Restoring Degraded Old Films with Recursive Recurrent Transformer Networks

Shan Lin
Waseda University
mountlin@fuji.waseda.jp

Edgar Simo-Serra
Waseda University
ess@waseda.jp

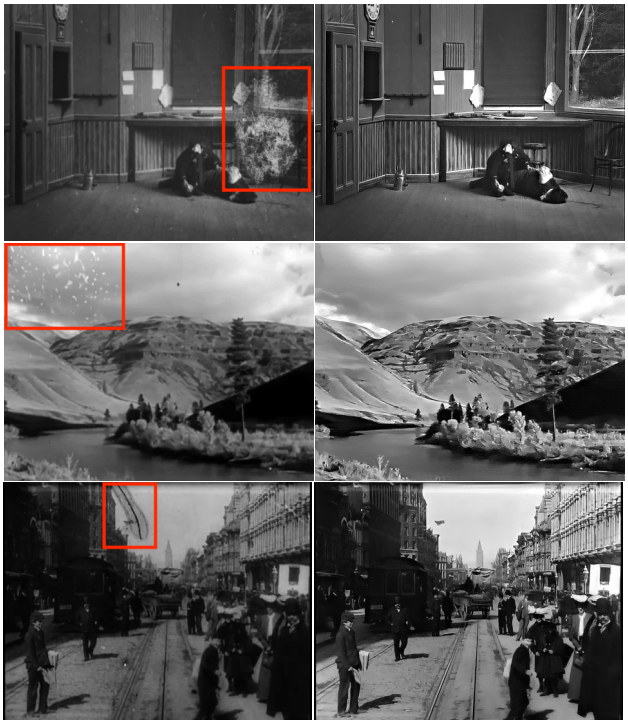
Abstract

There exists a large number of old films that have not only artistic value but also historical significance. However, due to the degradation of analogue medium over time, old films often suffer from various deteriorations that make it difficult to restore them with existing approaches. In this work, we proposed a novel framework called Recursive Recurrent Transformer Network (RRTN) which is specifically designed for restoring degraded old films. Our approach introduces several key advancements, including a more accurate film noise mask estimation method, the utilization of second-order grid propagation and flow-guided deformable alignment, and the incorporation of a recursive structure to further improve the removal of challenging film noise. Through qualitative and quantitative evaluations, our approach demonstrates superior performance compared to existing approaches, effectively improving the restoration for difficult film noises that cannot be perfectly handled by existing approaches. The code and model are available at https://github.com/mountln/RRTN-old-film-restoration.

1. Introduction

Due to the degradation of the analog medium over time, old films often suffer from various deteriorations. Many films with historical and artistic value are gradually forgotten by the public because of the bad image qualities. With the advancements in image processing technologies, efforts have been made to restore these degraded films. The restoration process involves physically removing dust and stains from the film, followed by scanning it into digital format and digitally restoring it using computers.

Digital restoration is a time-consuming and expensive task that requires manual frame-by-frame restoration by a team of experienced experts. Consequently, only a select few well-known works were chosen for restoration, leaving behind a large number of films that remained untouched due to limited resources. As a result, most of the old films available online suffer from significant degradation. In order to be able to restore these old films to their original appearance and show them in the best condition, an automatic restoration approach is essential. In recent years, the development of deep learning has made it possible.



(a) Input (b) Restoration Results

Figure 1. **Examples of restoration results achieved using our proposed approach.** Our purposed approach effectively removes challenging noise and significantly improves the overall image quality of frames. Important noise in the input is marked with red rectangles.

Despite the advancements in existing methods, there still exists no universal solution that can effectively remove most of the harder noise well. In order to efficiently remove complex noises, we propose a new framework called Recursive Recurrent Transformer Network (RRTN), which is based on a more complex RNN architecture with Transformers. Within RRTN, we have developed an accurate approach for estimating the film noise mask. We used both the previous and next frames of the frame being estimated, and used

the difference between them to estimate a film noise mask for guiding the restoration. Additionally, inspired by BasicVSR++ [3], we employ second-order grid propagation and flow-guided deformable alignment to efficiently leverage the information contained within the different frames of an old film. Furthermore, our model utilizes a recursive structure to efficiently handle complex noises in old films with varying degrees of degradation. When the video degradation is severe, our model performs more steps of recursion to remove the challenging film noise.

To assess the effectiveness of our approach, we conducted both quantitative and qualitative evaluations. The quantitative results show that the output of our approach has better performance than existing approaches in terms of overall frame quality. The qualitative result shows that our model can better remove some difficult film noise that cannot be removed by existing approaches. Furthermore, we performed an ablation study, verified the significance of each component in our method.

Our contributions can be summarized as follows:

- Explicit modelling of film noise mask with second-order grid propagation and flow-guided deformable alignment that allows removal of large film artefacts.
- Adaptive recursive architecture that encourages temporal coherence of the output, reducing flickering and other common film damage.
- In-depth comparisons with existing approaches that demonstrate the effectiveness of our approach.

2. Related Work

2.1. Video Restoration

Video restoration is a common task aimed at restoring low-quality videos to a higher quality, including aspects such as noise removal, sharpening, and inpainting. Since videos can have various degradation, many video processing tasks can be considered as subtasks of video restoration. Extensive research has been conducted in these subtasks, yielding various research results. For instance, video super-resolution [2, 3, 17, 40, 43] focuses on enhancing the spatial resolution of videos. Video denoising [31, 44, 47] addresses the removal of noise from videos. Video deblurring [36, 42, 58] aims to reduce blur and improve video sharpness. Video colorization [22, 51, 53] involves adding color to grayscale videos. Video inpainting [19, 28, 52] focuses on filling in missing regions in videos. Due to the similarity of these subtasks, which are all video-to-video transformations, there are also some works like [4, 23, 24, 49] that can restore different degradation with a single network architecture by training models using different data.

Unlike image restoration [6, 7, 25, 54–56], video restoration benefits from not only the spatial information within individual frames but also the temporal information across multiple frames. As a result, it is important to use information in different frames effectively. Consequently, propagation, alignment, and fusion act as essential steps in the video restoration process.

Propagation plays a key role in transferring information between frames in video restoration. While early approaches [1, 9, 14] primarily used temporal CNNs, were limited in capturing long-term information. To tackle this problem, the utilization of RNN structures has gained significant popularity in video restoration [10, 11, 15, 16, 35]. In addition to traditional one-way RNNs, there is a increasing adoption of bidirectional propagation approaches [2, 12, 13]. Furthermore, Chan *et al*. [3] introduced a higher-order grid propagation approach to further gather different levels of features.

Alignment is used to align the different features obtained by propagation. Compared to the method without alignment [10, 13, 15, 16], optical flow warping based approaches [2, 20, 39, 50] using the guide of the optical flow to warp the frames or features for alignment. In addition to optical flow warping based approaches, other approaches such as deformable convolution [3, 24, 27, 46, 49] and deformable attention [26] have also been utilized and shown to deliver strong performance in video restoration tasks.

Fusion aims to restore aligned features and integrate features. For fusion step, a common and lightweight option is to use convolutional layers [2, 3, 10, 13, 15]. Wang *et al*. [49] incorporates attention mechanism with convolution layers achieved good results. More recently, with the proposal vision Transformers [8, 29], Swin Transformer-based approach has been adopted by some works [26, 30, 48], which greatly improves the image quality of the final video.

2.2. Old Film Restoration

Old film restoration is also a type of video restoration that deals with a variety of degradation, including scratches, dust adhesion and fading. Unlike other video restoration tasks, old films may contain multiple types of degradation at the same time. The purpose of old film restoration is to remove these degradation and restore the film to its original state when it was completed. In the past, old films usually required a large number of experts to perform frame by frame manual restoration. However, recent advancements in deep learning based approaches made it possible to automate the restoration of old films.

DeepRemaster [14] is the first framework using a temporal CNN to restore old films. To generate training data, they simulated the degradation in old films by blending real film noise footage with high-quality videos and subsequently applying algorithms to introduce further degradation. Through this approach, they successfully trained a model that can be used for old film restoration.

Wan *et al*. [48] introduced a network based on bidirec-

tional RNN with Swin Transformer [29] for old film restoration. They incorporated perceptual loss [18] and adversarial loss to training loss function, resulting in significant enhancements in the overall final video quality.

Despite the notable progress achieved by existing methods, the varying levels of degradation present in old films pose a challenge. While the existing approaches can effectively handle mildly degraded films, there remains a lot of films with complex film noise that cannot be entirely removed using these approaches.

3. Approach

An overview of our proposed approach can be seen in Fig. 2. Our approach takes a sequence of degraded frames $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\mathrm{T}\}$ as the input and generates a sequence of restored frames $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_\mathrm{T}\}$ as the output, where $x_t^{ij}, y_t^{ij} \in [0.0, 1.0]$. The input frames are first used to estimate the film noise mask, then concatenated with the mask and fed into an encoder consisting of convolutional layers for feature extraction. The extracted features are subsequently passed through Alignment and Transformer Blocks in a second-order grid propagation strategy for alignment and feature restoration. Then, the features are decoded and added to the input to obtain primary output frames. If the output frames do not meet the predefined conditions, they will be processed again as input recursively.

3.1. Film Noise Mask Estimator

The details of film noise mask estimator can be seen in Fig. 3. In the analogue medium, each frame is entirely independent, resulting in film noise caused by contaminants such as dust attached to them being different in each frame. Exploiting this characteristic of degraded film, we developed a noise estimator to guide the restoration process. In [48], Wan *et al*. also tried a estimation method for masks, but they only use a single adjacent frame to compute the mask. The mask estimation often fails due to errors in optical flow estimation. Besides, if the reference frame itself contains noise, that would lead to an inaccurate mask. Therefore, we use both the previous frame $\mathbf{x}_{t-1}$ and the next frame $\mathbf{x}_{t+1}$ for estimating the t-th film noise mask $\mathbf{m}_t$ to improve the accuracy of the mask. We first warp the adjacent frames $\mathbf{x}_{t-1}$, $\mathbf{x}_{t+1}$ to the current frame $\mathbf{x}_t$ guided by optical flows, and calculate their variations $\mathbf{v}_{t-1,t}$, $\mathbf{v}_{t+1,t}$. Then, we calculate the element-wise geometric mean of the absolute variations $\mathbf{v}_{t-1,t}$, $\mathbf{v}_{t+1,t}$ to represent the film noise mask $\mathbf{m}_t$. The process can be represented by the following Eqs. (1) to (3).

$$\mathbf{v}_{t-1,t} = \mathcal{W}(\mathbf{x}_{t-1}, \mathbf{o}_{t-1\to t}) - \mathbf{x}_t \tag{1}$$

$$\mathbf{v}_{t+1,t} = \mathcal{W}(\mathbf{x}_{t+1}, \mathbf{o}_{t+1\to t}) - \mathbf{x}_t \tag{2}$$

$$\mathbf{m}_t = \left(\sqrt{|v_{t-1,t}^{ij}| \cdot |v_{t+1,t}^{ij}|}\right) \tag{3}$$

where $\mathbf{o}_{t-1\to t}$, $\mathbf{o}_{t+1\to t}$ denote the optical flow from $(t-1)$-th and $(t+1)$-th frames to t-th frame, $\mathcal{W}$ denotes the spatial warping operation. And to represent element-wise operations more clearly, tensor $\mathbf{v}$ is represented as $\left(v^{ij}\right)$ in Eq. (3).

3.2. Feature Propagation, Alignment and Restoration

Feature propagation plays an important role in the transmission of information between features. Wan *et al*. used bi-directional propagation in [48], this method propagates forward and backward once in each temporal direction, and the information in the adjacent features in both temporal directions can be utilized by the current feature, which performs better than the one-directional propagation of traditional RNN.

As shown by the orange arrows in Fig. 2, our RRTN uses grid propagation, which propagates forwards and backwards multiple times in temporal directions, which allows the previous and next features in both temporal direction can be better utilized than the single time bi-directional propagation. In addition to adjacent frames, we use second-order propagation, which uses more distant features to directly obtain information over a larger temporal range. This propagation method is shown in Fig. 2 with blue dashed arrows.

Through second-order grid propagation, features are passed between alignment and transformer blocks. Each alignment and transformer block consists of flow-guided deformable alignment module and Swin Transformer [29]. The purpose of these blocks are to align the passed features and to perform spatial restoration on the aligned features.

The details of the enlarged block in Fig. 2 can be represented by the following Eq. (4).

$$\mathbf{f}_{j+1}^{t+1} = \mathcal{R}\left(\mathbf{f}_j^{t+1} \frown \mathcal{A}\left(\mathbf{f}_{j+1}^{t-1}, \mathbf{f}_{j+1}^{t}, \mathbf{o}_{t-1\to t+1}, \mathbf{o}_{t\to t+1}\right)\right) \tag{4}$$

where $\mathbf{f}$s denote features, $\mathcal{R}$ denotes restoration operation by Swin Transformer, $\mathcal{A}$ denotes flow-guided deformable alignment and $\frown$ denotes the concatenation operation along the channel dimension.

3.3. Recursive Architecture

In Fig. 2, the recursive structure is represented by red dashed arrows. The recursive structure leverages another characteristic of old films, which is that the level of degradation varies greatly from one old film to another. Some well-preserved old films may have only a little noise, whereas some old films that have been poorly maintained may suffer from severe degradation. Using recursive structure to process old films with different levels of degradation for different times of recursions is highly beneficial in removing the film noise that is difficult to handle.

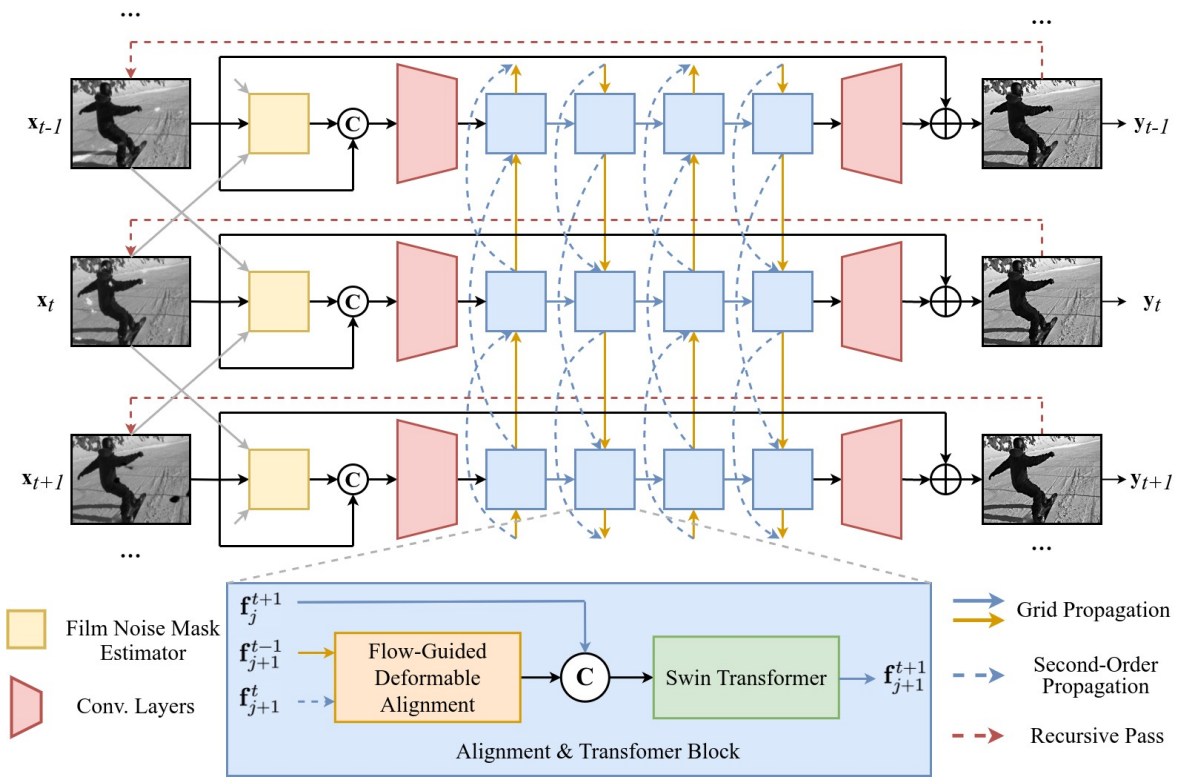


Figure 2. **Overview of proposed approach.** By using an RNN architecture with an explicit film noise mask estimator and second order propagation, our model is able to more accurately represent and remove the independent frame noise.

We trained the model twice to get two sets of parameters, denoted as θ_1 and θ_2. During the inference stage, we utilize these two sets of parameters for different recursion steps. θ_1 is used in first recursion, and is obtained by setting the number of recursive steps to 1 during training. Using θ_1 during inference allows for more aggressive frame processing, resulting in improved image quality. On the other hand, θ_2 is used in subsequent recursion, and is obtained by setting the number of recursive steps to 2 during training. Using θ_2 during inference leads to a more gentle frame processing which keeps the overall image appearance and specifically targeting the imperfectly processed film noise from the last recursion step. It is important to note that using θ_1 alone for all recursion steps would result in unnatural final output frames due to over-processing. In contrast, if we use θ_2 alone for all recursion steps, although the model does well in film noise removal and does not make the output frames look unnatural due to multiple times of processing, it has a poor performance on super resolution and deblurring. The final output frames look blurrier than the output frames obtained by θ_1.

Our method determines whether to stop the recursion based on the mean square error (MSE) of the input and the output data in the current recursion step. When the MSE of the current input and output is less than the threshold $\epsilon = 10^{-4}$, it means that the model has processed the data very little, at which point we stop the recursion. The value of ϵ can be determined based on the amount of film noise in the actual output frames and the number of recursions when process these frames. If ϵ is too large, the number of recursions will be small and the model will not be able to remove difficult film noise. Conversely, if ϵ is too small, the number of recursions will be too large and the inference time will be greatly increased. In practice, an better choice of ϵ results in a recursion number of 2 for frames with less film noise, and 3 for those with more difficult film noise.

4. Experiments and Results

To compare our approach with the state-of-the-art approaches, we trained models for each approach and conducted both quantitative and qualitative evaluations.

4.1. Training

4.1.1 Data

The training data is generated based on the REDS [34] dataset. In each iteration of training, we select 7 consecutive frames from the REDS dataset as the initial values of ground truth $\mathbf{y}$ and input video $\mathbf{x}$. Then the transformations listed in Tab. 1 are performed for data augmentation and

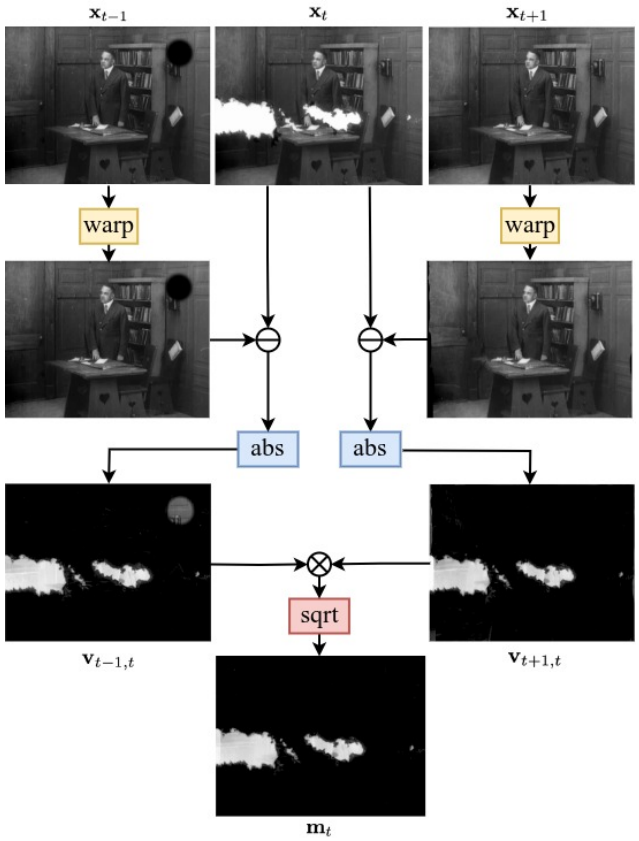


Figure 3. **Film noise mask estimator.** We warp adjacent frames to a reference image and compute the difference with the geometric mean to estimate the noise mask and allow more efficient noise removal.

(a) Example 1 (b) Example 2

Figure 4. **Two pair of generated frames using transformations in Tab. 1.** In each example, the left frame is input frame and the right frame is the corresponding GT.

degraded video synthesis.

In Tab. 1, when target is $(\mathbf{x}, \mathbf{y})$, it means the operation is performed on both $\mathbf{x}$ and $\mathbf{y}$ at the same time. At first, we convert the color image to grayscale and perform scaling and random crop operations to obtain a set of frames of size 128×128. Then, the data augmentation is further performed by random flip and rotation operations. The operations with target $\mathbf{x}$ are performed only on the input data $\mathbf{x}$, and is intended to degrade the data to simulate the degraded old film. When performing film noise blending, we used the footage provided in DeepRemaster [14]. A final generated data example is shown in Fig. 4.

Name	Target	Prob.	Parameters
Grayscale	$(\mathbf{x}, \mathbf{y})$	100%	-
Scaling	$(\mathbf{x}, \mathbf{y})$	100%	$h : \mathcal{U}(128, 720)$
Random Crop	$(\mathbf{x}, \mathbf{y})$	100%	$w : 128, h : 128$
Vertical Flip	$(\mathbf{x}, \mathbf{y})$	50%	-
Horizontal Flip	$(\mathbf{x}, \mathbf{y})$	50%	-
Rotation	$(\mathbf{x}, \mathbf{y})$	50%	$\pm 90°$
Film Noise Blending	$\mathbf{x}$	100%	$\alpha : \mathcal{U}(0.6, 1.0)$
Brightness	$\mathbf{x}$	50%	$\mathcal{U}(0.8, 1.2)$
Contrast	$\mathbf{x}$	50%	$\mathcal{U}(0.9, 1.0)$
Gaussian Blur	$\mathbf{x}$	100%	$\sigma : \mathcal{U}(0.0, 1.0)$
Gaussian Noise	$\mathbf{x}$	50%	$\sigma : \mathcal{U}(0.0, 0.04)$
Speckle Noise	$\mathbf{x}$	50%	$\sigma : \mathcal{U}(0.0, 0.04)$
Downsampling	$\mathbf{x}$	100%	$h : \mathcal{U}(0.25, 1.0)$
JPEG	$\mathbf{x}$	100%	$q : \mathcal{U}(40, 100)$

Table 1. **Transformations for data generation.** The "Parameters" column specifies the parameters used for each transformation operation. $\mathcal{U}(a, b)$ denotes a value which is obtained by sampling from a uniform distribution over the closed interval $[a, b]$.

4.1.2 Objective Function

We use Eq. (5) to compute loss during training.

$$\mathcal{L} = \lambda_c \mathcal{L}_c + \lambda_{perc} \mathcal{L}_{perc} + \lambda_{adv} \mathcal{L}_{adv} \tag{5}$$

where $\lambda_c, \lambda_{perc}, \lambda_{adv}$ represent the respective weights of each loss term. We set the weights as: $\lambda_c = 1$, $\lambda_{prec} = 1$, and $\lambda_{adv} = 0.01$. The loss terms included are as follows: $\mathcal{L}_c$ refers to Charbonnier loss [5]. $\mathcal{L}_{perc}$ denotes perceptual loss [18]. $\mathcal{L}_{adv}$ represents spatial-temporal adversarial loss which is used in [48]. For feature extraction in the perceptual loss, we employed a pretrained VGG19 [41] model, utilized the features extracted after layers relu2_2 to relu5_2 of the VGG19 model.

4.1.3 Training Details

We trained all models with the same setting. We utilized the Adam optimizer [21] with a learning rate of 2e-4 for the initial 100,000 iterations, and linearly decayed the learning rate after 100,000 iterations.

For optical flow estimation, according to the original implementation, we used SPyNet [38] for training RVRT [26] and BasicVSR++ [3], used RAFT [45] for training the model proposed by Wan *et al*. [48]. For training our own model, we used RAFT [45] for optical flow estimation.

The models with the lowest validation loss were selected as the final models after 200,000 training iterations.

4.2. Quantitative Evaluation

For the quantitative evaluation, we conducted comparisons on both the synthesized data and real old film data.

	PSNR↑	SSIM↑	LPIPS↓
Input	19.3584	0.6864	0.4389
DeepRemaster [14]	20.9723	0.7156	0.3606
BasicVSR++ [3]	22.0617	0.7626	0.3610
RVRT [26]	22.1256	0.7404	0.3545
Wan *et al*. [48]	21.8243	0.7732	0.3636
Wan *et al*. [48]†	22.9341	0.7774	0.3487
Our method	**23.1208**	**0.7879**	**0.3185**

Table 2. **Quantitative comparison on synthetic video dataset.** Best results are highlighted in **bold**. † denotes using a pre-trained model trained with 256 × 256 frames.

	BRISQUE↓	NIQE↓
Input	46.2065	17.6175
DeepRemaster [14]	39.1887	17.5010
BasicVSR++ [3]	25.8349	17.4555
RVRT [26]	33.1265	17.0986
Wan *et al*. [48]	24.1537	17.1495
Wan *et al*. [48]†	17.3608	17.3182
Our method	**15.6102**	**16.6216**

Table 3. **Quantitative comparison on real old film dataset.** Best results are highlighted in **bold**. † denotes using a pre-trained model trained with 256 × 256 frames.

We used the same transformations in Tab. 1 generate test synthesized data based on DAVIS [37] dataset, and collected some real old films from the Internet.

On the synthesized data, we employed PSNR, SSIM, and LPIPS [57] as evaluation metrics. On the real old film data, we employed two no-reference metrics, BRISQUE [32] and NIQE [33] as evaluation metrics, since we cannot obtain the ground truth data of the real films.

The quantitative evaluation results are shown in Tabs. 2 and 3, with the best results highlighted in bold. The results obtained using the pre-trained model provided by Wan *et al*. [48] were also included in the tables. Note that this model was trained under different conditions than the other models. It can be seen that our method has significantly better results than the existing methods for both the synthesized data and the real old film data.

4.3. Qualitative Evaluation

For the qualitative evaluation, we used real old films to assess the performance of our method. As a result, our method shows its effectiveness in removing those noises that cannot be removed by other methods. In Fig. 5, we selected 4 frames with hard-to-remove noise to show the results of each method. In the variation image on the right column, the film noise that needs to be removed is highlighted by red box. As we can see, our RRTN still performs well for the noise that cannot be perfectly removed by BasicVSR++ [3] and the method Wan *et al*. proposed [48].

	Parameters (M)	Runtime (ms)
DeepRemaster [14]	9.9	43
BasicVSR++ [3]	5.8	272
RVRT [26]	8.5	402
Wan *et al*. [48]	6.2	267
Ours(recursion=1)	7.5	421
Ours(recursion=2)	7.5	876
Ours(recursion=3)	7.5	1322

Table 4. **Comparison of model size and inference time.** Inference time was measured using an RTX 2080Ti GPU with a frame size of 640 × 368.

	(A)	(B)	(C)	(D)	Ours
noise mask		✓			✓
prop. & align.			✓		✓
recursion				✓	✓
PSNR↑	21.82	22.16	22.70	21.77	**23.12**
SSIM↑	0.773	0.771	0.785	0.757	**0.788**
LPIPS↓	0.364	0.334	0.327	0.358	**0.319**
BRISQUE↓	24.15	21.65	23.73	30.32	**15.61**
NIQE↓	17.15	16.86	17.10	17.08	**16.62**

Table 5. **Ablation study of the components.** Best results are highlighted in **bold**.

4.4. Comparison of Model Size and Inference Time

We measured the inference time of each approach. The results, along with the size of each model, are shown in Tab. 4. Our approach takes more inference time than other approaches. Additionally, as the number of recursions increases, the inference time increases linearly.

Although using recursion significantly increases inference time, in cases where multiple recursions are required, since using other approaches cannot remove difficult film noise perfectly, the time spent on multiple recursions is acceptable compared to having an expert restore it manually which would take far more time.

5. Ablation Study

In order to evaluate the significance of each component in our method, we conducted an ablation study. The results are presented in Tab. 5 and Fig. 6. We selected Wan *et al*.'s method [48] as the baseline, denoted as (A). Method (B∼D) integrate the baseline with each corresponding components. *Ours* refers to our method that combines all the compo-

Figure 5. **Comparison on real films.** Our approach is able to remove challenging large noise and significantly improve the original degraded film. Important noise in the input is marked with red rectangles.

nents. For each component, *noise mask* represents the film noise mask, *prop. & align.* stands for second-order grid propagation and flow-guided deformable alignment, and *recursion* refers to the recursive structure.

From columns (B) and (C) of Tab. 5, it can be seen that the incorporation of film noise mask or prop. & align. enhances the overall frame quality of the final video relative to the baseline. However, the results in column (D) indicate

Input (A) baseline (B) noise mask (C) prop. & align. (D) recursion Ours

Figure 6. **Ablation study on real films.** Important noise in the input is marked with red rectangles. Compared to the baseline (A), integrating any component (B∼D) contributes to film noise reduction. With all the components (Ours), the best output frame is obtained.

that the use of recursion diminishes the quality. This decline can likely be attributed to excessive recursive calls, which magnify the inherent shortcomings of the under-performing model, thus deteriorating the results. Yet, when recursion is combined with the other two components, recursion plays a very important role in dealing with noise that is difficult to remove.

The results in Fig. 6 further show that the addition of any component aids in reduce film noise. However, for challenging inputs as seen in the first and second rows, slight residues of film noise remain. These residues are not obvious in the frame, but can be clearly seen in the video due to the interruption of the consistency of consecutive frames. The best results are achieved by combining all the components.

6. Conclusion

In this paper, we proposed a novel framework, called Recursive Recurrent Transformer Network, for degraded film restoration. We obtain better restoration performance than the existing methods by use of a more accurate film mask estimator, a more efficient feature propagation, alignment and spatial restoration strategy, and a recursive structure that can handle more difficult film noise.

Even though RRTN offers the possibility to restore old films that cannot be perfectly restored before, the network uses more computation time than existing approaches due to require additional operations. Although more complex film noises can be handled by recursive processing, the computation time required increases as the number of recursions increases, which leads to more time spent on restoration for degraded films with more noise. To solve this problem, in future research, we believe we can improve the processing time by reducing the number of parameters appropriately.

Input Output

Figure 7. **An example of failure case.** Some of the smoke was mistakenly removed as noise. Important difference is marked with a red rectangle.

Additionally, there are failure cases where non-noise is mistakenly removed as noise, which can lead to undesired alterations in the original content or result in over-smoothed textures. As illustrated in Fig. 7, some of the smoke from the input frame was wrongly removed as noise, leading to an output frame with noticeably less smoke than the original. Thus, reducing noise misclassification is also an important future research direction.

References

[1] Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4778–4787, 2017. 2

[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. 2

[3] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5972–5981, 2022. 2, 5, 6, 7

[4] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. On the generalization of basicvsr++ to video deblurring and denoising. *arXiv preprint arXiv:2204.05308*, 2022. 2

[5] Pierre Charbonnier, Laure Blanc-Feraud, Gilles Aubert, and Michel Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In *Proceedings of 1st international conference on image processing*, volume 2, pages 168–172. IEEE, 1994. 5

[6] Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen. Hinet: Half instance normalization network for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 182–192, 2021. 2

[7] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, pages 184–199. Springer, 2014. 2

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[9] Yuchen Fan, Jiahui Yu, Ding Liu, and Thomas S Huang. An empirical investigation of efficient spatio-temporal modeling in video restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 2

[10] Dario Fuoli, Shuhang Gu, and Radu Timofte. Efficient video super-resolution through recurrent latent space propagation. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3476–3485. IEEE, 2019. 2

[11] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3897–3906, 2019. 2

[12] Yan Huang, Wei Wang, and Liang Wang. Bidirectional recurrent convolutional networks for multi-frame super-resolution. *Advances in neural information processing systems*, 28, 2015. 2

[13] Yan Huang, Wei Wang, and Liang Wang. Video super-resolution via bidirectional recurrent convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):1015–1028, 2017. 2

[14] 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, 5, 6, 7

[15] Takashi Isobe, Xu Jia, Shuhang Gu, Songjiang Li, Shengjin Wang, and Qi Tian. Video super-resolution with recurrent structure-detail network. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 645–660. Springer, 2020. 2

[16] Takashi Isobe, Fang Zhu, Xu Jia, and Shengjin Wang. Revisiting temporal modeling for video super-resolution. *arXiv preprint arXiv:2008.05765*, 2020. 2

[17] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3224–3232, 2018. 2

[18] 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. 3, 5

[19] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Deep video inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5792–5801, 2019. 2

[20] Tae Hyun Kim, Mehdi SM Sajjadi, Michael Hirsch, and Bernhard Scholkopf. Spatio-temporal transformer network for video restoration. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 106–122, 2018. 2

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[22] Chenyang Lei and Qifeng Chen. Fully automatic video colorization with self-regularization and diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3753–3761, 2019. 2

[23] 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*, pages 9822–9832, 2023. 2

[24] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool.

Vrt: A video restoration transformer. *arXiv preprint arXiv:2201.12288*, 2022. 2

[25] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 2

[26] 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. 2, 5, 6, 7

[27] Jiayi Lin, Yan Huang, and Liang Wang. Fdan: Flow-guided deformable alignment network for video super-resolution. *arXiv preprint arXiv:2105.05640*, 2021. 2

[28] Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fuseformer: Fusing fine-grained information in transformers for video inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14040–14049, 2021. 2

[29] 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*, pages 10012–10022, 2021. 2, 3

[30] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 2

[31] Mona Mahmoudi and Guillermo Sapiro. Fast image and video denoising via nonlocal means of similar neighborhoods. *IEEE signal processing letters*, 12(12):839–842, 2005. 2

[32] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012. 6

[33] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 6

[34] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 4

[35] Seungjun Nah, Sanghyun Son, and Kyoung Mu Lee. Recurrent neural networks with intra-frame iterations for video deblurring. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8102–8111, 2019. 2

[36] Dongwon Park, Dong Un Kang, and Se Young Chun. Blur more to deblur better: Multi-blur2deblur for efficient video deblurring. *arXiv preprint arXiv:2012.12507*, 2020. 2

[37] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 6

[38] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4161–4170, 2017. 5

[39] Mehdi SM Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6626–6634, 2018. 2

[40] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 2

[41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[42] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1279–1288, 2017. 2

[43] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *Proceedings of the IEEE international conference on computer vision*, pages 4472–4480, 2017. 2

[44] Matias Tassano, Julie Delon, and Thomas Veit. Dvdnet: A fast network for deep video denoising. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1805–1809. IEEE, 2019. 2

[45] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 5

[46] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3360–3369, 2020. 2

[47] Gregory Vaksman, Michael Elad, and Peyman Milanfar. Patch craft: Video denoising by deep modeling and patch matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2157–2166, 2021. 2

[48] 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, 3, 5, 6, 7

[49] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2

[50] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented

flow. *International Journal of Computer Vision*, 127:1106–1125, 2019. 2

[51] Liron Yatziv and Guillermo Sapiro. Fast image and video colorization using chrominance blending. *IEEE transactions on image processing*, 15(5):1120–1129, 2006. 2

[52] Yanhong Zeng, Jianlong Fu, and Hongyang Chao. Learning joint spatial-temporal transformations for video inpainting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 528–543. Springer, 2020. 2

[53] Bo Zhang, Mingming He, Jing Liao, Pedro V Sander, Lu Yuan, Amine Bermak, and Dong Chen. Deep exemplar-based video colorization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8052–8061, 2019. 2

[54] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155, 2017. 2

[55] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3262–3271, 2018. 2

[56] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 649–666. Springer, 2016. 2

[57] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[58] Chao Zhu, Hang Dong, Jinshan Pan, Boyang Liang, Yuhao Huang, Lean Fu, and Fei Wang. Deep recurrent neural network with multi-scale bi-directional propagation for video deblurring. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 3598–3607, 2022. 2