Perceptual Loss based Approach for Analogue Film Restoration

Daniela Ivanova, Jan Paul Siebert and John Williamson
School of Computing Science, University of Glasgow, Glasgow, U.K.

Keywords: Image Restoration, Perceptual Loss, Restoration Evaluation.

Abstract: Analogue film restoration, both for still photographs and motion picture emulsions, is a slow and laborious manual process. Artifacts such as dust and scratches are random in shape, size, and location; additionally, the overall degree of damage varies between different frames. We address this less popular case of image restoration by training a U-Net model with a modified perceptual loss function. Along with the novel perceptual loss function used for training, we propose a more rigorous quantitative model evaluation approach which measures the overall degree of improvement in perceptual quality over our test set.

1 INTRODUCTION

Photographic film emulsion, because of its physical nature, is prone to degradation due to improper storage and handling or simply over time (Chambah, 2019). One way to ensure the longevity and wider availability of images (and movies) captured on film is to digitise them through scanning. Scanning film often causes random analogue artifacts such as dust and scratches of different shape, size and colour, to also be transferred to the digital domain. Dust and scratch artifacts can occlude a varying degree of the content of the image and decrease its overall perceptual quality.

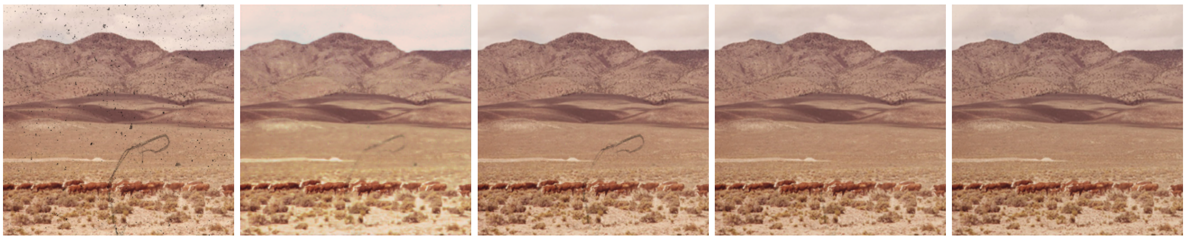
Image restoration in this context refers to the identification of such artifacts in film image scans and the subsequent in-painting of the affected area. While traditional image processing tools that aim to automate the task such as Kodak's Digital ICE (Image Correction and Enhancement) do exist, they can only be applied to a limited number of colour film emulsion types and introduce a significant additional cost through requiring specialised hardware (Fielding, 2008). Furthermore, such an approach can only go so far as to detect the artifacts; the identified areas would still need to be digitally in-painted.

While convolutional neural network approaches have been utilised for various image restoration tasks, such as super-resolution, JPEG artifact removal, deraining and denoising with great success, we hypothesise that *for such approaches to successfully be applied to film restoration, meaningful differentiation between artifacts and useful high frequency image features has to be learned by the network during training*. In addition, upon evaluation, it is crucial to quantify the loss of information introduced by the network, as an ideal network will not only have to remove the artifacts present in the input, but also learn the identity function for inputs where there are no artifacts.

Informed by the above insight, our main contribution is a perceptual loss function better suited to the image statistics of dust and scratches. We utilise the shallower layers of a pre-trained feature extraction network, and include an additional loss term based on the Structural Similarity Index (SSIM) perceptual quality metric. We demonstrate that the restoration network trained with our novel perceptual loss formulation improves the achieved perceptual quality of restored images. Additionally, we show that the network targets analogue artifacts specifically and the loss of useful information (such as fine detail for colour shifts) is decreased. We also describe a more rigorous way to quantitatively evaluate restoration quality, taking into account whether the restoration network introduces new degradation if non-damaged images are passed as input. As we found that data sets which could be used for benchmarking model performance on this specific task are unavailable, our final contribution is a data set of clean-damaged pairs, which we produced by applying synthetic artifact damage to "clean" image scans; we used this data set both for training and evaluation.

The paper is organised as follows: in Section 2, we review comparable restoration tasks, as well as state-of-the-art deep learning approaches for analogue ar-

Ivanova, D., Siebert, J. and Williamson, J.
Perceptual Loss based Approach for Analogue Film Restoration.
DOI: 10.5220/0010829300003124
In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022) - Volume 4: VISAPP*, pages 126-135
ISBN: 978-989-758-555-5; ISSN: 2184-4321



(a) Input.

(b) Perceptual loss prediction.

(c) Modified perceptual loss prediction (ours).

(d) Modified perceptual loss with SSIM term prediction (ours).

(e) Target.

Figure 1: Input (a) has a large scratch artefact and several small dust artefacts. The model which was trained with perceptual loss (b) blurred the image, introduced a colour shift and lost a lot of fine detail, without removing the large scratch. The model which was trained with our modified perceptual loss (c) has preserved detail but has not removed the scratch. The final model, trained with our combined modified perceptual loss with an SSIM loss term (d), has successfully removed most of the large scratch, while preserving detail and without introducing additional loss of information.

tifact restoration, and consider the difficulty of evaluating a restoration task. In Section 3, we detail our approach, including the architecture used, data pre-processing, training and evaluation methodologies and perceptual loss formulations. Finally, in Section 4, we discuss the results of our experiments, including qualitative and quantitative comparisons with both existing perceptual losses used for colourisation and super-resolution and a state-of-the-art old photo restoration approach.

2 RELATED WORK

Some of the most exciting state-of-the-art deep learning approaches have captured the attention of both researchers and the general public through demonstrating compelling results in the task of colourisation of old black-and-white photographs (Zhang et al., 2016; Zhang et al., 2017; Antic, 2020). These approaches leverage a learned prior over low frequency, global image context features to generate missing colour information. The successful application of similar Convolutional Neural Network techniques has also been demonstrated in low level image restoration tasks, such as in-painting (Mao et al., 2016; Ulyanov et al., 2018), denoising (Mao et al., 2016; Ulyanov et al., 2018), deraining (Meng Tang et al., 2018; Fan et al., 2018), superresolution (Mao et al., 2016; Ulyanov et al., 2018; Ledig et al., 2017). Virtually all such state-of-the-art approaches, with the exception of Deep Image Prior, rely on training the networks on large datasets of ”natural” images, such as ImageNet.

However, research on film artifact removal is scarce. In the literature, Strubel et al. proposed a SegNet encoder-decoder architecture which is trained with a cross entropy loss to remove dust and scratches for a purpose-built data set of black and white image pairs of dusty scans and their matching manually repaired versions (Strubel et al., 2019). Mironica et al. presents an approach based on generative adversarial training, in which the generator is trained to restore artifacts using a perceptual loss function inspired by style transfer approaches (Johnson et al., 2016) as the reconstruction loss (Mironică, 2020). Perceptual loss in a GAN setting has also been applied to colourisation in the DeOldify project (Antic, 2020). An alternative approach based on Variational Autoencoders (VAEs) is proposed by Wan et al. The authors train two VAEs on the domains of damaged and restored images respectively, and use their learned latent spaces as an in-between domain for the translation from a damaged to a restored image. The networks are trained using synthetic paired data (Wan et al., 2020). To our knowledge this is the state-of-the-art approach for old photographs restoration - a task which is the most closely aligned with film scan artifact restoration.

Automated digital image restoration quality evaluation is another understudied topic. It is a difficult one as many metrics depend on the existence of ground truth restored images to compare to, and in real life restoration scenarios those are not readily available. Other metrics which utilise models of the human visual system to measure perceived quality may miss distortions which are below the threshold of visibility (Chambah, 2019). Finally, the quality of restoration is dependent on objective properties such as the size and detectability of artifacts with respect to image resolution, but also to the subjective definition and scope of what makes a perceptually good restoration: should we only in-paint artifacts, or additionally correct colour shifts, should we remove or preserve grain, etc. In any case it is clear that a good restoration approach should minimise additional loss of information. Neural networks can introduce checkerboard artifacts (Aitken et al., 2017), GAN-artifacts (Zhang et al., 2019), while VAEs producing blurry outputs is an empirically observed and extensively addressed problem (Bousquet et al., 2017). All of the above restoration methods mentioned in the literature use standard denoising evaluation metrics, such as Peak signal-to-noise ratio (PSNR) and Sructural similarity index measure (SSIM) to quantitatively evaluate the quality of restoration. While the results shown in these works are impressive, there is a lack of quantitative evaluation in terms of novel damage or loss of information introduced by the restoration networks. Our experiments with passing non-damaged images through the restoration networks and measuring the perceptual quality of the outputs aims to address this gap.

3 APPROACH

3.1 Restoration Network Architecture

A restoration network's aim is to translate the input corrupted image to a ”restored” version of it; a state-of-the-art architecture used for this type of task is U-net with skip connections. U-net falls within the encoder-decoder family of network architectures, where the encoder is responsible for downsampling the input to a compact feature-vector form, which is then used by the second part, the decoder, to translate it back to the image domain. In our case, the target domain is the domain of ”clean”, natural images with no dust and scratches present. The skip-connections between each up- and down-sampling stages in the U-Net present an advantage over simpler encoder-decoder architectures, effectively aiding the network in preserving intermediate multi-scale representations of the input created during the down-sampling phase, which are then used during up-sampling to better model high-frequency features at larger resolutions. This is empirically demonstrated by U-Net's initial success in medical segmentation tasks that require a highly precise image (segmentation mask) output.

We leverage the U-net's morphological separation of clearly defined encoder and decoder sections by using pre-trained weights for the ResNet34 architecture provided by PyTorch as the encoder part. Two additional convolutional layers are added as the bottleneck of the ”U” shape in order to transition to the decoder part of the overall network. The decoder is tasked with utilising the encoded natural features to separate out the artifacts and generate image data of high perceptual quality. Inspired by the DeOldify implementation (Antic, 2020), that can be achieved by using sub-pixel convolution for the upsampling blocks that make up the decoder (Shi et al., 2016a; Shi et al., 2016b). To minimise checkerboard artefacts that can be introduced during the upsampling process, self-attention (Zhang et al., 2018) is added to the second upsampling block, counting from the bottleneck. The input image itself is also concatenated with a dense cross-connection to the input of the last upsampling block in the decoder, to further address the problem of preserving fine detail. Finally, sigmoid range activation is applied to produce a 3-channel RGB image with the same spatial size as the input.

3.2 Dustified Data Set

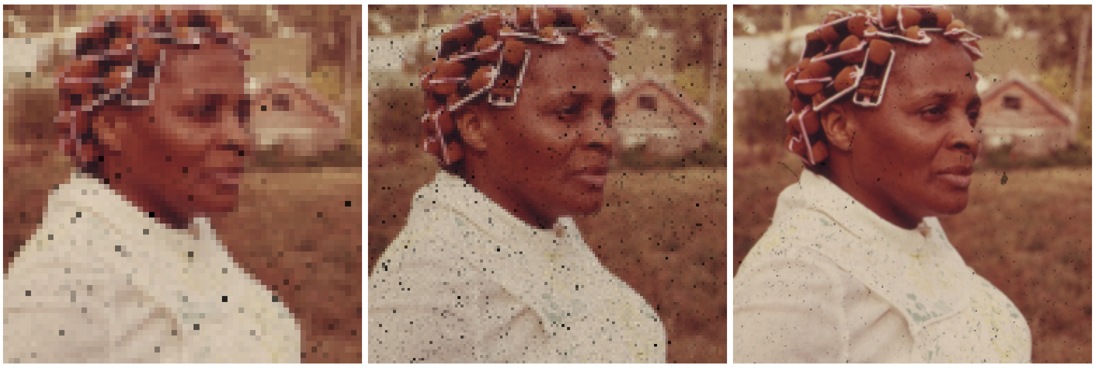
For our data set, we used digitised versions of Kodachrome slides from the *Documerica*[1] series made available by The US National Archives through Flickr. Overall, 6232 *Documerica* images were collected using the Flickr API. The first pre-processing step is to resize and center-crop the raw images. Resizing is carried out by using inter area interpolation, so that there is minimal image quality loss. The complete data set is available at https://archive.org/details/documerica.

To generate damaged versions of the collected images, a randomly selected patches from a set of dust and scratch overlays[2] were applied to each of the 6232 ”raw” images. Since the images we use are slide scans, the colour of the artifacts we simulate is black. While the random nature of artefacts like dust and scratches is impossible to perfectly replicate in a deterministic way, the following simple approach was devised to generate a unique scratch overlay for every clean image: randomly select a dust overlay, crop a random square patch from it, warp and invert it, then apply to the image. For each image, this was performed twice, so that the dust pattern on each image is a different random combination of two dust overlays. This step was performed for each corresponding target size, i.e. 64 by 64, 128 by 128 and 256 by 256 pixels, resulting in different patterns for the three different sizes of the same clean image. An example is provided in Figure 2.

We split the data set using a 8:1:1 training-validation-test ratio, resulting in 4895 pairs in the training set, 623 in the validation set, and 624 in the test set.

[1] https://www.flickr.com/photos/usnationalarchives/collections/72157620729903309/

[2] https://blog.spoongraphics.co.uk/freebies/30-free-film-dust-textures-add-dirty-effects-work



(a) 64x64 pixels.

(b) 128x128 pixels.

(c) 256x256 pixels.

Figure 2: Three dustified versions of the same image from the data set at three different resolutions. Notice that the damage ”pattern” is different for each size as to help the network generalise better.

3.3 Proposed Perceptual Loss Function

Instead of explicitly comparing two images in the pixel domain, another network can be used as a comparison tool. That is, we aim at comparing the generated output with the target via a pre-trained network's feature space to allow for a more comprehensive expression of the difference. Perceptual loss as used for style transfer is defined by two terms, representing the style and content of the generated and the target image through the feature activations of the pre-trained loss network's hidden layers. The content feature loss is the sum of the element-wise difference between the feature maps extracted from each of the ReLU layers of a pre-trained VGG16 network (Johnson et al., 2016). The authors chose Euclidean distance in the original paper, whereas DeOldify uses Manhattan distance (Antic, 2020) when adapting the perceptual loss approach to a colour restoration task. The style loss is calculated in a similar way, however, the feature maps need to be transformed to a spatially-invariant form first - the idea of style loss is to measure the difference between the *distributions* of the feature map activations, as well as the correlation between features within each feature map that is produced (Johnson et al., 2016). To find the correlation between features within a feature mapping, the Gram matrix of the feature map is calculated; the Gram matrix is the dot product between each pair of flattened feature vectors in the feature map. Therefore, the Gram matrix measures which features tend to activate together.

The terms corresponding to feature activations extracted from each layer which comprise the style and content loss sums can also be weighted. The weights used in the DeOldify project are 0, 0, 20, 70, 10 for each respective ReLU layer in the VGG16 network (Antic, 2020); the activations extracted from the first two ReLUs are ignored. We propose that these layers are indeed relevant for the task of detecting small scale analogue artifacts as well as preserving high frequency image features, and conduct a small preliminary experiment, based on which we revise the weights for both the style and content loss sum terms.

In addition we propose a term representing another measure of perceptual quality, based on the SSIM index (Wang et al., 2004), which we call self-similarity loss. As we seek to maximise the SSIM index during restoration, we have defined the self-similarity loss as:

$$D_{SSIM}(x,y) = 1 - SSIM(x,y), \tag{1}$$

where x and y can be images or feature maps.

Since the feature maps produced by each layer of the feature extractor do hold spatially relevant information, we decide to employ the self-similarity loss as a distance measure between feature representations for each layer in the content loss term in our final loss formulation. We trained the restoration network with three different perceptual loss functions, all based on the activations extracted from a pre-trained VGG16:

Model 1: We assigned the layer weights as used in the DeOldify project, i.e. 0, 0, 20, 70, 10 (Antic, 2020) corresponding to layers 5, 12, 22, 32, 42 from VGG16 (ReLUs). We also used the same distance function as DeOldify, i.e. mean absolute error (MAE).

Model 2: Based on our proposition that earlier layer activations are relevant to the analogue artifact restoration task, we assign the following weights each of the five ReLUs: 2, 4, 5, 6, 6. Again, we use MAE as the distance measure for both style loss and content loss.

Model 3: We train with the weights of Model 2, and also add our own distance metric - SSIM loss (see Equation 1) with a window size of 5 by 5 pixels. We use the SSIM loss as a distance measure between the prediction and target images, as well as the distance function in the content loss term; for the style loss term, we use MAE, and we also measure the MAE distance between the target and the prediction.

3.4 Training Procedure

For each of the three perceptual loss configurations in our experiments, a network with our proposed architecture is trained progressively over three different input sizes for 10 epochs. The training and validation data is loaded as matched image pairs of 3-channel images from the clean and dustified classes at each respective resolution. The data is normalized and scaled using the ImageNet data set's overall mean and standard deviation values (per-channel) (Deng et al., 2009). We applied the following data augmentation on the training set, with specific probabilities of applying the transformation listed: random horizontal flip, probability 50%; symmetric warp with magnitude between (-0.25, 0.25), probability 75%; random zoom up to x1.5, probability 75%; brightness variation between (0.25, 0.75), probability 75%; contrast scaling between (0.5, 2.0), probability 75%.

During training, the ResNet34 encoder is frozen - therefore the feature space learned from ImageNet is preserved and used to encode the damaged input. When decoder layers are initially appended to a pre-trained backbone, their weights are randomly initialised. We use One cycle training policy (Smith and Topin, 2019) to drastically reduce training times: the learning rate is increased for 80% of the iterations where the encoder is frozen. An additional training phase at the highest resolution is performed for another 20 epochs, by unfreezing the backbone and fine-tuning the encoder, again using One cycle training policy, this time increasing the learning rates only 50% of the time. The loss landscape is optimised via the Adam optimiser, with the PyTorch default betas and weight decay set to $1e-3$ (Kingma and Ba, 2014). Python notebooks containing the training, evaluation and links to the final trained models' weights are available at https://github.com/daniela997/DustScratchRemoval.

4 EXPERIMENTS

4.1 Experimental Methodology

For our experiments, we calculate the SSIM scores for each clean-dustified pair, in the training set and plot the distribution of the result. The distribution of the obtained values is displayed in Figure 3. We also show that the SSIM between every clean image and itself is equal to 1. Therefore, a restoration network's task is to squish the left histogram so that it resembles the one on the right as much as possible, for both damaged and clean input.

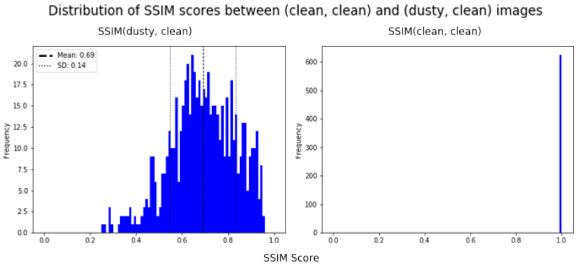


Figure 3: Distribution of SSIM scores for items in the test set. The SSIM score is obtained for each image pair in the test set.

The damage present in the test set is varied, as evidenced by the wide range of SSIM scores obtained by the set of clean-dustified pairs: lowest score is 0.2 SSIM (i.e. a large degree of damage is present), while the highest score is 0.97 SSIM (i.e. very little damage is present). As all clean-dustfied pairs produced SSIM scores of less than 1, we can confirm that all damaged images in each pair have some degree of damage.

We hypothesise that a well-trained restoration network should be able to map multiple versions of an image with varying degrees of damage to the same restored version. Furthermore, we observe that if an already non-damaged image is passed through the restoration network, the SSIM score between the input and the output should be 1, i.e. a perfect restoration network should learn the identity transform and not change the input, since no damage is present. For this, we designed an experiment inspired by the above observation. We, therefore, pass *clean* images through the restoration network, and measure how much damage is introduced by the network. Additionally, we seek to quantify the degree of *improvement* seen across the test set when predictions are made on dustified images. We calculate the difference between the SSIM score of each clean-dustified pair and the SSIM score of each corresponding clean-restored pair to produce the $\delta SSIM$ measure.

4.2 Experiment on Layer Response to the Presence of Dust and Scratches

The activations extracted from the first two ReLUs of the VGG16 loss network are ignored (their corresponding weights are set to 0) in the DeOldify project implementation for perceptual loss (Antic, 2020). This could be explained with the fact that low-level features learned by earlier layers are not relevant to the task of colourisation, since to colourise something

Table 1: Number of different filters within the top 9 filter sets for the clean and dustified version of each image. If all filters in both top 9 sets are the same (irrespective of ranking), the result would be 0 different filters, and on the contrary, if they are all different, the result would be 18.

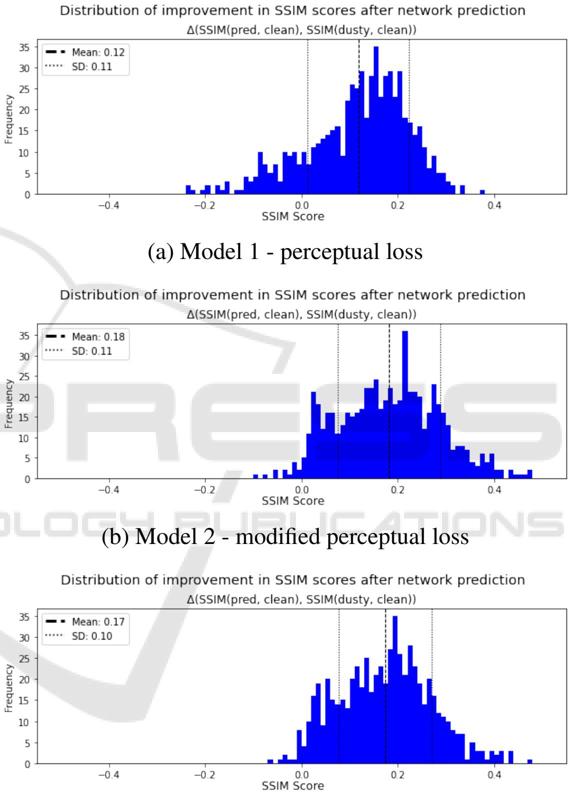
	Image 1	Image 2	Image 3
ReLU 5	0	4	0
ReLU 12	10	10	6
ReLU 22	6	8	2
ReLU 32	14	16	6
ReLU 42	14	12	12

in a correct way, a higher-level, semantic knowledge is required, and that is learned by the deeper layers of the VGG16 network. On the other hand, we hypothesise that the activations for shallower layers could be relevant to the task of artefact removal and should not be discarded. At the same time, high-level semantic knowledge from the learned image prior of the deeper layers is still important in distinguishing dust and scratches from contextually meaningful high frequency features with similar statistics, such as edges, fine lines, strands of a person's hair.

For this preliminary experiment, we selected three image pairs from our data set with varying degree of damage, which we pass through the pre-trained feature extractor loss network - VGG16. For each photo, we extract the activations of the ReLU layers of the VGG16 network which are used in the perceptual loss formulation - i.e. layers 5, 12, 22, 32, 42. Each of those will have 64, 128, 256, 512 and 512 feature maps (filters), as per the VGG16 architecture definition. The mean activations for each filter and the 9 most activated filters are recorded and shown in Table 1. These results, while limited, demonstrate that the activations from earlier layers in the feature extractor network are indeed responsive to the presence of dust and scratches, and therefore relevant to the perceptual loss formulation.

Table 2: Summary of mean and standard deviation values for the SSIM scores distribution obtained by each perceptual loss model on the test set of 624 image pairs.

	SSIM of Predictions on dustified images		SSIM of Predictions on clean images	
	Mean	SD	Mean	SD
MSE Model	0.84	0.07	0.95	0.04
Model 1	0.80	0.08	0.77	0.08
Model 2	0.88	0.05	0.94	0.04
Model 3	0.88	0.05	0.98	0.02



(a) Model 1 - perceptual loss

(b) Model 2 - modified perceptual loss

(c) Model 3 - modified perceptual loss with SSIM term

Figure 4: Distribution of improvement ($\delta SSIM$) for predictions on the test set for the three perceptual loss models. These histograms show us by how much images with artifacts present increased (or decreased) their SSIM scores after being restored by each network.

4.3 Quantitative Comparison of Perceptual Losses

We assess the ability to restore artefacts, and measure what degree of new damage is introduced to clean images, for each model, through calculating the SSIM score between the model's prediction and the target clean image. We compare the obtained distributions to the SSIM score distributions of the clean-dustified pairs in the test set (see Figure 3), as well as a baseline model trained and fine-tuned using per-pixel Mean Squared Error loss. Results are summarised in Table 2 and compared to the model trained using MSE pixel loss as a baseline. A mean of 1.0 and standard deviation of 0.0 is ideal for both predictions on clean and dustified images - it would mean that the model has perfectly repaired the dustified images, and has perfectly preserved the clean images.

Model 1 obtained worse scores than the model trained using simple pixel MSE as loss function. It obtained an especially low mean value for the SSIM scores of predictions made on clean images. This suggests that the network has a destructive effect on non-artefact high frequency features, as only activations from deeper VGG16 layers are included in the loss function. Model 2 has a higher mean value than both the baseline and Model 1 for SSIM scores on predictions over dustified images, as well as smaller spread of the overall distribution. The mean SSIM scores for predictions on clean images is comparable to that of the baseline and much better than those obtained by Model 1, which demonstrates that the inclusion of the first two ReLU layers in the loss function aided Model 2 in learning to preserve non-artefact image information much better than Model 1. Model 3 includes the addition of SSIM loss term between prediction and target, as well as as a distance measure for comparing feature activations in the content loss term of the feature loss. The mean and standard deviation for the SSIM scores are the same as Model 2 - however, this final model obtained the highest mean SSIM score on predictions over clean test images, 0.98, along with smallest standard deviation, 0.02, demonstrating that it was by far the least intrusive out of all networks we trained.

We use our $\delta SSIM$ measure to quantify the degree of ”improvement” the network introduced in predictions on dustified images, shown in Figure 4. Model 3 achieved the smallest spread of $\delta SSIM$. Both Model 2 and 3 and produced a smaller number of predictions which obtained lower SSIM after they were restored by the network (i.e. they have $\delta SSIM$ below 0).

To summarise, we found that the inclusion of activations from the earlier ReLU layers from the VGG16 network helped preserve detail, minimise colour shift and improve the quality of in-painting. Additionally, the introduction of SSIM loss as part of the the perceptual loss to compare both predictions and target, and the feature activations from each ReLU layer in the content loss, allowed the the network to generalise over different artefact sizes and shapes, and made it minimally invasive to areas with no damage.

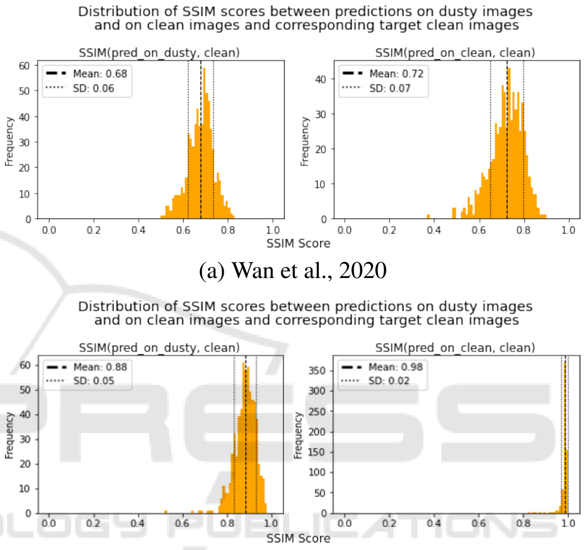
4.4 Comparison with State-of-the-Art Restoration Approaches

We also compare our approach against an alternative deep learning image restoration approach which involves training two Variational Autoencoders on damaged and restored photos respectively, and using their learned latent spaces to transform an image from one domain to the other (Wan et al., 2020). The method based on Deep Latent Space Translation via VAEs frames the restoration problem not only as dust and scratch removal, but also include other types of restoration such as colour correction and smoothing. Still, the authors provide pre-trained model weights which specifically target scratches; we therefore used these model weights in our experiment. A visual comparison of the achieved restorations is shown in Figure 6. We can see that the VAE-based model has smoothed the image, removed small dust specks and shifted the overall tint of the image towards green. The latter is expected as Wan et al. (Wan et al., 2020) trained their model on sepia-toned and discoloured examples as well, in the case of which restoration is understood as colour correction.

We quantitatively evaluated the approach of Wan et al. against ours using the methodology based on the SSIM metric described in Section 3.5. A summary of the achieved scores accross the test set is provided in Table 3 and Figure 5.

Table 3: Summary of mean and standard deviation values for the SSIM scores distribution obtained by a state-of-the-art restoration approach and by our approach on the test set of 624 image pairs.

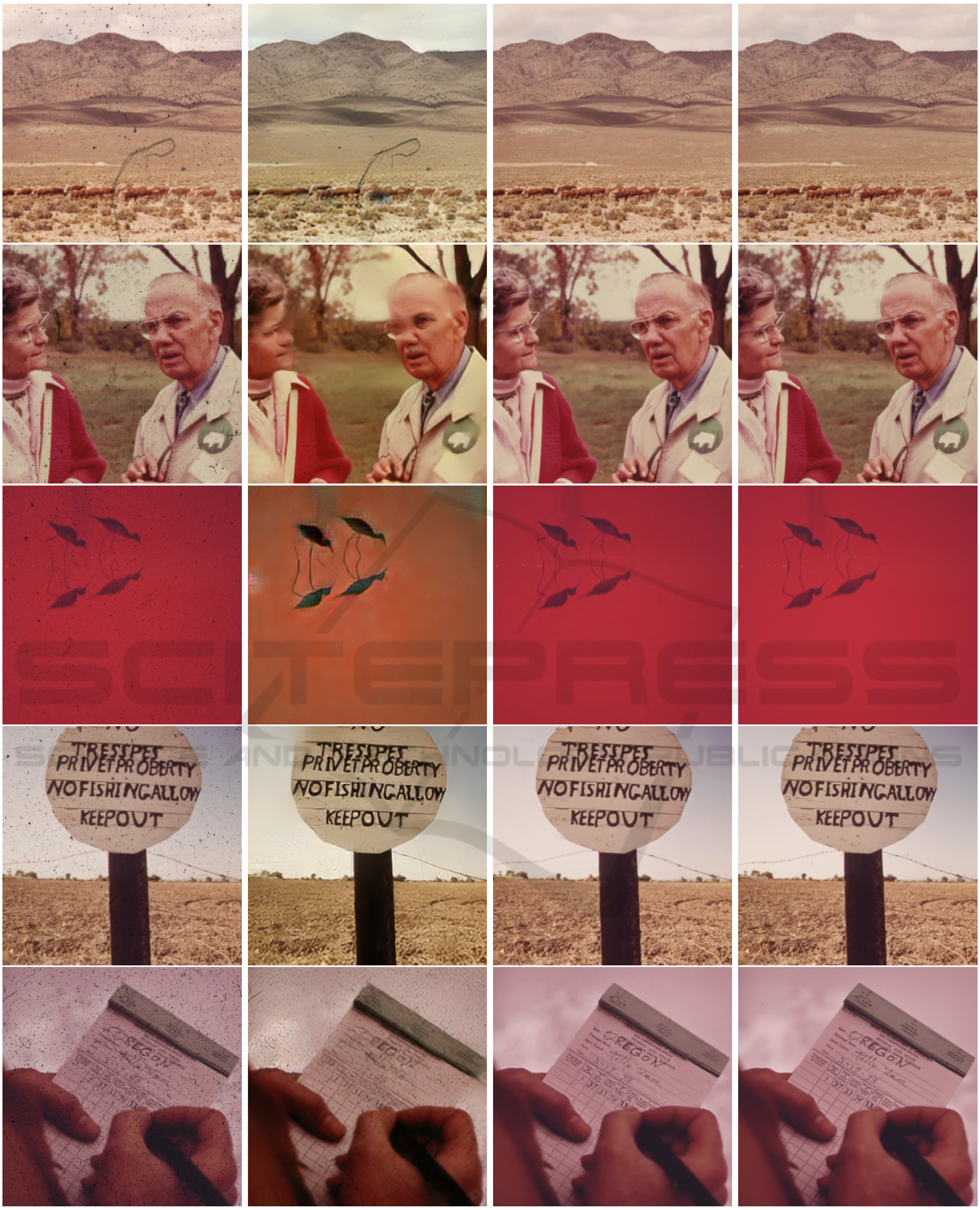
	SSIM of Predictions on dustified images		SSIM of Predictions on clean images	
	Mean	SD	Mean	SD
MSE Model	0.84	0.07	0.95	0.04
Wan et al.	0.68	0.06	0.72	0.07
Ours	0.88	0.05	0.98	0.02



(a) Wan et al., 2020

(b) Ours

Figure 5: Comparison between a state-of-the-art restoration approach against our approach based on achieved SSIM scores of restorations.



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

Figure 6: Visual comparison between the restored predictions for (a) damaged inputs generated by (b) VAE-based Deep Latent Space Translation (Wan et al., 2020) and (c) our approach, where (d) are the clean targets. Image in the first row is from the validation set, rows 2-5 are from the test set.

Comparing the histograms against the ones for the initial test set scores in Figure 3, we can see that our model performend better at improving the SSIM scores on the damaged test images. Additionally, we found that the VAE-based approach was highly damaging to clean images compared to ours. This is explained partially by the colour shifts, but also through the loss of fine detail and inability to remove larger scratches, as shown in Figure 5 (first row). Additionally, the Wan et al. method introduced checkerboard artifacts to some examples Figure 5 (third row) and struggled to reconstruct faces Figure 5 (second row). In the case of signs and handwriting, no meaningful differentiation is made between artifacts and lines forming the letters, which results in failure to restore writing Figure 5 (fourh and fifth row). On the other hand, our model has successfully targeted only existing artifacts, and minimised the introduction of new damage or loss of information.

5 CONCLUSION

The work presented in this paper demonstrates that our approach achieves improved quality of restoration at the task of automated dust and scratch removal for analogue film scans when compared to state-of-the-art. We adapt an architecture and training techniques from the literature, and use those along with our perceptual loss comprising of both exracted VGG16 feature activations and SSIM-based terms. By combining the learned natural prior of a pre-trained CNN-based architecture with a perceptual quality metric which targets image degradation in our loss formulation, we allow the network to meaningfully differentiate between dust and scratches and useful high-frequency image features. Our model achieved better SSIM scores compared to the VAE-based method of Wan et al.; while this can be attributed to our approach explicitly optimising for SSIM during training, our qualitative results demonstrate that our approach is much more reliable in both restoring dust and scratches, and preserving high frequency image detail.

Additionally, we discuss a more comprehensive approach to evaluating restoration quality, which also includes measuring the information loss or new artifacts introduced by the restoration networks. We also provide a data set of synthetically damaged slide film scans to be used for benchmarking of the specific task of dust and scratch removal for film.

As future work, we plan to collect a data set of wild damaged film scans to evaluate our approach and other existing approaches on real damaged input. Additionally, when training on synthetic data where the ground truth clean scan is available, we plan to explicitly incorporate our requirement that the network should not damage clean inputs in the loss formulation.

ACKNOWLEDGEMENTS

We thank Gerardo Aragon-Camarasa for his valuable comments and feedback while preparing this paper.

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/RS13222/1].

REFERENCES

Aitken, A., Ledig, C., Theis, L., Caballero, J., Wang, Z., and Shi, W. (2017). Checkerboard artifact free sub-pixel convolution: A note on sub-pixel convolution, resize convolution and convolution resize. *arXiv preprint arXiv:1707.02937*.

Antic, J. (2020). Deoldify: A deep learning based project for colorizing and restoring old images (and video!).

Bousquet, O., Gelly, S., Tolstikhin, I., Simon-Gabriel, C.-J., and Schoelkopf, B. (2017). From optimal transport to generative modeling: the vegan cookbook.

Chambah, M. (2019). Digital film restoration and image quality. In *ICA-BELGIUM Colour Symposium*, Ghent, Belgium.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Fan, Z., Wu, H., Fu, X., Huang, Y., and Ding, X. (2018). Residual-guide feature fusion network for single image deraining. *CoRR*, abs/1804.07493.

Fielding, G. (2008). Digital ice: Defect detection and correction using infrared-enabled scanners.

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.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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.

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*.

Meng Tang, L., Hong Lim, L., and Siebert, P. (2018). Removal of visual disruption caused by rain using cycle-consistent generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0.

Mironică, I. (2020). A generative adversarial approach with residual learning for dust and scratches artifacts removal. *Proceedings of the 2nd Workshop on Structuring and Understanding of Multimedia heritAge Contents*.

Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., and Wang, Z. (2016a). Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883.

Shi, W., Caballero, J., Theis, L., Huszar, F., Aitken, A., Ledig, C., and Wang, Z. (2016b). Is the deconvolution layer the same as a convolutional layer? *arXiv preprint arXiv:1609.07009*.

Smith, L. N. and Topin, N. (2019). Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006, page 1100612. International Society for Optics and Photonics.

Strubel, D., Blanchon, M., and David, F. (2019). Deep learning approach for artefacts correction on photographic films. In Cudel, C., Bazeille, S., and Verrier, N., editors, *Fourteenth International Conference on Quality Control by Artificial Vision*, volume 11172, pages 156 – 161. International Society for Optics and Photonics, SPIE.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. (2018). Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9446–9454.

Wan, Z., Zhang, B., Chen, D., Zhang, P., Chen, D., Liao, J., and Wen, F. (2020). Bringing old photos back to life.

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.

Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. (2018). Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*.

Zhang, R., Isola, P., and Efros, A. A. (2016). Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer.

Zhang, R., Zhu, J.-Y., Isola, P., Geng, X., Lin, A. S., Yu, T., and Efros, A. A. (2017). Real-time user-guided image colorization with learned deep priors. *ACM Transactions on Graphics (TOG)*, 9(4).

Zhang, X., Karaman, S., and Chang, S.-F. (2019). Detecting and simulating artifacts in gan fake images. In *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE.