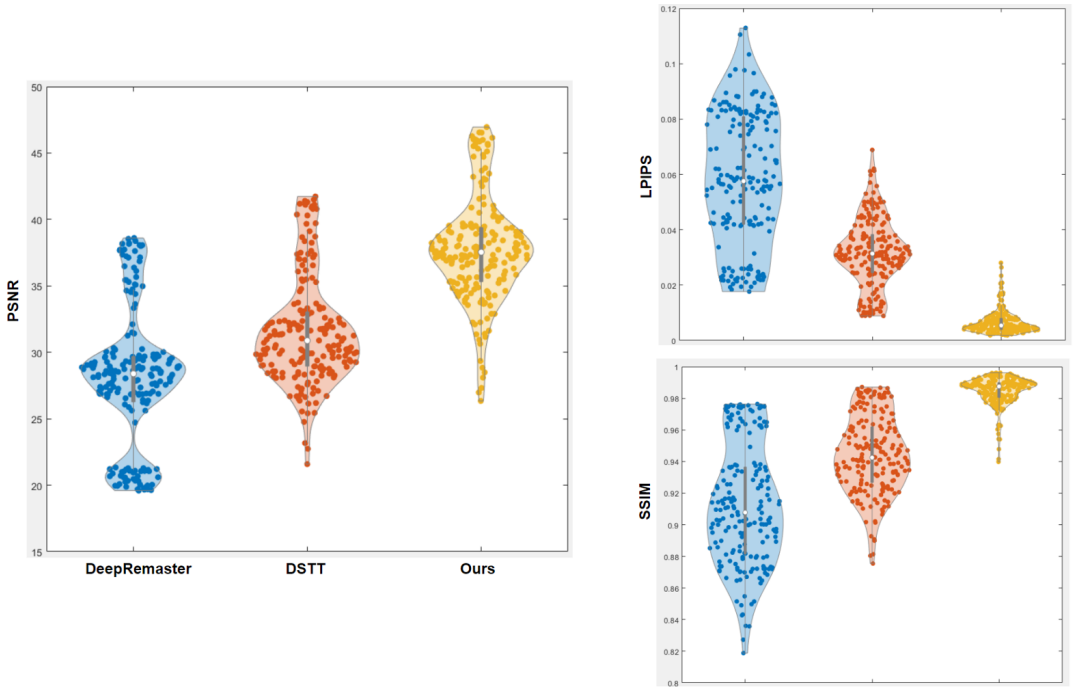

Figure 5.5: *Violin plots depicting the performance of the DeepRemaster (23), DSTT (35) and proposed DSTT+MARB model on the 200 test images with black artifacts with respect to PSNR, SSIM and LPIPS Image Quality Assessment (IQA) metrics. The X-axis indicates the models' name and Y-axis refers to the IQA metrics*

models DeepRemaster (23) and DSTT (35).

5.1.2 Video Inpainting Approach

Similar to the testing of the denoising algorithm, a test image set with 400 test images is targeted to the video inpainting evaluation process of the proposed model. However, the process of generating test images is non-identical to the denoising algorithm. Along with the 400 test clean frames, a set of 400 mask images are selected to apply to the clean frames to simulate masked frames with missing regions. Figure 5.7 first row depicts the artificially generated masked frames that are required to be restored by the proposed model trained in a video inpainting manner. It should be mentioned that 200 test images and masks are selected for generating each of the black and white masked frames. Anyway, Table 5.5 and Figure 5.6 demonstrate the psnr, ssim and lpips values of the restored images by the model. It can be observed that the video inpainting model achieves respectively 30.339 (black) and 33.969 (white) in psnr, 0.934 (black) and 0.959 (white) in ssim,

Table 5.5: *A performance of proposed DSTT+MARB model trained as video inpainting approach in terms of three reference-based image quality metrics PSNR, SSIM, LPIPS*

	PSNR		SSIM		LPIPS	
	Black	White	Black	White	Black	White
DSTT+MARB +hole-valid+ Perceptual loss	30.339	33.969	0.934	0.959	0.0437	0.0249

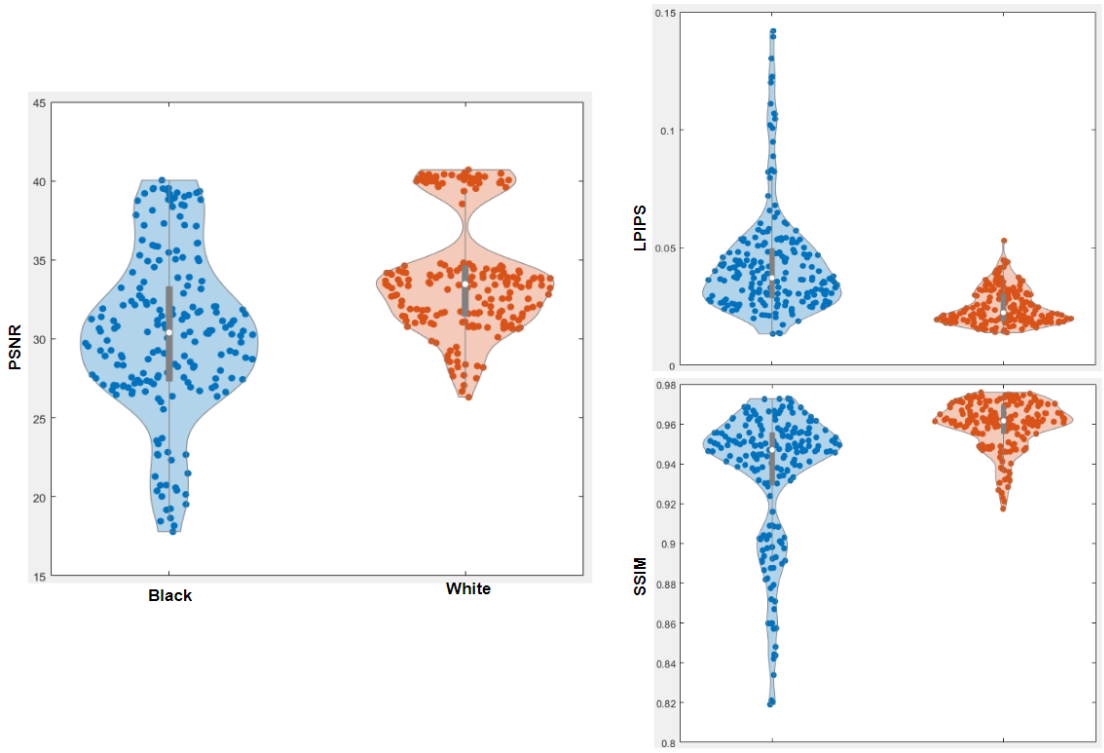

Figure 5.6: *Violin plots depicting the performance of the proposed DSTT+MARB model trained as video inpainting approach with respect to PSNR, SSIM and LPIPS Image Quality Assessment (IQA) metrics. The X-axis indicates the type of artifacts and the Y-axis refers to the IQA metrics*

0.0437 (black) and 0.0249 (white) in the lpips image quality metrics.

Figure 5.7: *The first row represents the masked frames that are artificially generated and the second row refers to the corresponding restored frames by the proposed model trained using the video inpainting approach. The last row depicts the SSIM difference map among the masked and restored frames*

5.2 Qualitative Analysis

In this section, the performance of Figure 5.8 demonstrates the input frame (a) from an old film 'Den-forsvundne-polsemaker', the restored frame by the proposed model trained in video inpainting manner (b) and another restoration of the same frame by the same model trained as a denoising algorithm (c). Along with that, a small rectangular frame region is cropped out to have a better visualization. Finally in the last row of Figure 5.8 represents the structural similarity index measure (SSIM) difference map between (a), (b) and (a), (c). These difference maps help us to visualize the artifacts removal location in the restored frame. Although the video inpainting approach works well on the simulated test images, it is not effective on real-world old film, as we can observe from Figure 5.8 (b). Basically, the trained model is not paying attention to the artifacts. One of the reasons that the inpainting model is not capable of removing artifacts could be:

- **The nature of the old film artifacts:** Inpainting model is objective to fill up the missing pixel of the target frame. By 'missing pixel', it means that the region does not contain any information; thus, the pixel values are either set to 0 (black) or 1 (white). So, the model is trained to restore those missing values. However, real-world old film noise is non-identical to the artifacts seen in the case of video inpainting. If we can observe Figure 4.2, typical film noise is not confined to a specific region. In fact, the noise is distributed throughout the frame. Dust and scratches have a higher and lower intensity

Figure 5.8: *A scene from the film 'Den-forsvundne-polsemaker' where the first row of (a) refers to the input frame and (b), (c) are the restored frames by the proposed model trained in video inpainting and denoising manner, respectively. The third row of the figure indicates the SSIM difference map of the restored frames with respect to the input frame*

compared to the other noise part, causing them to be more visible to the observer.

The restored frame obtained by the proposed model trained as a denoising algorithm depicts that the denoising approach is effective in the case of old film restoration as it removes most of the old film artifacts. Apart from that, if we can observe the difference map, the resolve of the old film artifacts is spotted noticeably. So, here we can have a clear understanding that the denoising method is a more feasible approach to follow in the case of the old film restoration task.

Table 5.2 demonstrates that DSTT+DeformConv+MARB shows higher psnr, ssim and lpips values compared to the DSTT+MARB model where Conv2D layers are used instead of deformable Conv2D. However, the reason for not proceeding further with DSTT+DeformConv+MARB is shown in Figure 5.9. Here, Figure 5.9(a) represents the a input frame from a old film and Figure 5.9(b) and Figure 5.9 (c) depict the the restored frames produced by the DSTT+DeformConv+MARB and DSTT+MARB models respectively. As we can observe from the reconstructed frame with the deformable convolution, DSTT+DeformConv+MARB produces unpleasant

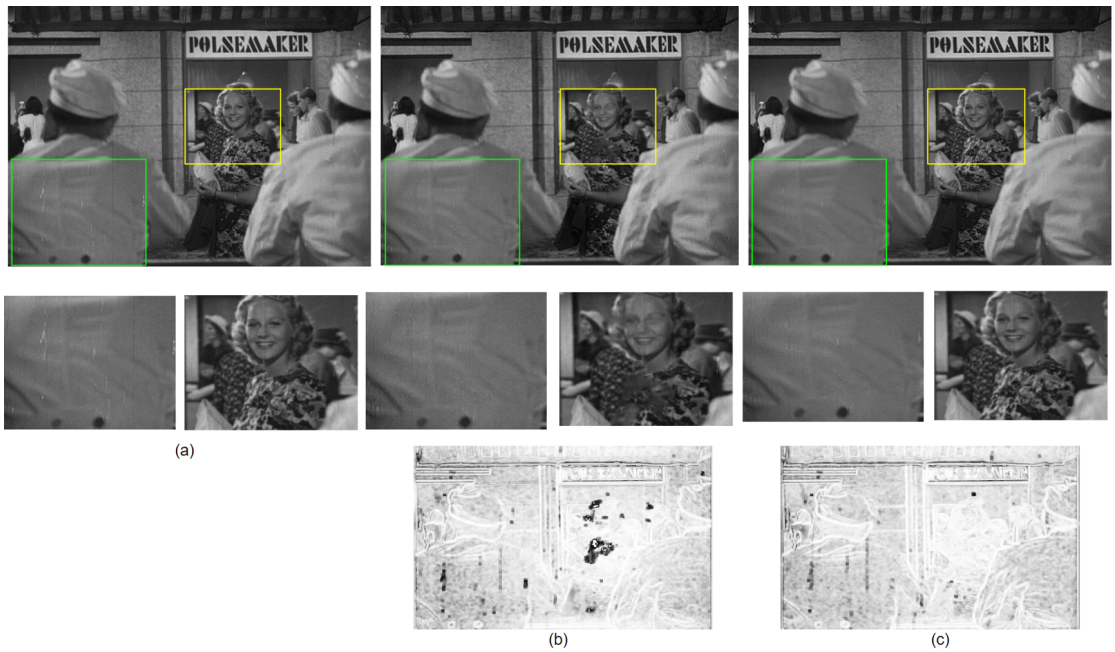

Figure 5.9: *(a) refers to the input frame and (b), (c) are the restored frames by the DSTT+DeformConv+MARB and DSTT+MARB models, respectively. The third row of the figure indicates the SSIM difference map of the restored frames with respect to the input frame*

additional artifacts in the restored frame, although it achieves better visual results in the other part of the frame (green region) compared to DTT+MARB. The DSTT+DeformConv+MARB model is unable to preserve the texture information in the restored frame. One of the reasons could be:

- As it is mentioned that deformable convolution employs a deformed kernel feature extractor shaped by learnable offsets, the kernel shape adjusts more or narrows down to the more identical artifact regions. Due to that, it is unable to extract global contextual information and so performs inferior when we have a high-frequency region in the frame.

The limitation of the DSTT+DeformConv+MARB is visible in the ssim difference map where high disparity appears in the non-artifacts region Figure 5.9 (b). On the other side, the DSTT+MARB model is able to preserve texture information and also restore the artifacts region by removing them. From the ssim difference map Figure 5.9 (c) also, it is clear that high differences exist only in the artifacts regions.

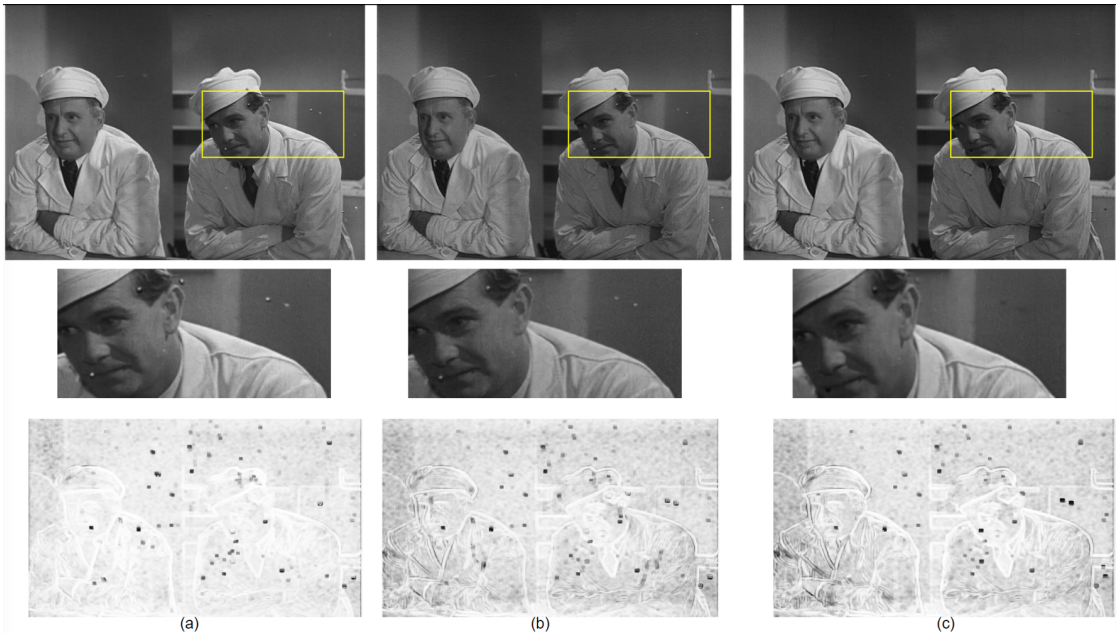

Figure 5.10: *(a), (b), (c) are the restored frames by the DeepRemaster (23), DSTT (35) and proposed DSTT+MARB models respectively. The third row of the figure indicates SSIM difference map of the restored frames with respect to the input frame*

Figure 5.10 demonstrates a visual comparison of the proposed model with the current state-of-the-art model, DeepRemaster (23) and DSTT (35), in the old film restoration task. If we analyze the restored frames by the DeepRemaster and DSTT models, the DSTT has better restoration than the DeepRemaster model, which is observable from Figure 5.10 (a) and Figure 5.10 (b). However, any of the models are able to completely remove the artifacts. On the other hand, the proposed model successfully achieves superior performance among the models as it attempts all the artifacts in the frame and restores the underline information.

The real-world old film 'Den-forsvundne-polsemaker' that the model is evaluated on, does not have any clean reference frames to assess PSNR, SSIM or LPIPS. Due to that reason, a no-reference image quality assessment metrics Brisque (discussed in 4.7.4) is included for model evaluation. Table 5.6 and Figure 5.11 demonstrate a comparison of the proposed model against the state-of-the-art models DeepRemaster (23) and DSTT (35) model on the basis of brisque image quality matrics. It is undoubtedly visible that the proposed model, which is DSTT+MARB trained in a denoising manner, has a lower mean brisque value (42.871) than the other two models. It should be mentioned that the brisque value range is [0, 100], where the

Figure 5.11: *Violin plot depicting the models' performance comparison in terms of Brisque values of the restored frames*

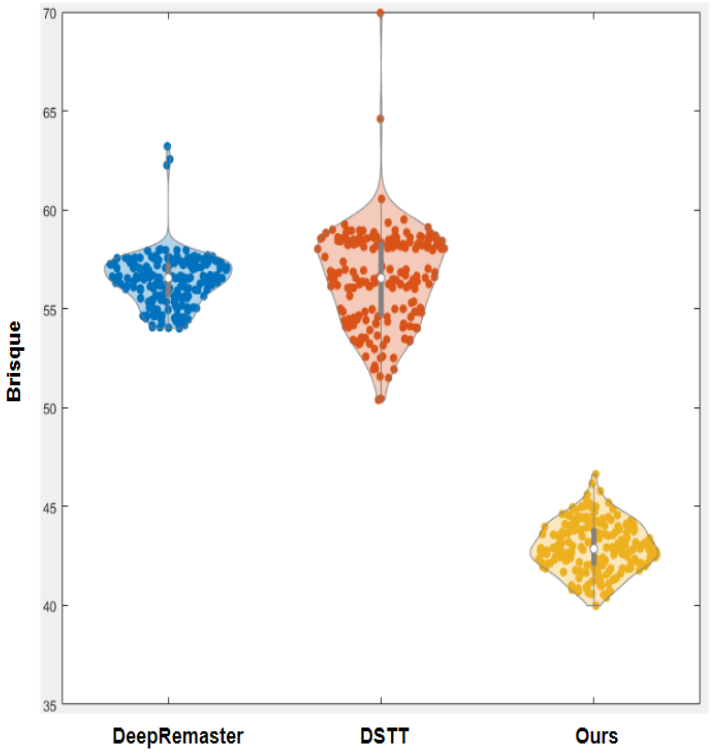

Table 5.6: *Models' comparison based on Brisque IQA matrics*

	Brisque
Deep-Remaster (23)	56.552
DSTT (35)	56.546
Ours	**42.871**

lower values indicate better human perceptuality.

5.3 Limitation

Despite having a remarkable performance of the proposed MSARB-SST model in the old film restoration task, it has a limitation which we discuss in this section. The model is trained on varieties of old film degradations (e.g., dust). So, during the restoration process of the train, it targets the dust in the input frame and removes those artifacts. The exception happens when the model misjudges a frame content as artifacts and removes that content in the restored frame. That means in the restored frames, some information is disappeared, which is not supposed to be removed. Figure 5.12 is a good example of this kind of scenario. Here, we can observe that the proposed model removed the reflection of the light in the background since it is misjudged as dust by the model. As a result, the light

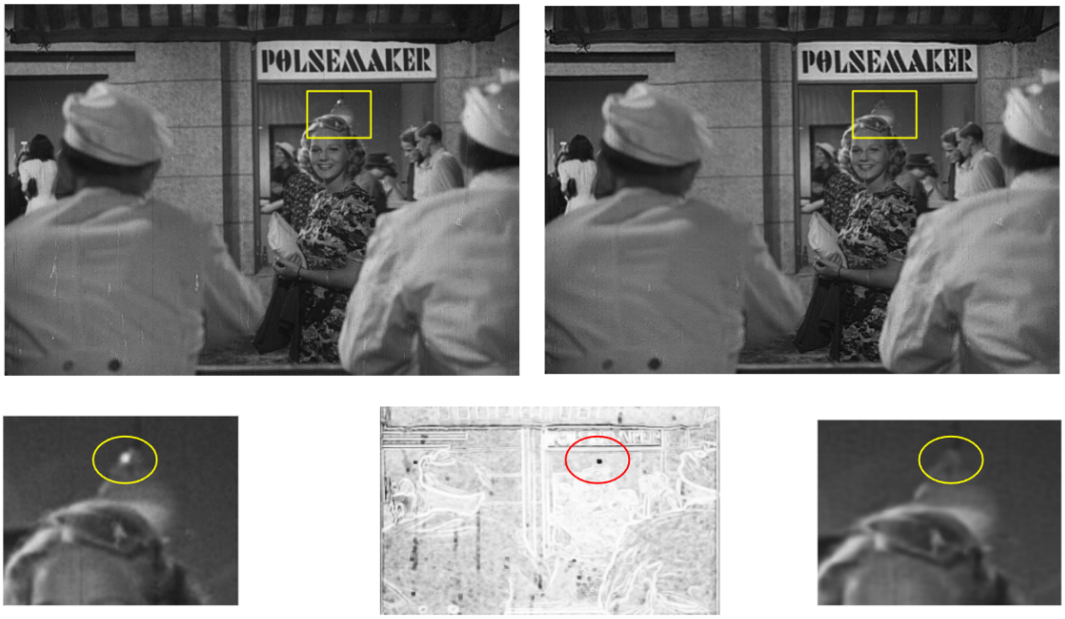

Figure 5.12: *The use case when the model fails to preserve the frame content misjudging light reflection as artifact*

reflection does not exist in the restored frame. This can be addressed as a limitation of the model. Wan et al. (59) recommend a solution to this kind of limitation that is to include more images in the dataset so that a more generalized and well-trained model can be obtained that have a better understanding of the old film artifacts.

5.4 Conclusion

Among the research questions that are presented in section 1.1, one is to find out an effective approach to follow between the video inpainting and denoising techniques. Although video inpainting is a well-established method to reconstruct missing information in a video, it is not impactful in the case of old film restoration. The real-world old film artifacts are more diverse than the reconstruction of the missing portion of the image. Because the old film noise patterns are not confined to a particular region, they are distributed over the whole frame region without following a uniform distribution. So, old film restoration tasks can be identified as more relatable to the video denoising algorithm. However, old film degradation is also dissimilar to the typical image noise like gaussian noise, blurriness, and compression that follows a certain statistical distribution. Old film noise patterns are completely random in nature as the appearance of these kinds of artifacts (for

example, dust, scratches) on a frame can not be predefined spatially and temporally. In this research, both spatial and temporal coherence is preserved during video reconstruction by involving two transformer blocks, taking into account both spatial and temporal aspects of the video. That is how the temporal consistency of the restored video sequences is preserved without involving optical flow estimation, which answers the second research question of this research. This work is highly inspired by the Decoupled Spatial-Temporal Transformer for Video Inpainting (DSTT) model proposed in (35) which is basically a GAN architecture, and it is developed for the video inpainting context. We suggest an improvement to the DSTT model by introducing Multi-scale Attention Residual Block (MARB) in both encoder-decoder parts of the generator. The advantage of the inclusion of MARB lies in having more comprehensive features that help models understand content at various scales.

For the quantitative evaluation of the proposed methodology tested on the artificially generated noisy test images, reference-based image quality metrics PSNR, SSIM, LPIPS are incorporated. A separate but similar evaluation process is done for both black and white artifacts. The experiments are conducted among the several versions of the DSTT where the single-scale channel and spatial attention residual block, fused MBConv block, deformable convolution, and MARB are the candidate components. In addition to that, loss-based experiments are also included, where perceptual loss and DISTS loss are combined separately with the L1 loss function.

Apart from the quantitative analysis of the model, a visual output analysis of the proposed method is also included in the evaluation process. State-of-the-art models DeepRemaster (23), DSTT (35) and DSTT-MARB model are tested on a Norwegian comedy film named "Den-forsvundne-polsemaker" released in 1941. The proposed DSTT-MARB model achieves superior performance in no-reference image quality assessment metrics (BRISQUE) than the other models. However, as mentioned earlier, the proposed model still has the limitation that sometimes fails to distinguish the video content from artifacts, and as a result, it removes the content when it is not supposed to be removed. One solution to that could be to include more diverse videos in the dataset during training so that the model overcomes this kind of shortcoming and it can be addressed as one of the future research objectives to improve the old film restoration system.

A | Appendix

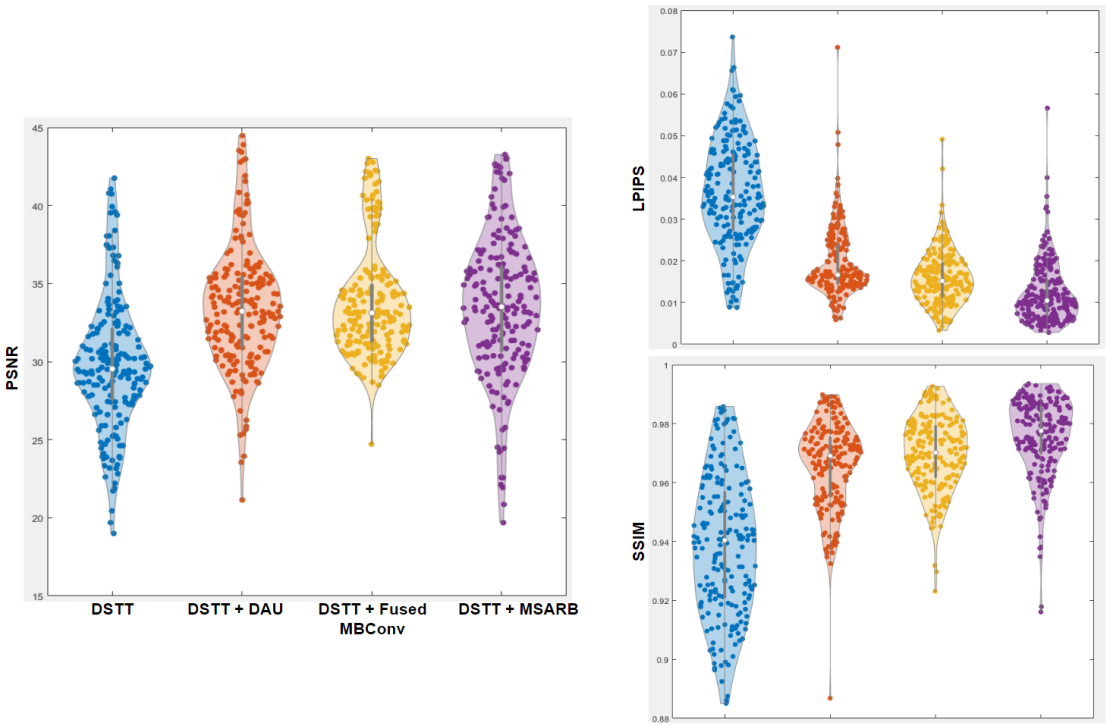

Figure A.1: *Violin plots depicting the performance of the candidate models on the 200 test images with white artifacts with respect to PSNR, SSIM and LPIPS Image Quality Assessment (IQA) matrices. X-axis indicates to the models' name and Y-axis refers to the IQA matrices*

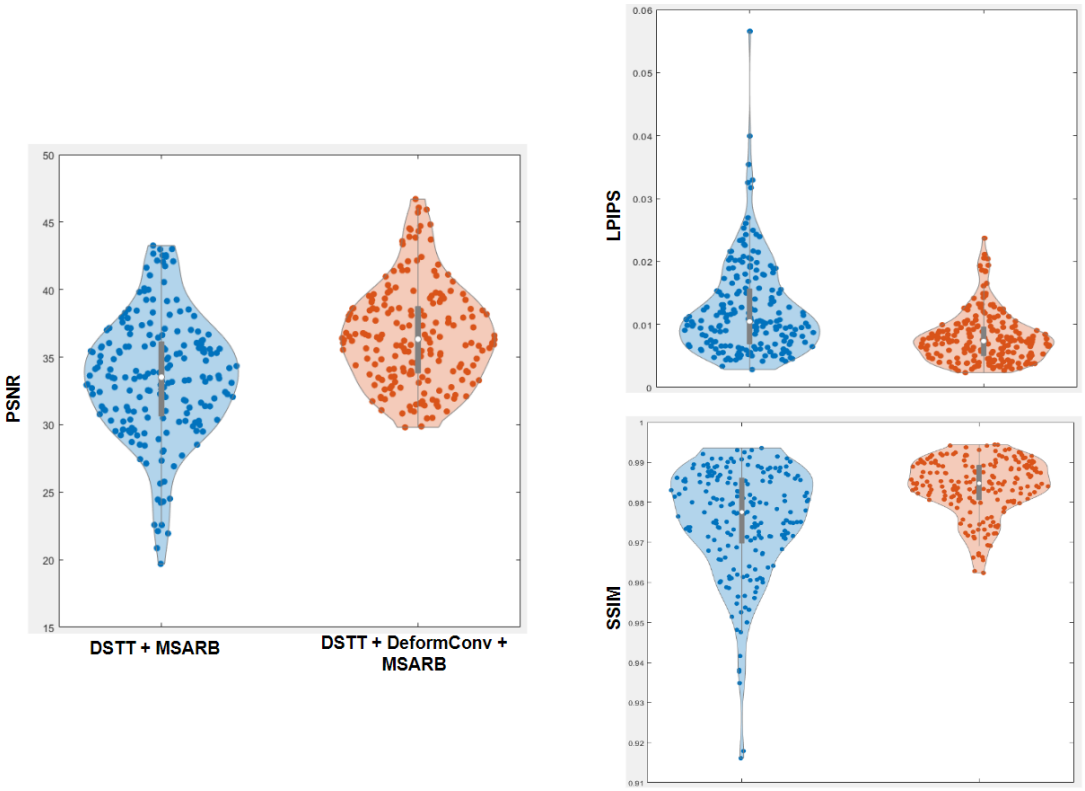

Figure A.2: *Violin plots depicting the performance of the ConV2D and DeformConV2D on the 200 test images with white artifacts with respect to PSNR, SSIM and LPIPS Image Quality Assessment (IQA) matrices. X-axis indicates to the models' name and Y-axis refers to the IQA matrices*

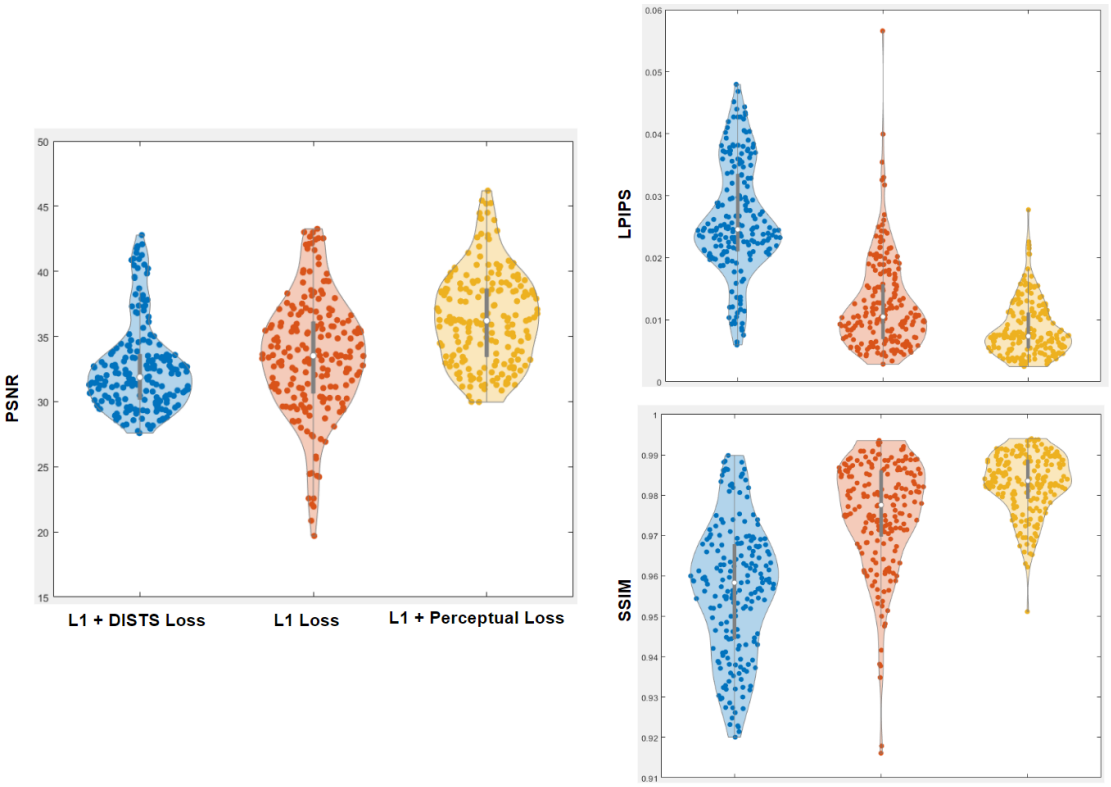

Figure A.3: *Violin plots depicting the performance of the proposed model trained with L1, L1+DISTS and L1+Perceptual loss functions on the 200 test images with white artifacts with respect to PSNR, SSIM and LPIPS Image Quality Assessment (IQA) matrices. X-axis indicates to the models' name and Y-axis refers to the IQA matrices*

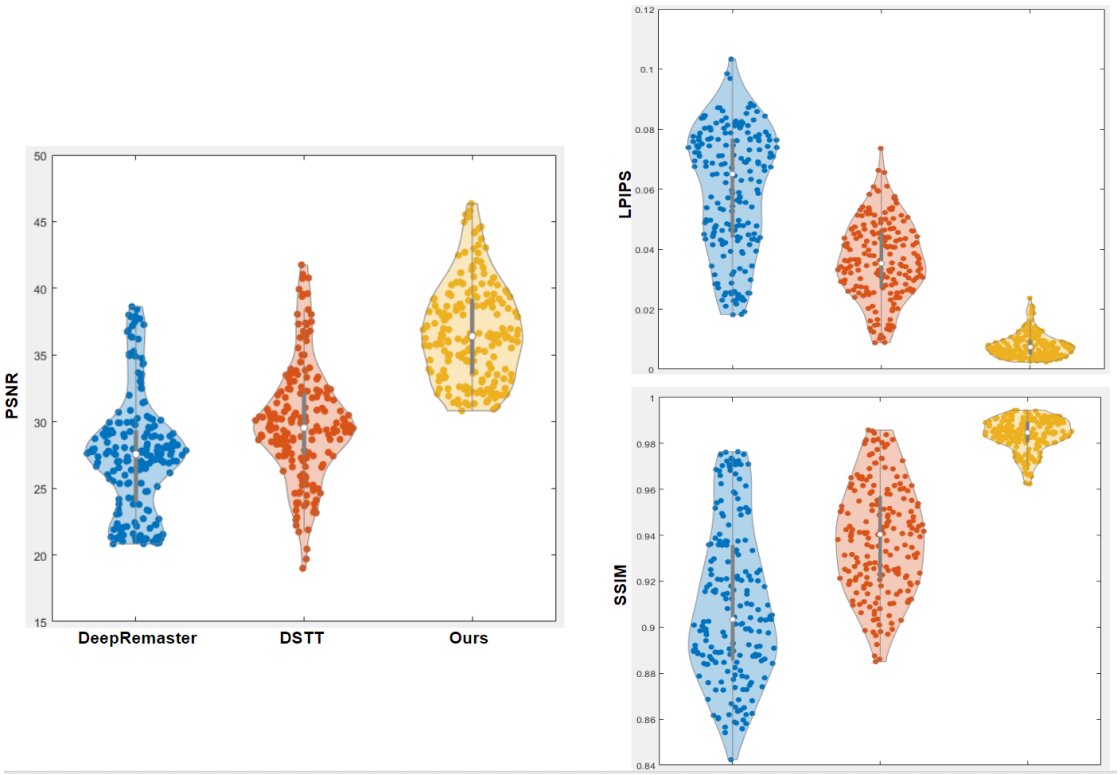

Figure A.4: *Violin plots depicting the performance of the DeepRemaster (23), DSTT (35) and proposed DSTT+MARB model on the 200 test images with white artifacts with respect to PSNR, SSIM and LPIPS Image Quality Assessment (IQA) matrices. X-axis indicates to the models' name and Y-axis refers to the IQA matrices*

Bibliography

[1] Adobe photoshop. `https://www.adobe.com/no/`. Accessed: 2022-07-24. (cited on page 8)

[2] Amazon web services. `https://aws.amazon.com/`. Accessed: 2022-07-24. (cited on page 61)

[3] Brisque image quality assessment. `https://learnopencv.com/image-quality-assessment-brisque/`. Accessed: 2022-07-24. (cited on pages 59 and 93)

[4] Den-forsvundne-polsemaker. Mercury Film A/S. (cited on pages 2 and 91)

[5] Youtube. `https://www.youtube.com/`. Accessed: 2022-07-24. (cited on page 37)

[6] Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., and Vijayanarasimhan, S. (2016). Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*. (cited on page 36)

[7] Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495. (cited on pages 26 and 92)

[8] Chambah, M. (2008). Reference-free image quality evaluation for digital film restoration. *Colour: Design & Creativity*, 4(3):1–16. (cited on page 7)

[9] Chang, Y.-L., Liu, Z. Y., Lee, K.-Y., and Hsu, W. (2019). Free-form video inpainting with 3d gated convolution and temporal patchgan. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9066–9075. (cited on page 30)

[10] Chen, X., Huang, Y., and Xu, L. (2020). Multi-scale attentive residual dense network for single image rain removal. In *Proceedings of the Asian Conference on Computer Vision*. (cited on pages 26, 27, and 92)

[11] Criminisi, A., Pérez, P., and Toyama, K. (2004). Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on image processing*, 13(9):1200–1212. (cited on page 4)

[12] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee. (cited on page 50)

[13] Ding, K., Ma, K., Wang, S., and Simoncelli, E. P. (2020). Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*. (cited on page 50)

[14] Dixon, W. W. (2009). Treasures iv: American avant-garde film, 1947–1986. national film preservation foundation/image entertainment, 2009. *Quarterly Review of Film and Video*, 26(3):263–264. (cited on page 7)

[15] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. (cited on pages 28 and 29)

[16] Du, Y., Xu, J., Qiu, Q., Zhen, X., and Zhang, L. (2020). Variational image deraining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2406–2415. (cited on page 27)

[17] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27. (cited on page 24)

[18] Hintze, J. L. and Nelson, R. D. (1998). Violin plots: a box plot-density trace synergism. *The American Statistician*, 52(2):181–184. (cited on page 65)

[19] Hore, A. and Ziou, D. (2010). Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE. (cited on page 56)

[20] Hossam, M., Le, T., Papasimeon, M., Huynh, V., and Phung, D. (2021). Text generation with deep variational gan. *arXiv preprint arXiv:2104.13488*. (cited on page 24)

[21] Hu, J., Shen, L., and Sun, G. (2018a). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141. (cited on page 46)

[22] Hu, Y., He, H., Xu, C., Wang, B., and Lin, S. (2018b). Exposure: A white-box photo post-processing framework. *ACM Transactions on Graphics (TOG)*, 37(2):1–17. (cited on page 14)

[23] Iizuka, S. and Simo-Serra, E. (2019). Deepremaster: temporal source-reference attention networks for comprehensive video enhancement. *ACM Transactions on Graphics (TOG)*, 38(6):1–13. (cited on pages 18, 19, 37, 43, 68, 70, 75, 76, 78, 82, 91, 93, 94, 95, and 97)

[24] Iizuka, S., Simo-Serra, E., and Ishikawa, H. (2017). Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14. (cited on pages 15 and 91)

[25] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134. (cited on page 14)

[26] Johnson, J., Alahi, A., and Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer. (cited on pages 49 and 93)

[27] Keleş, O., Yılmaz, M. A., Tekalp, A. M., Korkmaz, C., and Doğan, Z. (2021). On the computation of psnr for a set of images or video. In *2021 Picture Coding Symposium (PCS)*, pages 1–5. IEEE. (cited on page 57)

[28] Kim, C., Kim, S., and Paik, J. (2010). Digital image processing techniques for restoring old damaged films and their applications to korean film restoration. In *SMPTE Annual Tech Conference & Expo, 2010*, pages 2–11. SMPTE. (cited on pages 17, 18, and 91)

[29] Kim, D., Woo, S., Lee, J.-Y., and Kweon, I. S. (2019). Deep video inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5792–5801. (cited on page 3)

[30] Kim, K.-t. and Kim, E. Y. (2010). Film line scratch detection using texture and shape information. *Pattern recognition letters*, 31(3):250–258. (cited on page 7)

[31] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. (cited on pages 53 and 61)

[32] Kuiper, A. and Sigmund, M. (2005). Simulating of authentic movie faults. In *EUROCON 2005-The International Conference on" Computer as a Tool"*, volume 2, pages 1015–1018. IEEE. (cited on page 7)

[33] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690. (cited on page 14)

[34] Liu, G., Reda, F. A., Shih, K. J., Wang, T.-C., Tao, A., and Catanzaro, B. (2018). Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European conference on computer vision (ECCV)*, pages 85–100. (cited on pages 15 and 91)

[35] Liu, R., Deng, H., Huang, Y., Shi, X., Lu, L., Sun, W., Wang, X., Dai, J., and Li, H. (2021a). Decoupled spatial-temporal transformer for video inpainting. *arXiv preprint arXiv:2104.06637*. (cited on pages 16, 28, 29, 51, 60, 63, 64, 68, 69, 70, 75, 76, 78, 82, 92, 93, 94, 95, and 97)

[36] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021b). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022. (cited on page 21)

[37] Mafi, M., Martin, H., Cabrerizo, M., Andrian, J., Barreto, A., and Adjouadi, M. (2019). A comprehensive survey on impulse and gaussian denoising filters for digital images. *Signal Processing*, 157:236–260. (cited on page 3)

[38] Mao, X.-J., Shen, C., and Yang, Y.-B. (2016). Image restoration using convolutional auto-encoders with symmetric skip connections. *arXiv preprint arXiv:1606.08921*. (cited on page 3)

[39] Maurer, R. (2007). Comprehensive solutions for removal of dust and scratches from images. (cited on pages 11, 12, 13, and 91)

[40] Mironică, I. (2020). A generative adversarial approach with residual learning for dust and scratches artifacts removal. In *Proceedings of the 2nd Workshop on Structuring and Understanding of Multimedia heritAge Contents*, pages 15–22. (cited on page 14)

[41] Mittal, A., Moorthy, A. K., and Bovik, A. C. (2012). No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708. (cited on page 59)

[42] Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*. (cited on page 30)

[43] Nah, S., Hyun Kim, T., and Mu Lee, K. (2017). Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3883–3891. (cited on page 3)

[44] Ohuchi, T., Seto, T., Komatsu, T., and Saito, T. (2000). A robust method of image flicker correction for heavily-corrupted old film sequences. In *Proceedings 2000 International Conference on Image Processing (Cat. No. 00CH37101)*, volume 2, pages 672–675. IEEE. (cited on page 8)

[45] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32. (cited on page 61)

[46] Perez, L. and Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*. (cited on page 41)

[47] Raijada, M., Patel, D., and Prajapati, P. (2015). A review paper on image quality assessment metrics. *Journal of Emerging Technologies and Innovative Research (JETIR)*, 2:130–132. (cited on page 56)

[48] Robins, D. and Ye, J. (2003). Method and apparatus for detection and removal of scanned image scratches and dust. US Patent App. 09/939,094. (cited on page 11)

[49] Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*. (cited on page 54)

[50] Saito, T., Komatsu, T., Ohuchi, T., and Seto, T. (2000). Image processing for restoration of heavily-corrupted old film sequences. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 3, pages 13–16. IEEE. (cited on page 18)

[51] Salem PhD, N. M. F. (2021). A survey on various image inpainting techniques. *Future Engineering Journal*, 2(2):1. (cited on page 4)

[52] Schallauer, P., Pinz, A., and Haas, W. (1999). Automatic restoration algorithms for 35mm film. *J. Computer Vision Res*, 1(3):59–85. (cited on pages 7 and 8)

[53] Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. (cited on page 50)

[54] Suganuma, M., Liu, X., and Okatani, T. (2019). Attention-based adaptive selection of operations for image restoration in the presence of unknown combined distortions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9039–9048. (cited on page 22)

[55] Tan, M. and Le, Q. (2021). Efficientnetv2: Smaller models and faster training. In *International Conference on Machine Learning*, pages 10096–10106. PMLR. (cited on pages 48 and 93)

[56] Tian, C., Fei, L., Zheng, W., Xu, Y., Zuo, W., and Lin, C.-W. (2020). Deep learning on image denoising: An overview. *Neural Networks*, 131:251–275. (cited on page 3)

[57] Tomasi, C. and Manduchi, R. (1998). Bilateral filtering for gray and color images. In *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*, pages 839–846. IEEE. (cited on page 11)

[58] van Roosmalen, P. M., Lagendijk, R. L., and Biemond, J. (1999). Correction of intensity flicker in old film sequences. *IEEE Transactions on Circuits and Systems for Video Technology*, 9(7):1013–1019. (cited on page 8)

[59] Wan, Z., Zhang, B., Chen, D., and Liao, J. (2022). Bringing old films back to life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17694–17703. (cited on pages 21, 22, 77, and 91)

[60] Wan, Z., Zhang, B., Chen, D., Zhang, P., Chen, D., Liao, J., and Wen, F. (2020). Bringing old photos back to life. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2747–2757. (cited on pages 1, 20, 21, and 91)

[61] Wang, X., Xie, L., Dong, C., and Shan, Y. (2021). Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1905–1914. (cited on page 1)

[62] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612. (cited on pages 57, 58, and 93)

[63] Werlberger, M., Pock, T., Unger, M., and Bischof, H. (2011). Optical flow guided tv-l 1 video interpolation and restoration. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 273–286. Springer. (cited on page 3)

[64] Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19. (cited on page 46)

[65] Wu, Y., Chrysos, G. G., and Cevher, V. (2022). Adversarial audio synthesis with complex-valued polynomial networks. *arXiv preprint arXiv:2206.06811*. (cited on page 24)

[66] Xu, B., Wang, N., Chen, T., and Li, M. (2015). Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*. (cited on page 31)

[67] Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., and Huang, T. (2018). Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*. (cited on pages 35, 36, and 92)

[68] Xu, R., Li, X., Zhou, B., and Loy, C. C. (2019). Deep flow-guided video inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3723–3732. (cited on page 3)

[69] Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., and Huang, T. S. (2018). Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514. (cited on pages 15, 30, and 91)

[70] Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., and Huang, T. S. (2019). Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4471–4480. (cited on page 30)

[71] Yu, T. and Zhu, H. (2020). Hyper-parameter optimization: A review of algorithms and applications. *arXiv preprint arXiv:2003.05689*. (cited on page 53)

[72] Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., Yang, M.-H., and Shao, L. (2020). Learning enriched features for real image restoration and enhancement. In *European Conference on Computer Vision*, pages 492–511. Springer. (cited on pages 46, 47, and 93)

[73] Zhang, K., Liang, J., Van Gool, L., and Timofte, R. (2021a). Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4791–4800. (cited on page 1)

[74] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings*

of the IEEE conference on computer vision and pattern recognition, pages 586–595. (cited on page 58)

[75] Zhang, R., Wu, J., Ding, Y., and Hao, P. (2009). The correction of intensity flicker in archived film. In *2009 International Conference on Information Technology and Computer Science*, volume 2, pages 402–405. IEEE. (cited on page 8)

[76] Zhang, Y., Liu, T., Cattani, C., Cui, Q., and Liu, S. (2021b). Diffusion-based image inpainting forensics via weighted least squares filtering enhancement. *Multimedia Tools and Applications*, 80(20):30725–30739. (cited on page 4)

List of Figures

LIST OF FIGURES

List of Tables