Training a Deep Remastering Model

Abdelaziz Djelouah
abdelaziz.djelouah@disney.com
DisneyResearch|Studios
Switzerland

Andrew J. Wahlquist
andrew.j.wahlquist@disney.com
The Walt Disney Company
USA

Sally Hattori
sally.hattori@disney.com
The Walt Disney Company
USA

Christopher Schroers
christopher.schroers@disney.com
DisneyResearch|Studios
Switzerland

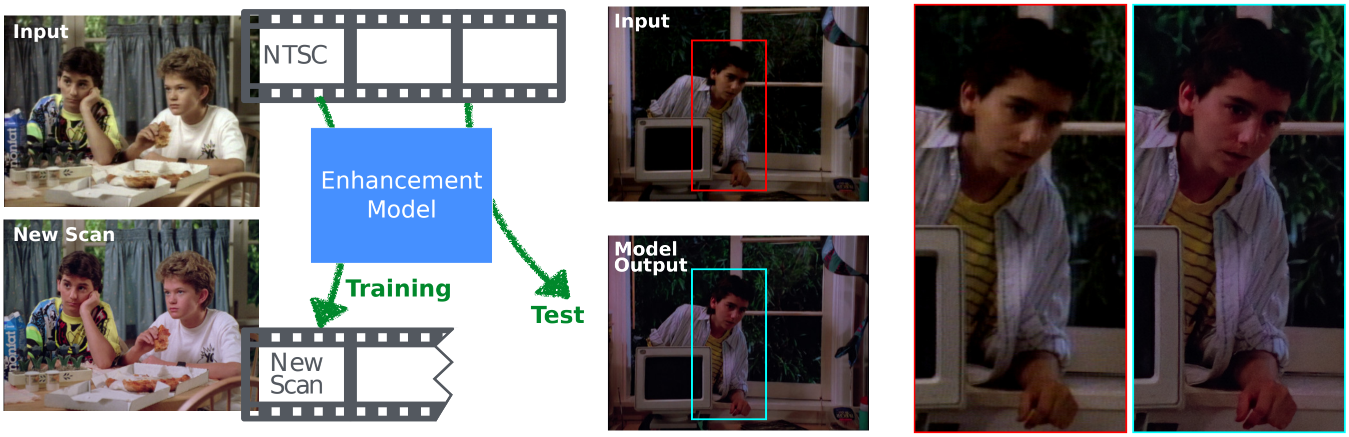

Figure 1: One possible strategy for video restoration is to rescan original film, however film reels may be missing or destroyed. We propose a framework to train an image enhancement model that benefits from the available film. This model is then applied to the broadcast version to produce new high quality frames.

ABSTRACT

The success of video streaming platforms has pushed studios to make available TV shows from legacy catalog, and there is an increased demand for remastering this content. Ideally, film reels are re-scanned with modern devices directly into high quality digital format. However this is not always possible as parts of the original film reels can be damaged or missing, and the content is then available in its entirety only in the broadcast version, typically NTSC. In this work, we present a deep learning solution to bring the NTSC version to the new scan quality levels, which would be otherwise impossible with existing tools.

ACM Reference Format:
Abdelaziz Djelouah, Andrew J. Wahlquist, Sally Hattori, and Christopher Schroers. 2022. Training a Deep Remastering Model. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Talks (SIGGRAPH '22 Talks), August 07-11, 2022.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3532836.3536228

1 INTRODUCTION

The quality and size of the content catalog is one of the key factors valued by users of streaming platforms. In addition to the constant flow of original creations, it is also the opportunity for older films and TV shows to find new exposures to large audiences. This content is however often noisy, blurry and in lower visual quality in general. Video remastering is then needed to bring this legacy data to today's standards. It is possible to do so by using video processing toolboxes which are slowly integrating progress made in the scientific community on image enhancement problems such as denoising [Tassano et al. 2020], super-resolution [Cornillere et al. 2019; Wang et al. 2018] or the combination of multiple degradations. Another alternative is to rescan original film and benefit from the improvement made over the years in terms of sensors and hardware in general. This is illustrated in the left part of Figure 1. One can notice the gap in quality between the available NTSC format, which was used for TV broadcast, and the result obtained from rescanning the tapes with modern technology.

However there is an issue: it is possible, and it often is the case, that original film reels are missing or damaged. In such situations, the only remaining option was to rely on human users to combine existing image enhancement techniques (denoising, color grading, etc.). Given the important gap in quality, this would require important efforts and may not even be possible.

SIGGRAPH '22 Talks, August 07-11, 2022, Vancouver, BC, Canada
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9371-3/22/08.
https://doi.org/10.1145/3532836.3536228

In this work, we develop a framework to address this scenario, by leveraging the parts of the film that have been successfully rescanned as training data for an image enhancement model. This model learns a mapping from the broadcast version to the high quality target. Once trained, it can be used to process any frame as illustrate in Figure 1 to perform a complex enhancement that includes color grading, artifact removal, denoising and grain synthesis.

2 METHOD

2.1 Enhancement Model

We use Generative Adverserial Networks (GANs) for this enhancement problem. In particular we use a neural network architecture similar to [Wang et al. 2018] for the generator model. It consists of several dense compression units using residual connections. Additionally we apply a pixel-shuffle operation on the input. This enhancement model is denoted as G with parameters θ_g. It should predict the high quality output I^* given the low quality input I_l.

$$I^* = G(I_l \mid \theta_g). \tag{1}$$

In the first stage of the training, we only use an ℓ_1 loss on the difference of the FFT decomposition (F_{fft}) of the target image I and the model prediction.

$$\mathcal{L}_{\text{fft}} = \ell_1(F_{\text{fft}}(G(I_l)), F_{\text{fft}}(I)). \tag{2}$$

After 500k optimization steps, we switch to an adversarial training regime, using the discriminator architecture similar to [Wang et al. 2018]. The discriminator D is trained to predict if the input image is real (output is 1) or fake (output is 0). We use the least-squares formulation and additionally use the VGG-loss to enforce texture similarity. During this GAN training stage, the generator is trained to optimize the loss

$$\mathcal{L}_g = \mathcal{L}_{\text{fft}} + \lambda_1 (D(G(I_l)) - 1)^2 + \lambda_2 \sum_k ||\Phi_k(G(I_l)) - \Phi_k(I)||^2 \tag{3}$$

where Φ_k is the k-th pooling layer of the VGG architecture. λ_1 and λ_2 are respectively set to 3.0 and 0.5. The discriminator training step optimizes for the following loss

$$\mathcal{L}_d = D(G(I_l))^2 + (D(I) - 1)^2 \tag{4}$$

2.2 Alignment Procedure

The enhancement model assumes that the low quality input and the high quality targets are already aligned which is not the case. A first step is to deinterlace the broadcast version. For this we can use inverse telecine toolboxes or more recent deep deinterlacing techniques [Bernasconi et al. 2020]. A second step is temporal alignment. Complex temporal alignment algorithms can be considered such as [Wang et al. 2014] but in our use cases a fixed offset was sufficient. Finally geometric alignment is needed. Indeed the video telecine process that is used for the original production, relies on a camera lens to capture a projected film image onto video tape. This is different than the film scanning today, in which a datacine would be scanning similarly to a flatbed image scanner. It is also not possible to compute a single distortion map for the entire show, as the broadcast version depends on how the telecine camera lens was positioned, which can is different cut to cut. Our solution is to first compute a sparse set of spatial correspondences between two temporally matching frames and used that to optimize for a global translation and non uniform scaling. We perform then another dense alignment step to eliminate remaining misalignment.

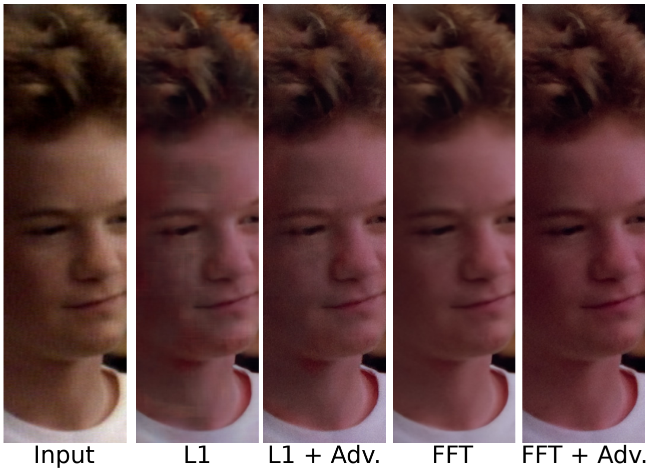

Figure 2: Results with Different Losses. As the input has many artifacts such as low frequency noise, grain and blur, using the FFT loss is needed to produce best results.

3 RESULTS

At test time the broadcast version is re sampled according to the desired output transform: simple scaling or more complex. We use bilinear interpolation at this stage. This resampled low quality input is then processed by our image enhancement model. Figure 1 shows an example of both training data and results. One can notice the complex enhancement that happens in the image which consists in color grading, added details and noise pattern.

To demonstrate the importance of the proposed $\mathcal{L}_{\text{fft}}$ loss we train the same neural network model with the L_1 loss commonly used for super-resolution. We can notice in Figure 2 the artifacts due to the low frequency noise present in the input. These are reduced with the adversarial but it is clearly not sufficient. The model trained with the $\mathcal{L}_{\text{fft}}$ loss is more robust to the artifacts present in the input. In this case, adding the adversarial training stage really helps creating both texture details and noise pattern that better match the target film quality. After review, we have confirmed the quality of the results and only a minor post-processing was applied before integration with material that was fully remastered from film.

REFERENCES

Michael Bernasconi, Abdelaziz Djelouah, Sally Hattori, and Christopher Schroers. 2020. Deep deinterlacing. In *SMPTE Annual Technical Conf. Exhibition.*

Victor Cornillere, Abdelaziz Djelouah, Wang Yifan, Olga Sorkine-Hornung, and Christopher Schroers. 2019. Blind image super-resolution with spatially variant degradations. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 1–13.

Matias Tassano, Julie Delon, and Thomas Veit. 2020. FastDVDnet: Towards Real-Time Deep Video Denoising Without Flow Estimation. In *CVPR.*

Oliver Wang, Christopher Schroers, Henning Zimmer, Markus Gross, and Alexander Sorkine-Hornung. 2014. Videosnapping: Interactive synchronization of multiple videos. *ACM Transactions on Graphics (TOG)* 33, 4 (2014), 1–10.

Yifan Wang, Federico Perazzi, Brian McWilliams, Alexander Sorkine-Hornung, Olga Sorkine-Hornung, and Christopher Schroers. 2018. A Fully Progressive Approach to Single-Image Super-Resolution. In *CVPR Workshops.*