

Finetuning Variational Autoencoders for the Restoration of Sports Video Highlights

Michael Dobioli Siddharth Sharma



Introduction

- Advances in Computer Vision have opened the possibility of re-constructing older images at a higher resolution.
- This is an especially salient problem given the number of old images that requires immediate restoration or risk being lost forever.
- Sports highlights from the pre-2000s era are quite poor by modern standards.



Figure 1. Past Football Footage



Figure 2. Old Tennis Footage

- We use Variational Autoencoders (VAEs) to restore past highlights with higher quality.
- We approach this problem by splitting videos into frames, using a finetuned model on each frame, and then applying postprocessing video normalization methods (sharpening and denoising) to ensure consistent and better quality.

Background

- Prior to the deep learning era, image restorations were done by geometric method such as Bicubic or Lanczos filtering that either emphasized or dampened certain edges and features.
- However, such interpolation methods often fail to capture the full complexity of the image, meaning that they are often incomplete for the purposes of generating realistic restored images. We can do better by leveraging the training data in order to create a model that specializes in restoring images within the problem space.
- We reframe the problem of restoration as learning a mapping between a space of low resolution and high resolution images.
- Learning a mapping between domains lies firmly in the field of generative models. GANs have been used by Ledig et al. in a two-step process: a generator first creates high-resolution images, and then a discriminator differentiates the generated super-resolution image from a high resolution image

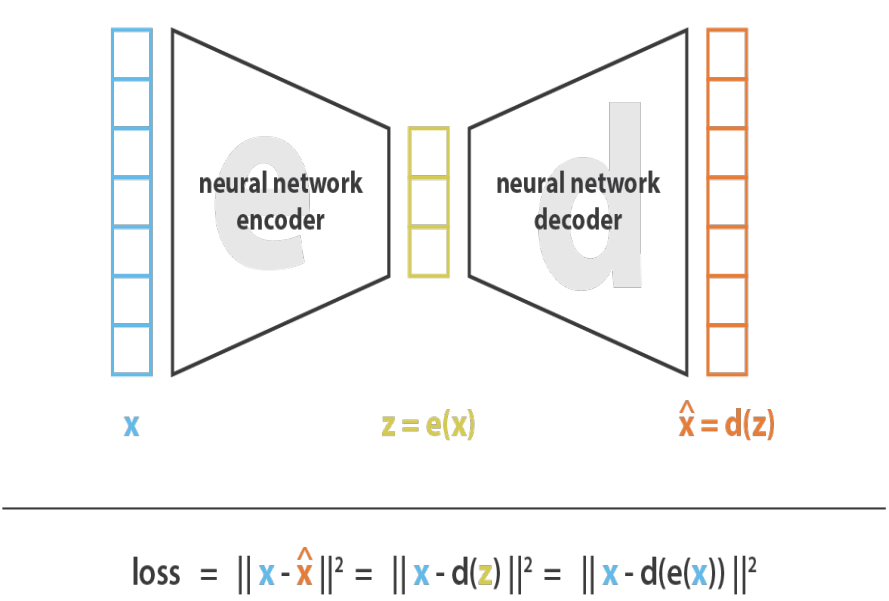


Figure 3. Autoencoder

References

[1] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. 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, 2017.

[2] Ziyu Wan, Bo Zhang, Dongdong Chen, Pan Zhang, Dong Chen, Jing Liao, and Fang Wen. Bringing old photos back to life. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2747–2757, 2020.

Methods

- We use Wan et al.'s model as a baseline upon which we finetune.
- Typically when training generative models for the purposes of restoration, the model uses pairs of synthetically generated pictures and compares it relative to the original picture.
- The key insight Wan et al. is that synthetically generated old images and real old images can differ quite significantly, as **Figure 4** and **Figure 5** show.



Figure 4. Real old image



Figure 5. Synthesized old image

- Hence, a mapping from the synthesized old images to the real modern images may not generalize to mapping real old images to restored versions.
- The key is that instead of learning how to map synthesized old images to real images, we learn a mapping between latent spaces generated by VAEs. Upon mapping old images to one latent space and modern images to a different latent space, we learn a mapping between these two latent spaces.

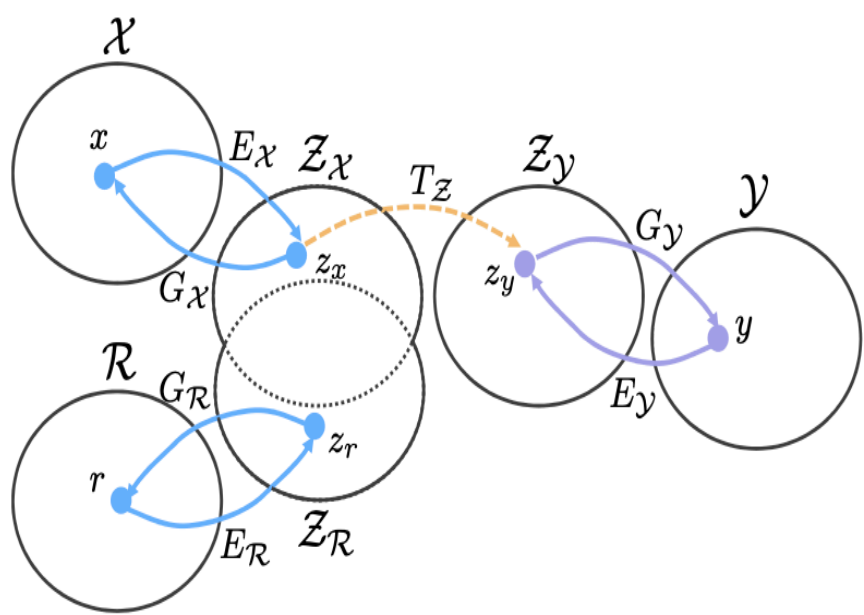


Figure 6. Mapping of Latent Spaces. $\mathcal{R}$ represents the real old image space, $\mathcal{X}$ represents the space of synthesized old images, and $\mathcal{Y}$ represents the space of real modern images. $\mathcal{Z}_X, \mathcal{Z}_R, \mathcal{Z}_Y$ are the corresponding latent spaces.

Finetuning

- Data-based finetuning:** We construct three datasets: one containing non-synthesized old images in RGB format, one containing non-synthesized old images in grayscale, and one containing modern sports images in RGB format.
- Data was preprocessed and extracted from YouTube (2,000 frames for train, 300 for validation, and 300 for test) via the Wimbledon Championships.
- We chose tennis for a few reasons: camera angles are relatively consistent, the crowd and background image features are out of the main frame during play, and data for long matches is publicly available for extraction.
- Hyperparameter Optimization:** We manipulate parameters such as the learning rate (α), batch size, and weight initialization. We keep the Adam optimizer constant.
- Postprocessing:** For image sharpening, we apply a 2D filter based on a given kernel and for Gaussian denoising, we use opencv's bilateral filter. These techniques reduce noise while maintaining edges, thus giving our restorations a smoother look.
- We also considered other techniques such as Sobel image derivatives, Laplacians, contrast and brightness scaling, and histogram equalization.

Experiments

- We ran several experiments with varied hyperparameters. Note that the default LR was $2 \cdot 10^{-4}$ and due to CUDA memory constraints, we could only operate with batch sizes of 16 and 32.
- We chose our hyperparameters based on deviation from the default parameters used by Wan et al. We used learning rates such as $2 \cdot 10^{-3}$, $2 \cdot 10^{-4}$, $2 \cdot 10^{-5}$ and we used weight initialization methods such as Xavier and Kaiming.
- Following experiments, we performed inference (restorations) using the checkpoints from the three training steps (domain A, domain B, mapping)

Results

- We used FID scores (Fréchet inception distance) to calculate how well model performed at restoring blurred images. A lower FID score means better restorations.
- The finetuned model performed significantly better than the baseline model and postprocessing further lowered the FID score.

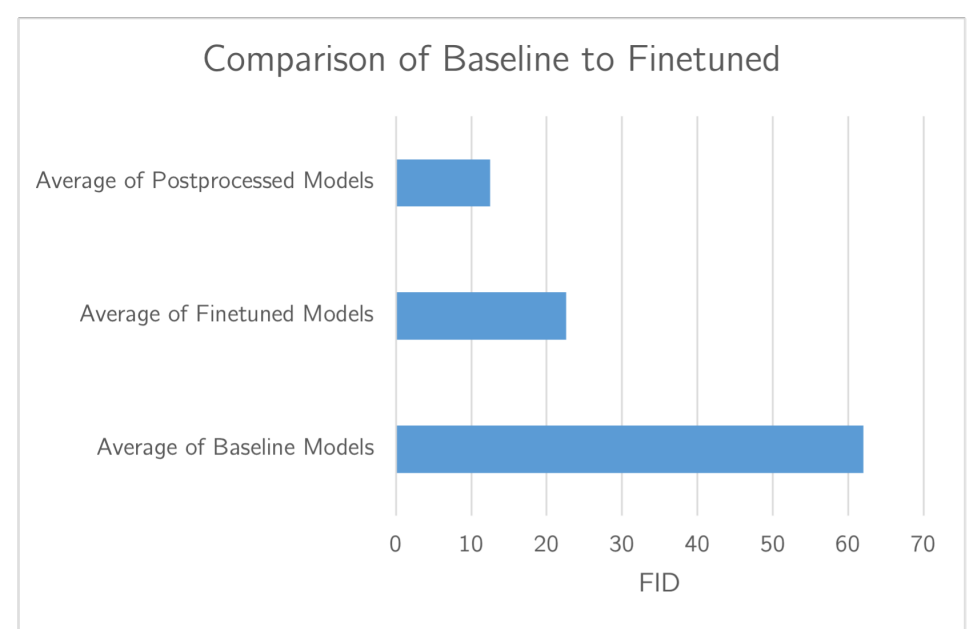


Figure 7. FID Score comparison for Model averages

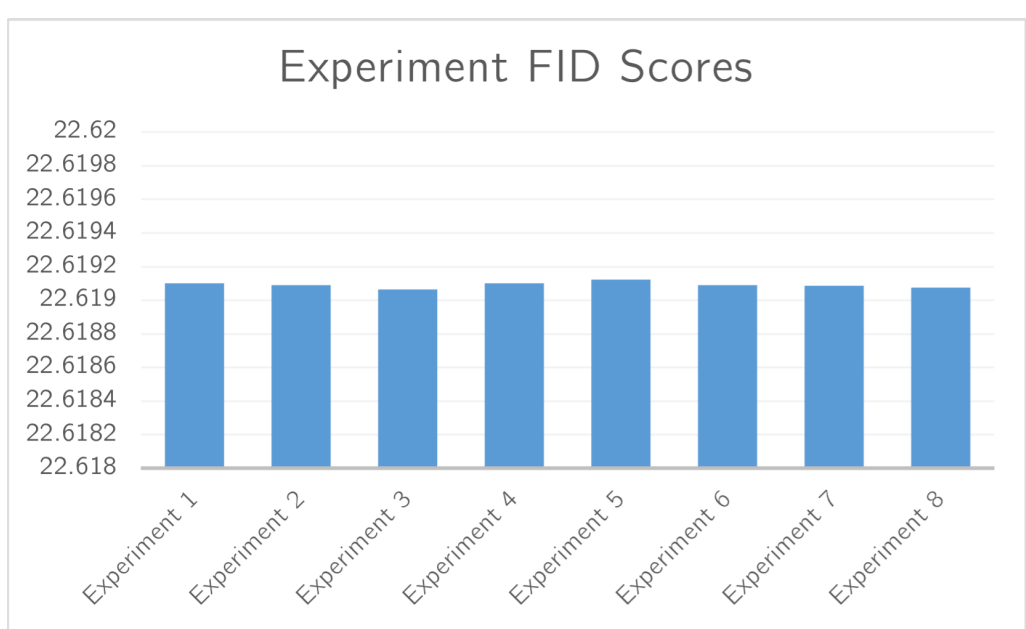


Figure 8. FID Scores for Finetuned Models

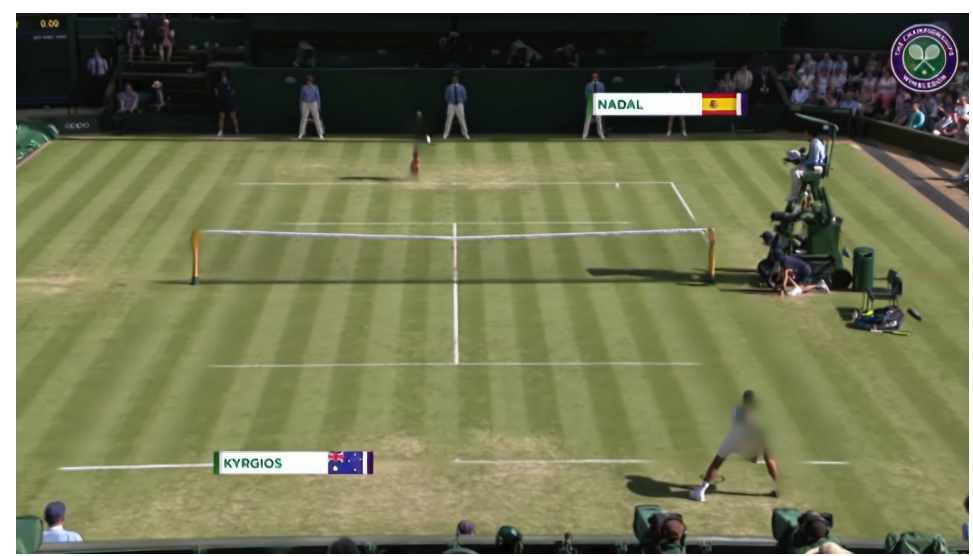


Figure 9. Baseline model restoration

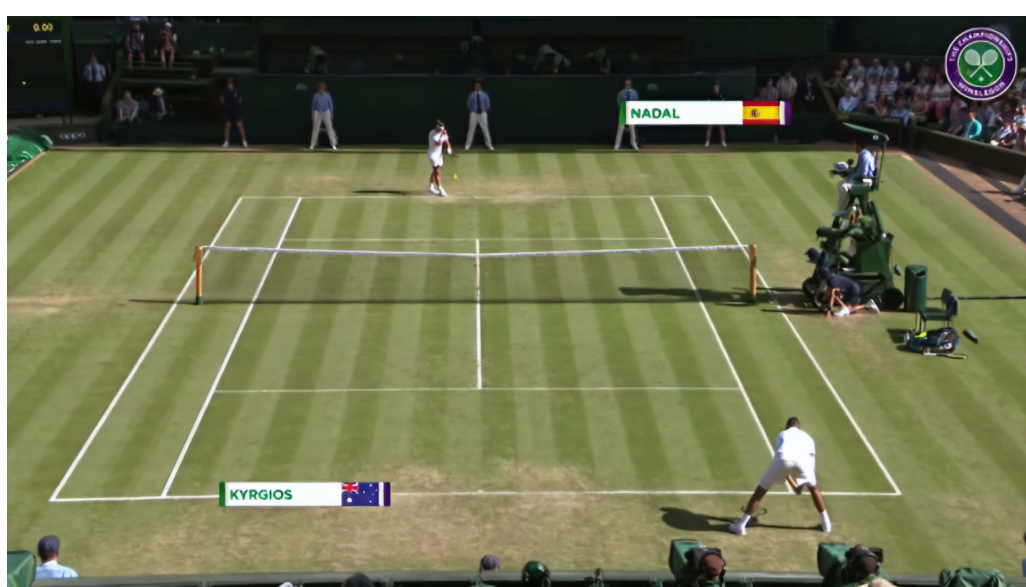


Figure 10. Finetuned model restoration

Conclusion

- In this work, we establish that finetuning VAEs can be used to restore old sports highlights compared to previous baseline results.
- The model had consistent performance across all parameters, often demonstrating very similar FID scores.
- With additional postprocessing methods, videos can achieve even lower FID scores than using a finetuned model alone.
- These same techniques can be applied to other court types and perhaps even other sports.
- Based on quantitative metrics and qualitative analysis, we are confident that our work can be used to restore old sports highlights for modern audiences to appreciate.
- Some extensions of this project could be to add more training epochs, diversify and enlarge the training set, and experiment with restoration of different sports.