Restoration of Analog Videos Using Swin-UNet

Lorenzo Agnolucci
lorenzo.agnolucci@unifi.it
Università di Firenze
Italy

Leonardo Galteri
leonardo.galteri]@unifi.it
Università di Firenze
Italy

Marco Bertini
marco.bertini@unifi.it
Università di Firenze
Italy

Alberto Del Bimbo
alberto.delbimbo@unifi.it
Università di Firenze
Italy

(a): Input **(b): Restored** **(c): Ground Truth**

Figure 1: Qualitative results of video restoration on the synthetic dataset

ABSTRACT

In this paper we present a system to restore analog videos of historical archives. These videos often contain severe visual degradation due to the deterioration of their tape supports that require costly and slow manual interventions to recover the original content. The proposed system uses a multi-frame approach and is able to deal also with severe tape mistracking, which results in completely scrambled frames. Tests on real-world videos from a major historical video archive show the effectiveness of our demo system.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision tasks**; **Reconstruction**; • **Information systems** → **Digital libraries and archives**.

KEYWORDS

Old Videos Restoration, Analog Videos, Swin Transformer, UNet

ACM Reference Format:
Lorenzo Agnolucci, Leonardo Galteri, Marco Bertini, and Alberto Del Bimbo. 2022. Restoration of Analog Videos Using Swin-UNet. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3503161.3547730

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10.
https://doi.org/10.1145/3503161.3547730

1 INTRODUCTION AND RELATED WORKS

Historical videos constitute an important part of the cultural heritage of a society. This content often is hampered by numerous artifacts and degradations due to technological limitations and aging of the recording support that limit its distribution and fruition by the general public. Normally the restoration of these videos is conducted frame by frame by experienced archivists with commercial solutions, thus at great economic and time cost.

For this reason, some works tried to restore historical video archives more rapidly and without human aid. [1] is an open-source tool for old films restoration that presents a NoGAN training approach. [7] relies fully on 3D convolutions and on source-reference attention for frame colorization. [11] proposes a recurrent transformer network that localizes defects in an unsupervised manner.

Istituto Luce Cinecittà, an Italian society responsible for the preservation and distribution of the *Archivio Storico Luce*, the largest Italian historical video archive dating from throughout the 1900s and comprising a variety of sources, provided us with some analog videos from this archive. These videos present several system intrinsic and aging-related types of degradations typical of analog video tapes. The related works focus on standard structured defects such as scratches and cracks, so they are not capable of restoring the particular types of artifacts that these videos present. Unfortunately, being real-world videos, there is no clean high-quality version of them to use as ground truth for supervised learning. Consequently we created a synthetic dataset as similar as possible to the real-world videos to train our system.

Table 1: Quantitative results on the synthetic dataset

Method	PSNR↑	SSIM↑	LPIPS↓
DeOldify [1]	11.56	0.451	0.671
Ours	**34.78**	**0.939**	**0.063**

2 PROPOSED APPROACH

2.1 Synthetic Dataset

In order to train a restoration model, we created a synthetic dataset as similar as possible to the real-world videos. Starting from high-quality videos of the Harmonic dataset [2] we used Adobe After Effects [4] to randomly add several types of degradations, such as:

- Gaussian noise, resembling the tape noise that is typical of analog videos;
- white artifacts simulating tape dropouts;
- cyan, magenta and green horizontal lines resembling chroma fringing;
- horizontal displacements, similar to tape mistracking artifacts; this is the most complex error that can be encountered.

As with the real-world videos, all these artifacts vary over time and occur with different intensity, positions and combinations for each frame.

We ended up with 26392 frames that we divided into training and validation sets with an 80-20 ratio. Then this synthetic dataset was used to train the model described in section 2.2.

2.2 Network Architecture

Inspired by [3], we developed a Swin-UNet architecture presented in Figure 2. Differently from [3] our network works on videos, so we converted it to a multi-frame approach. In this way, the model enhances T frames at once exploiting spatio-temporal information. Moreover, we employed 3D convolution for partitioning the input into patches and pixel shuffle for the patch expanding layer. Following [6], a skip connection between the degraded input and restored output makes the network learn the residual of each frame. This choice reduces the overall training time and improves its stability.

The training loss is a weighted sum of a pixel loss (in particular, the Mean Square Error) and a perceptual loss [5, 8, 9] defined on the VGG-19 [10] feature space. The network was trained with 256×256 patches cropped randomly from the input frames. During training and testing the number of frames T processed by the model was fixed to 5.

3 RESULTS

We measured the performance of our method using three standard full-reference visual quality metrics: 1) PSNR; 2) SSIM [12]; 3) LPIPS [13]. The quantitative results obtained for the restoration of the 512×512 central crop on the synthetic dataset are reported in Tab. 1. For a fair comparison, we re-trained DeOldify [1] from scratch using our training data. Our model achieves the best performance.

The qualitative results for the synthetic and real-world dataset are presented in Figure 1 and 3, respectively. Our model proves to be able to restore a lot of details lost with the heavy degradations to which the input frames had been subjected. Indeed, thanks to the spatio-temporal information captured by our multi-frame approach the network can exploit the time-varying nature of the artifacts and address the most severe tape mistracking.

To let users restore degraded videos with similar artifacts we developed a Flask-based demo web app accessible through a web browser. Our platform supports the upload of video files and provides the user with the downloadable restored result, as well as with a comparison with the original video. Alternatively, the user can choose one of our example videos just to see what our model is capable of.

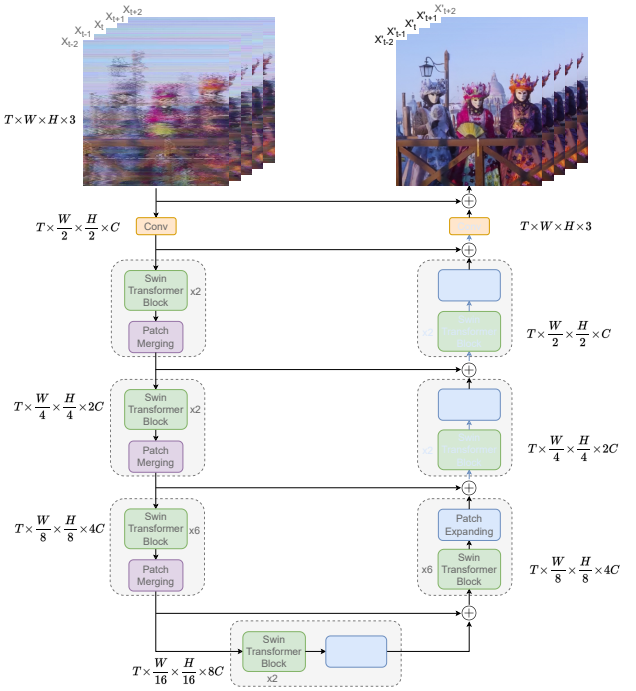

Figure 2: Proposed network architecture.

(a): Input **(b): Restored**

Figure 3: Qualitative results on the real-world video of the test dataset

4 CONCLUSIONS

In this work we focused on restoring analog videos of historical archives. We created a synthetic dataset to train the Swin-UNet network we designed. Tests on synthetic and real-world videos prove the effectiveness of our approach. We also developed a demo web-app to let users restore videos with similar artifacts.

ACKNOWLEDGMENTS

This work was partially supported by the European Commission under European Horizon 2020 Programme, grant number 101004545 - ReInHerit.

REFERENCES

[1] 2018. DeOldify. https://github.com/jantic/DeOldify

[2] 2019. Harmonic free 4K demo footage. https://www.harmonicinc.com/free-4k-demo-footage/

[3] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. 2021. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537* (2021).

[4] Mark Christiansen. 2013. *Adobe After Effects CC Visual Effects and Compositing Studio Techniques.* Adobe Press.

[5] Alexey Dosovitskiy and Thomas Brox. 2016. Generating images with perceptual similarity metrics based on deep networks. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS).*

[6] Leonardo Galteri, Marco Bertini, Lorenzo Seidenari, Tiberio Uricchio, and Alberto Del Bimbo. 2020. Increasing Video Perceptual Quality with GANs and Semantic Coding. In *Proc. of ACM International Conference on Multimedia (ACM MM)* (Seattle, WA, USA). 9 pages. https://doi.org/10.1145/3394171.3413508

[7] Satoshi Iizuka and Edgar Simo-Serra. 2019. DeepRemaster: Temporal Source-Reference Attention Networks for Comprehensive Video Enhancement. *ACM Transactions on Graphics (Proc. of SIGGRAPH Asia 2019)* 38, 6, Article 176 (2019), 13 pages.

[8] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Proc. of European Conference on Computer Vision (ECCV).*

[9] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[10] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[11] Ziyu Wan, Bo Zhang, Dongdong Chen, and Jing Liao. 2022. Bringing Old Films Back to Life. *CVPR* (2022).

[12] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing (TIP)* 13, 4 (2004), 600–612.

[13] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*