Supplementary Material - Making Old Film Great Again: Degradation-aware State Space Model for Old Film Restoration

First Author
Institution1
Institution1 address
firstauthor@i1.org

Second Author
Institution2
First line of institution2 address
secondauthor@i2.org

1. Overview

In the supplementary material, we provide additional experiments and detailed information referenced in the paper, including:

- Comprehensive definitions of the loss functions discussed in Section 3.5, as detailed in Sec. 2.
- Further details on data synthesis (in Sec. 3) and additional examples of real-world scenes (in Sec. 4) included in the proposed datasets, elaborated in Section 4.
- Specific network parameter configurations, as outlined in Section 5.1 (in Sec. 5).
- Results of the user study conducted on the real-world portion of the dataset (in Sec. 6).
- Additional examples showcasing visual comparisons of restoration results (in Sec. 7).

2. Loss Functions

To optimize our proposed method and restore visually pleasing results from old films, following the general setup in the restoration tasks, we employ the following loss functions.

L1 Loss The mean absolute deviation loss is always used for pixel-wise reconstruction, which is formulated as:

$$\mathcal{L}_1 = \frac{1}{T}\sum_{i=1}^{T} \|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\|_1, \tag{1}$$

where T is the length of the input sequence.

Perceptual Loss To improve the perceptual quality of the results, we employ the perceptual loss [5]. Compared to pixel-level reconstruction, perceptual quality optimization can make the results more aligned with human visual preferences. The loss is defined as:

$$\mathcal{L}_{\text{P}} = \frac{1}{T}\sum_{i=1}^{T}\sum_{p\in P} \omega_p \|\Phi_p(\boldsymbol{x}_i) - \Phi_p(\hat{\boldsymbol{x}}_i)\|, \tag{2}$$

where Φ_p is the p-th layer of the pretrained VGG [5], while ω_p is the weight of each layer.

Spatial-Temporal Adversarial Loss Generative Adversarial loss [3] is widely used in restoration tasks [6, 9] to enhance the visual quality of the results. We employ the improved loss function from [2] to optimize our networks and the discriminator. The discriminator is optimized by hinge loss:

$$\mathcal{L}_D = \mathbb{E}_{x\sim \boldsymbol{X}}[\sigma(1 - D(x))] + \mathbb{E}_{\hat{x}\sim \boldsymbol{X}}[\sigma(1 + D(\hat{x}))], \tag{3}$$

where $D(\cdot)$ is the discriminator and σ is the ReLU function. Then, the restoration network is optimized by:

$$\mathcal{L}_G = -\mathbb{E}_{x\sim \boldsymbol{X}}[D(x)]. \tag{4}$$

The discriminator and generator are optimized in the same iteration using two optimizers each.

Total loss Therefore, the total loss is that:

$$\mathcal{L}_{\text{total}} = \lambda_1\mathcal{L}_1 + \lambda_p\mathcal{L}_{\text{P}} + \lambda_G\mathcal{L}_G. \tag{5}$$

The hyper-parameters $\lambda_1, \lambda_P, \lambda_G$ are empirically set to $\{1.0, 1.0, 0.01\}$.

Figure 1. Visual examples of the degraded video clips. The figure displays three levels of degradation, from top to bottom: mild, moderate, and extreme.

3. Data Synthesis

To train and evaluate the proposed MambaOFR, we follow existing works [4, 9, 10] to synthesize pairs of clean and degraded video sequences. The degradation details are as follows:

- **Structured defects** We use the old film effect templates provided in [9] to simulate structured defects (e.g., scratches, stains, banding noise, dust) commonly observed in old films. Random data augmentations, including affine transformations such as flipping and rotation, are applied to each template.
- **Noise** Gaussian and Speckle noise with a standard deviation $\sigma \in [5, 50]$ are used to introduce noise..
- **Blur** Isotropic and anisotropic Gaussian blur kernels are applied, with rotation angles $\theta \in [0, \pi]$, and principal axis standard deviations $\sigma_1, \sigma_2 \in (0, 1)$.
- **Compression** Random upsampling and downsampling in the range of $2x \sim 4x$, along with JPEG compression (Compress level $\in [40, 100]$), are utilized to mimic storage-related compression degradation.
- **Fading** Random brightness ($\gamma \in [0.8, 1.2]$) and contrast ($\beta \in [0.9, 1.0]$) jitter are employed to simulate fading effects.

In addition, to ensure consistency of degradation in the temporal domain, we apply the same set of degradations within each video clip. As shown in Fig. 1, the degradations in the testing set are categorized into three different intensity levels based on interval averages. This setup allows us to evaluate the method's performance under varying degrees of degradation.

4. Real-world Old Films

Compared to synthetic data, old real-world films exhibit more complex degradations that cannot be easily captured by simple degradation models. Therefore, in our proposed dataset, we have collected a large number of old real-world films. The diversity of degradation sample types in the dataset is extremely important for research in this field. These video clips not only encompass a rich variety of scenes but also include diverse degradations. In this section, we provide examples of different types of degradations to illustrate the richness of our proposed dataset.

Figure 2. Visual examples of real-world old films with different categories of degradation.

5. Details of Parameter Configure

In this section, we provide a detailed description of the parameter settings for our entire framework. The following is presented in pseudocode form:

- **Parameter Initialization**
  - `num_feat = 16`
- **Flow Estimator**
  - `self.spynet = Get_RAFT()`
- **Bidirectional Propagation Restoration Network (with DMRB)**
  - `self.forward_resblocks = Backbone(`
    `embed_dim=64, depths=[2, 2, 2], d_state=16,`
    `mlp_ratio=1.2, in_chans=num_feat, drop_rate=0., weight_group_size=16)`
  - `self.backward_resblocks = Backbone(`
    `embed_dim=64, depths=[2, 2, 2], d_state=16,`
    `mlp_ratio=1.2, in_chans=num_feat, drop_rate=0., weight_group_size=16)`
    - `--VMB(`
      `dim=embed_dim, depth=depths[ith_layer], d_state=d_state,`
      `mlp_ratio=self.mlp_ratio, drop_path=dpr[`
      `sum(depths[:ith_layer]):sum(depths[:ith_layer + 1])])`
    - `--DPB(weight_group_size=16, kernel_size=3, padding=1)`
- **Temporal Feature Alignment**
  - `self.Forward_Aggregation = FMDA(hidden_channels=num_feat, kernel_size=3, padding=1)`
  - `self.Backward_Aggregation = FMDA(hidden_channels=num_feat, kernel_size=3, padding=1)`
- **Pixel-Shuffle Upsampling**
  - `self.up1 = PSUpsample(num_feat*4, num_feat, scale_factor=1)`
  - `self.up2 = PSUpsample(num_feat, num_feat, scale_factor=1)`

- **Tail Layers**
  - `self.conv_hr = nn.Conv2d(num_feat, num_feat, kernel_size=3, stride=1, padding=1)`
  - `self.conv_last = nn.Conv2d(num_feat, 3, kernel_size=3, stride=1, padding=1)`
- **Global Residual Learning**
  - `self.img_up = nn.Upsample(scale_factor=1, mode='bilinear', align_corners=False)`
- **Activation Function**
  - `self.lrelu = nn.LeakyReLU(negative_slope=0.1, inplace=True)`
- **Computation of Reconstruction Layers**
  - `cat_feat = torch.cat([rlt[i], feat_prop], dim=1)`
  - `lq_feat = self.lrelu(self.concate(cat_feat))`
  - `lq_feat = self.lrelu(self.up1(lq_feat))`
  - `lq_feat = self.lrelu(self.up2(lq_feat))`
  - `lq_feat = self.lrelu(self.conv_hr(lq_feat))`
  - `lq_feat = self.conv_last(lq_feat)`
  - `base = self.img_up(lq_feat)`
  - `lq_feat += base`
  - `hq_img = torch.tanh(lq_feat)`

6. User Study

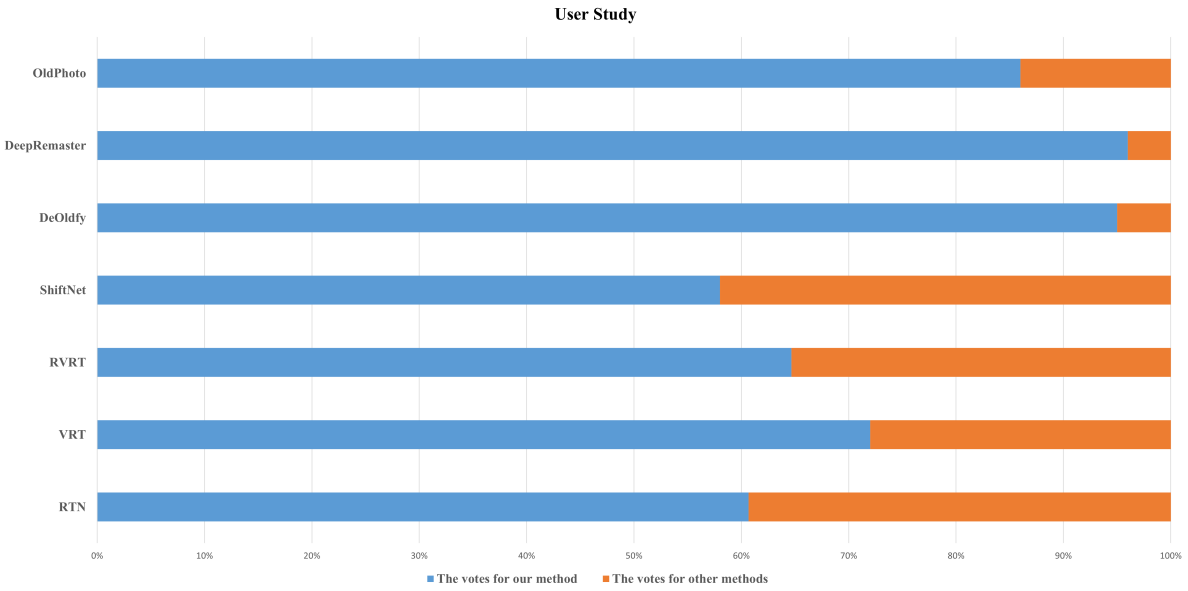

Figure 3. User Study. Voting statistics of different methods versus our method, including DeOldify[1], OldPhoto[9], DeepRemaster[4], RTN[10], VRT[8], RVRT[7], ShiftNet[6]

In real-world scenarios, where ground truth is unavailable, restoration performance can only be assessed using no-reference metrics. However, comparisons based on different evaluation metrics often yield inconsistent results. To alleviate this, we conducted a user study to compare the subjective quality of restoration outcomes.

For this study, we randomly selected 30 old film clips from our proposed dataset and performed pairwise comparisons of the restoration results produced by our method and the comparison methods. Participants were asked to vote for the result they considered to be of higher quality. A total of 15 participants were recruited for the study. The average of the 15 sets of votes was calculated, and the final statistical results are presented in Fig. 3. As shown in Fig. 3, our method demonstrates significant advantages in subjective quality compared to the comparison methods.

7. Extend Comparison Results

To further validate the effectiveness of our method, we present additional qualitative comparisons. Specifically, we provide restoration results for a sequence of continuous video frames to assess the temporal stability of our approach. As illustrated in Fig. 5, our method outperforms the comparison methods in terms of overall visual quality, including improvements in color accuracy, brightness restoration, and detail retrieval, as well as in addressing specific structural defects. Furthermore, as depicted in Fig. 4, the restoration results on real-world old films similarly underscore the effectiveness of our method.

Figure 4. Visual examples of restoration results on the **real-world** part of the proposed datasets, including DeOldify[1], OldPhoto[9], DeepRemaster[4], RTN[10], VRT[8], RVRT[7], ShiftNet[6]. **Zooming in for better comparison.**

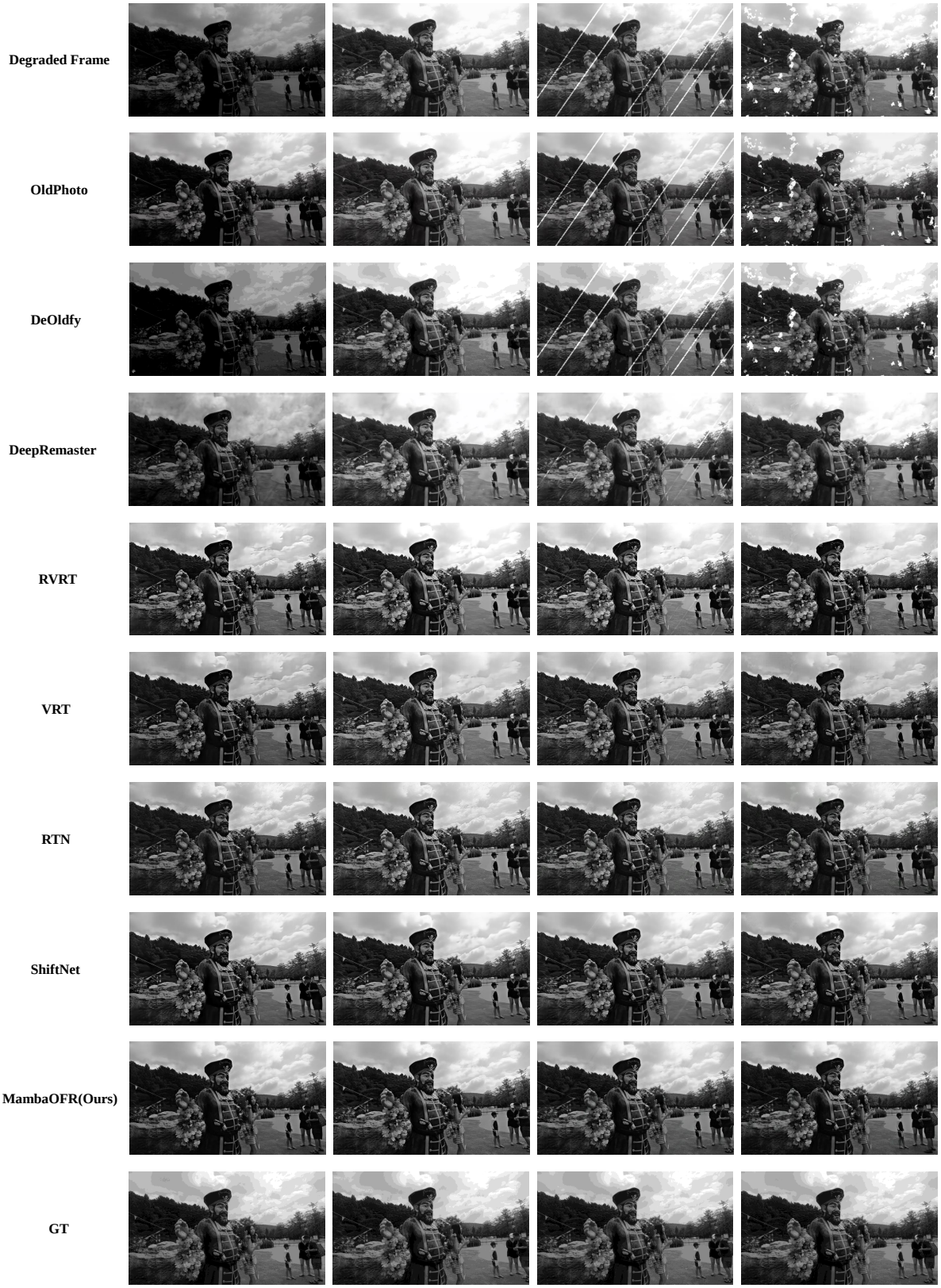

Figure 5. Visual examples of restoration results on the **synthetic** part of the proposed datasets, including DeOldify[1], OldPhoto[9], DeepRemaster[4], RTN[10], VRT[8], RVRT[7], ShiftNet[6]. **Zooming in for better comparison.**

References

[1] Deoldify, https://github.com/jantic/deoldify. 4, 5, 6

[2] Ya-Liang Chang, Zhe Yu Liu, Kuan-Ying Lee, and Winston Hsu. Free-form video inpainting with 3d gated convolution and temporal patchgan. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9066–9075, 2019. 1

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1

[4] Satoshi Iizuka and Edgar Simo-Serra. Deepremaster: temporal source-reference attention networks for comprehensive video enhancement. *ACM Transactions on Graphics (TOG)*, 38(6):1–13, 2019. 2, 4, 5, 6

[5] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 1

[6] Dasong Li, Xiaoyu Shi, Yi Zhang, Ka Chun Cheung, Simon See, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. A simple baseline for video restoration with grouped spatial-temporal shift. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9822–9832, 2023. 1, 4, 5, 6

[7] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc V Gool. Recurrent video restoration transformer with guided deformable attention. *Advances in Neural Information Processing Systems*, 35:378–393, 2022. 4, 5, 6

[8] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *IEEE Transactions on Image Processing*, 33:2171–2182, 2024. 4, 5, 6

[9] 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. 1, 2, 4, 5, 6

[10] Ziyu Wan, Bo Zhang, Dongdong Chen, and Jing Liao. Bringing old films back to life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17694–17703, 2022. 2, 4, 5, 6