Supplementary Material - Bringing Old Films Back to Life

Ziyu Wan[1] Bo Zhang[2] Dongdong Chen[3] Jing Liao[1*]
[1]City University of Hong Kong [2]Microsoft Research [3]Microsoft Cloud + AI
{raywzy,cddlyf}@gmail.com zhanbo@microsoft.com jingliao@cityu.edu.hk

1. Overview

In this supplemental material, additional experimental details and results are provided, including:

- More details about network architecture(Section 2);
- More qualitative restoration comparisons with baselines (Section 3);
- User study results (Section 4);
- Video comparison results. Please refer to *video.mp4*.

2. Network Architecture

Table. 1 and Table. 2 show the employed architecture of RTN and discriminator $\mathcal{D}$. We adopt Swin [4] as the transformer block for better efficiency, where the features are projected into 128 and then back to 64 dimensions via MLPs. For each convolution layer of $\mathcal{D}$, we use spectrum normalization (SN) [5] to stabilize the adversarial training procedure.

MODULE	LAYER	KERNEL SIZE / STRIDE	CHANNEL #	NON-LINEARITY
Spatial Restoration $\mathcal{R}$	2DConv	$3 \times 3/(1,1)$	$16 \rightarrow 32$	LeakyReLU(0.2)
	2DConv	$4 \times 4/(2,2)$	$32 \rightarrow 64$	LeakyReLU(0.2)
	Transformer$\times 8$	8×8 / Local Attention	$64 \rightarrow 128 \rightarrow 64$	GELU
	2DConv	$3 \times 3/(1,1)$	$64 \rightarrow 64$	LeakyReLU(0.2)
	Bilinear Upsample	–	$64 \rightarrow 64$	–
	2DConv	$3 \times 3/(1,1)$	$64 \rightarrow 32$	LeakyReLU(0.2)
	2DConv	$3 \times 3/(1,1)$	$32 \rightarrow 16$	–
Temporal Fusion $\mathcal{F}$	2DConv	$3 \times 3/(1,1)$	$3 \rightarrow 16$	LeakyReLU(0.2)
	2DConv	$3 \times 3/(1,1)$	$32 \rightarrow 8$	LeakyReLU(0.2)
	2DConv	$3 \times 3/(1,1)$	$8 \rightarrow 4$	LeakyReLU(0.2)
	2DConv	$3 \times 3/(1,1)$	$4 \rightarrow 1$	sigmoid
Pixel Decoder $\mathbf{D}$	2DConv	$3 \times 3/(1,1)$	$32 \rightarrow 16$	LeakyReLU(0.2)
	2DConv $\times 3$	$3 \times 3/(1,1)$	$16 \rightarrow 16$	LeakyReLU(0.2)
	2DConv	$3 \times 3/(1,1)$	$16 \rightarrow 3$	tanH

Table 1. **Detailed architecture of RTN.**

3. Qualitative Comparisons

We further present the qualitative comparisons with BasicVSR [2] and Real-ESRGAN [6] in this section and Figure. 1. BasicVSR [2], which leverage a recurrent architecture as well, have demonstrated great performance on video super-resolution. Nonetheless, as mentioned in the main submission, unlike video super-resolution, which mainly focuses on the unstructured

*Corresponding author.

MODULE	LAYER	KERNEL SIZE / STRIDE	OUTPUT CHANNEL	NON-LINEARITY
Discriminator $\mathcal{D}$	SN-3DConv	$3 \times 5 \times 5/(1,2,2)$	64	`LeakyReLU`(0.2)
	SN-3DConv	$3 \times 5 \times 5/(1,2,2)$	128	`LeakyReLU`(0.2)
	SN-3DConv	$3 \times 5 \times 5/(1,2,2)$	256	`LeakyReLU`(0.2)
	SN-3DConv	$3 \times 5 \times 5/(1,2,2)$	256	`LeakyReLU`(0.2)
	SN-3DConv	$3 \times 5 \times 5/(1,2,2)$	256	–

Table 2. **Detailed discriminator $\mathcal{D}$ structure.** SN: Spectral Normalization [5].

degradation, the mixed degradation issues make BasicVSR hard to generalize to old films, always leading to over-smoothed results and leaving the contaminants unresolved. Recently, Real-ESRGAN [6] have shown great potentials for single real-world image restoration. We extend it to video and further leverage TS [3] for temporal consistency. As shown in the third column of Figure. 1, without appropriately using the temporal clues, Real-ESRGAN [6] is not able to render reasonable texture details either. Compared with these baselines, our method, by contrast, could solve these degradations well and generate appealing frames.

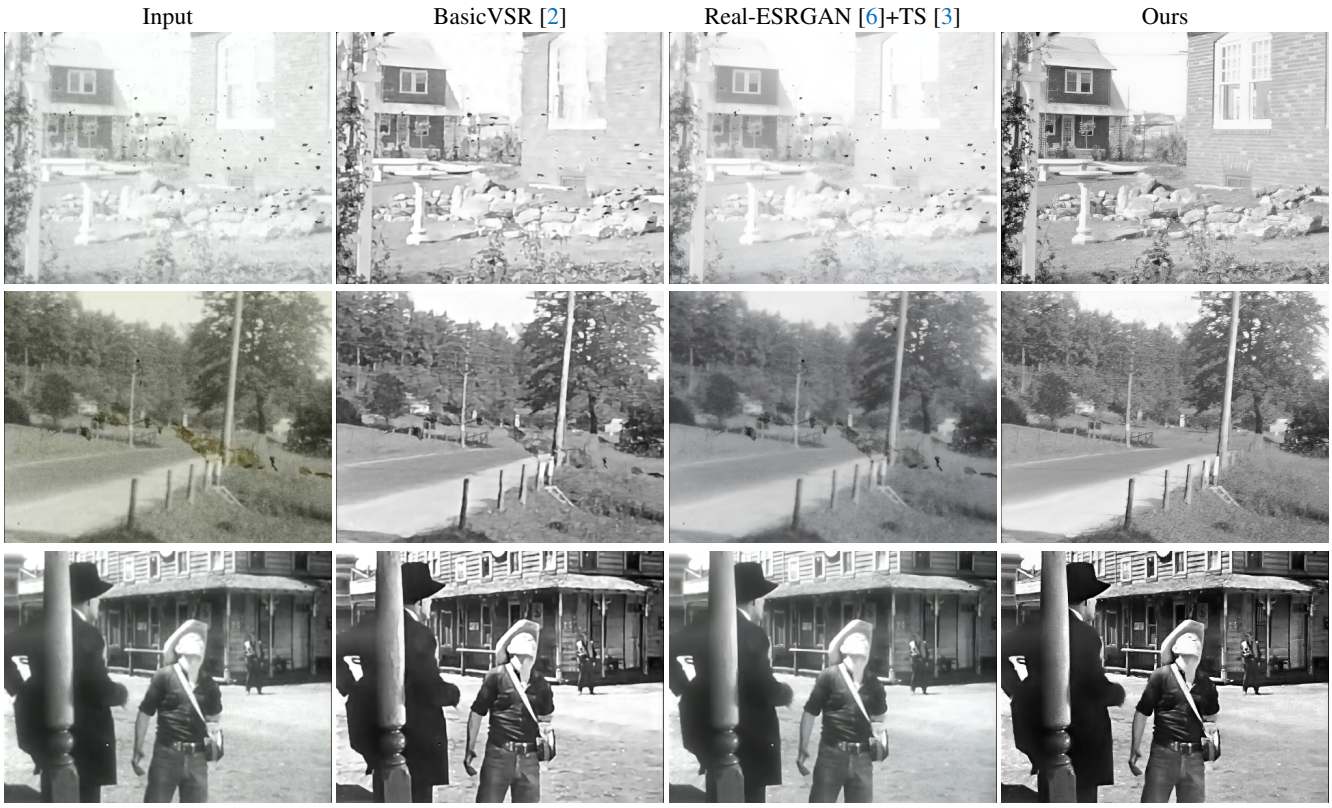

Figure 1. Qualitative restoration comparisons with BasicVSR [2] and Real-ESRGAN [6] on real-world old films. Our method could handle complicated degradations of old films. Zoom-in for more details.

4. User Study

To better evaluate the subjective old film restoration quality, we further conduct a user study to compare our method with other baseline methods. Specifically, we randomly select 20 old films from the collected test set. For each old film, we use our method and other baselines to restore it and then ask the participants to rank the six results from the highest one to the lowest one, based on comprehensive inspections of various aspects, including video temporal coherence, texture sharpness, noise degree, and scratch/dirt removal. We have collected the subject surveys from 25 participants and calculated the percentages of each method being selected as top 1,2,3, whose statistics is shown in Figure. 2. Our method demonstrates clear superiorities

over other methods. More than 58% users pick us as the best restoration result, surpassing the second percentage 11.1% of DeOldify [1] and BasicVSR [2] by a large margin.

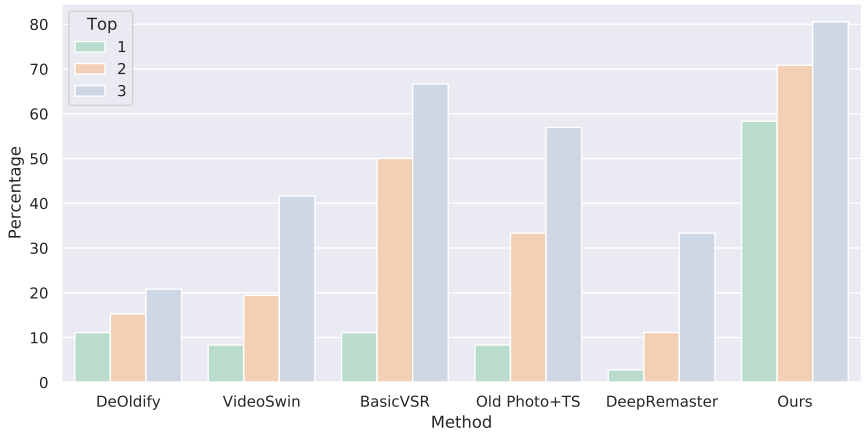

Figure 2. **Results of user study.**

References

[1] Deoldify. https://github.com/jantic/DeOldify. 3

[2] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4947–4956, 2021. 1, 2, 3

[3] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 170–185, 2018. 2

[4] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, October 2021. 1

[5] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 1, 2

[6] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. 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, 2021. 1, 2