Prompt-Based Video Frame Restoration

Abstract

Image restoration (IR) is a critical task in computer vision that aims to restore high-quality images from degraded input. PromptIR is a prompt-based multitasking image restoration model capable of enhancing various types of degraded images simultaneously. In this work, we extend PromptIR's capabilities to video frame restoration by introducing CRF compression, a common video degradation kernel, and fine-tuning the model on new datasets to validate its generalization ability. Additionally, we adjust the model's configuration and apply mixed-precision training to reduce memory overhead while maintaining high accuracy. Our experiments demonstrate that the enhanced PromptIR model effectively restores compressed video frames and achieves a balance between performance and resource efficiency, making it suitable for deployment on resource-constrained devices.

Keywords: Computer Vision, Image Restoration, Video Compression.

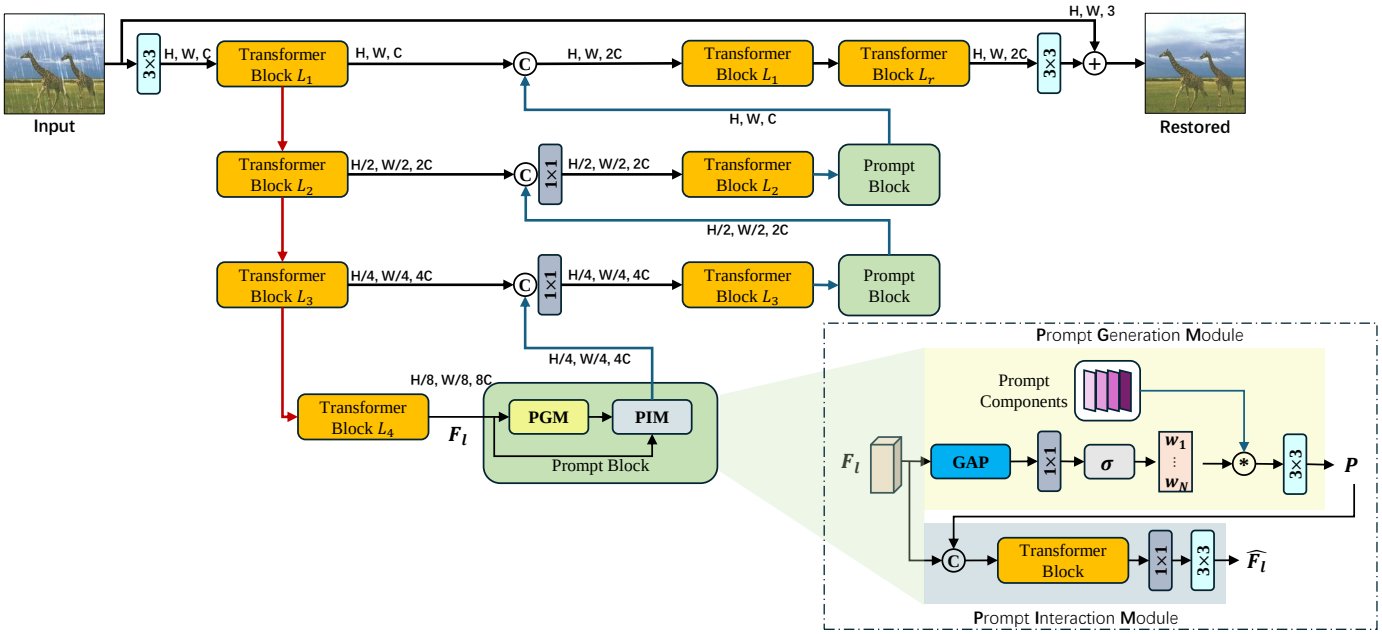

Figure 1. Overview of the PromptIR approach. The prompt block consists of two modules: the Prompt Generation Module (PGM) and the Prompt Interaction Module (PIM). The prompt generation module generates the input-conditioned prompt P, using the input features F_l and the Prompt Components. The prompt interaction module then dynamically adapts the input features using the generated prompt through the transformer block. The prompts interact with decoder features at multiple levels to enrich the degradation-specific context.

1 Introduction

In various computer vision tasks, image acquisition is the starting point of the whole pipeline, and the good quality of the raw images collected through cameras, camcorders, and other devices is crucial for performing downstream tasks. Image quality degradation (e.g., noise, blur, rain, snow, etc.) often occurs due to physical limitations of the camera or unsuitable environmental conditions, and image restoration, which is the process of restoring high-quality clean images from degraded images, is extremely challenging due to the existence of many different specific causes of degradation. Recently, deep learning-based restoration methods have emerged as a more effective choice compared to traditional methods.

Among these approaches, PromptIR [3] is a prompt-based model that interacts with input features through lightweight modules to perceive image degradation information and generates prompts to fuse with input features to guide the image restoration process. Although the model demonstrates powerful multitask image restoration capabilities, its large number of parameters leads to significant time and memory overheads in full-parameter training, which is not user-friendly for devices with limited resources, especially those deployed at the edge. In addition, the application of PromptIR is still limited to the image restoration field where the model processes independent and unrelated images, but in reality, the most common case is video frames with temporal continuity.

Based on the above observations, we enhanced existing work by expanding the capabilities of PromptIR to the video field and introducing a degradation kernel unique to video compression, then training the model to restore the video frame by frame. In addition, we reduced memory overhead during the training process by adjusting the model configuration and applying mixed-precision training strategy. Specifically, our work is listed as follows:

- We deployed PromptIR and verified its performance as claimed in the original paper.
- We created a dataset of compressed video frames and fine-tuned the model.
- We adjusted the configuration of the model and applied mixed-precision training strategy to reduce overhead while maintaining precision.

2 Related Works

2.1 Image Restoration

Image restoration has been widely studied in the field of computer vision and many methods have been proposed to address various challenges. In recent years, deep learning-based methods have gained widespread attention because of their ability to learn complex representations. For example, Hang Guo et al. [1] proposed MambaIR, an image restoration method based on a state-space model, which improved the performance and efficiency of image restoration through local enhancement and the channel attention mechanism. Xin Su et al. [4] proposed a Review Learning training method, which improved the performance of the image restoration model on multiple types of degradation through iterative learning and periodic review of challenging samples, while designing a lightweight SimpleIR network to achieve an efficient process of restoring 4K resolution degraded images. In the field of compressed image restoration, Wan et al. [5] proposed a feature consistency

training method to enhance the robustness of deep neural networks to compression artifacts by minimizing the feature distortion between the original image and its compressed JPEG version and demonstrated cross-task robustness to unseen distortions. Long Peng et al. [2] proposed a lightweight adaptive feature de-drifting module that effectively improved the accuracy of compressed image classification by learning the mapping relationship between feature drift and frequency distribution in the DCT domain, especially on resource-constrained mobile devices.

2.2 PromptIR

PromptIR proposes a plug-and-play module, and the main design is to utilize lightweight trainable parameters to interact with input features to perceive image degradation information. The module can be easily integrated into existing image restoration models, and the authors of PromptIR used Restormer [6] as the backbone of the restoration network, integrating scale-matched PromptIR modules into the three transformer blocks responsible for processing features of different scales in the decoder to enhance performance, as shown in Figure 1. The experiment covers image degradation caused by various weather factors, such as rain and snow, as well as interference caused by noise. The results show that the Restormer model integrated with PromptIR is capable of coping with different types of image degradation and restoring the quality of degraded images close to the original.

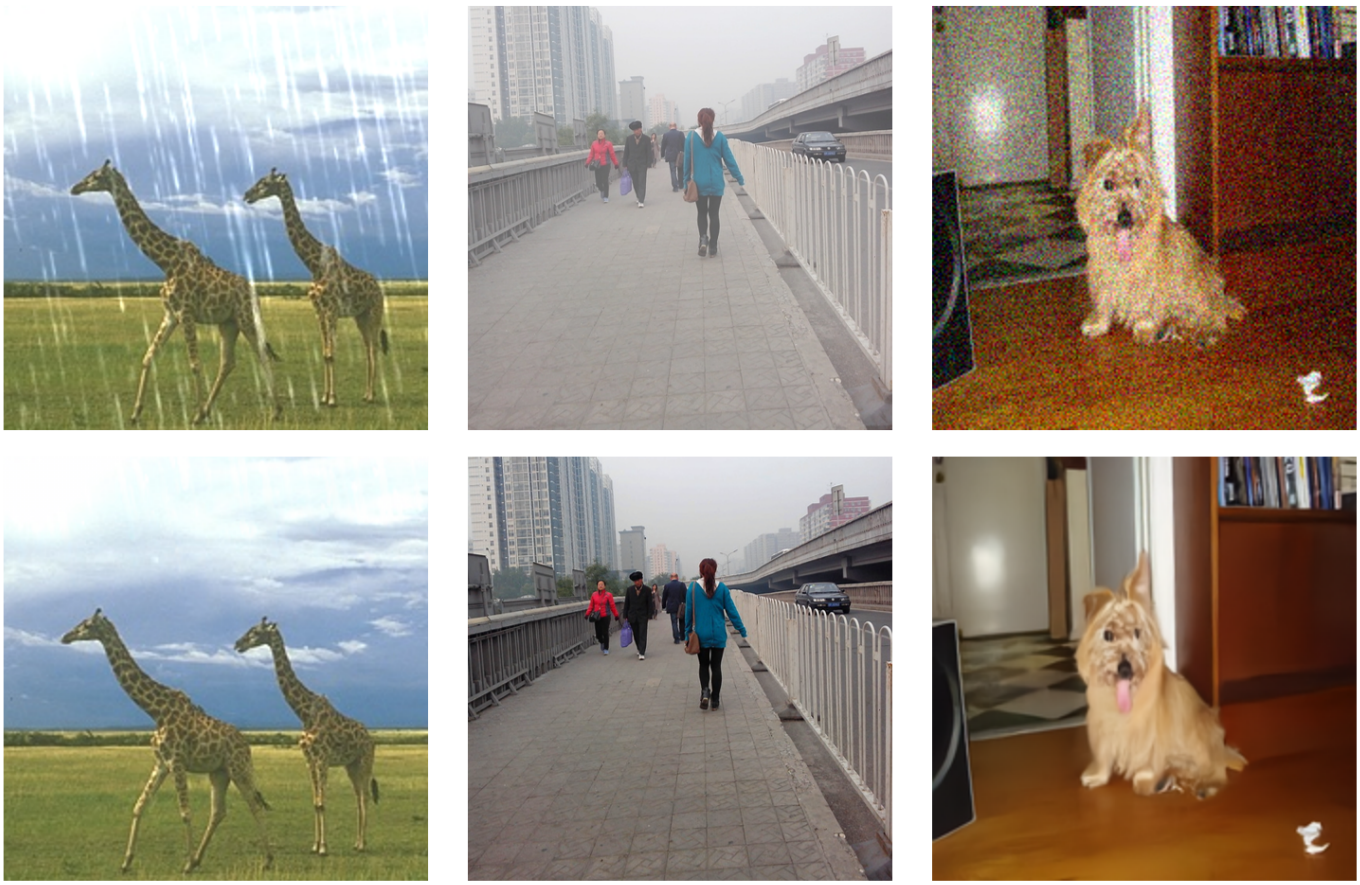

Figure 2. An all-in-one image restoration example of Restormer integrated with PromptIR. By perceiving the degradation information through prompts, the model can complete multiple types of image restoration tasks.

3 Method

3.1 Observation

Our work is based on existing open source code and improves potential problems in model design and training methods.

(a) Training Method: The authors claim that PromptIR is a "plug-and-play lightweight module", which means that using the additional prompt modules should not require retraining or fine-tuning the backbone network. However, in the implementation, the authors created a full model consisting of Restormer as well as PromptIR module, and trained all parameters from scratch. This is inconsistent with the effect claimed in the paper and will introduce huge training overhead. In addition, considering that Restormer itself is already capable of restoring various types of image degradation, such a training method does not reflect the capabilities of the PromptIR module.

(b) Parameter Scale: As a lightweight module, the parameter scale of PromptIR should be significantly smaller than that of the backbone network. It should not cause excessive overhead during training. But in practice, the authors integrated the transformer block of the Restormer network into the PromptIR module, which significantly increases the parameter size, as shown in Table 1. Compared with adding a "plug-in" module, it is more like simply increasing the size of the original network, which we think needs modification.

Module	Parameters	Proportion
Full Model	33.95M	-
Backbone	23.21M	68.4%
PromptIR	**10.74M**	**31.6%**

Table 1. Parameter distribution of the model

3.2 Enhancements

To address these issues, we design improvement schemes and additional experiments regarding both interpretability and overhead.

3.2.1 Interpretability

To verify that the PromptIR module works, we compared the performance between a) retraining the whole model and b) freezing the Restormer and only finetuning the PromptIR. Besides, considering that Restormer itself can already cope with several existing image degradations, we extended the restoration task to video by introducing a high CRF(Constant Rate Factor) compress, which is a video-specific degradation method, and verified the model's performance.

3.2.2 Overhead

We observed that the parameter scale of PromptIR comes mainly from the following parts:

(a) Deep Transformer Block: Due to the large number of deep feature channels in the network, the number

of parameters in the FFN layer of the transformer will increase significantly.
(b) Shallow PGM Module: The shallow features have fewer channels, but the prompt component in the PGM (essentially a set of trainable parameters equivalent to the size of the input) will become huge due to the large feature size.

Correspondingly, we first reduce the scale factor of the deep transformer block which determines the multiple of the feature dimension expansion to reduce the parameters of its FFN layer. At the same time, we adjust the size of the prompt component in the shallow PGM module by reducing it to half of the input feature and then restoring its size to the same as the input feature through bilinear interpolation subsequently to reduce redundancy. We tried various configurations and performed experiments to ensure that we did not significantly degrade performance when changing the model configuration.

4 Implementation Details

4.1 Dataset Preparation

To extend the capabilities of PromptIR to video, we chose the DAVIS 2016 dataset for experiments. This dataset consists of video frames from different video clips and was originally used for video object segmentation tasks. The dataset contains 3455 video frames, and we divide the training set and validation set in a ratio of 4:1. To create image degradation unique to the video domain, we use ffmpeg to re-encode the video frames, set a higher CRF value (CRF=36 in the experiment) to compress the original video, and then split it into independent degraded frames. CRF is an important parameter in video encoding and is mainly used to control the balance between video quality and file size. It is commonly used in H.264/H.265 encoders and is an efficient quality control method. Unlike constant bit rate (CBR) or constant quality (QP, Quantization Parameter), CRF achieves relatively constant perceptual quality by dynamically adjusting the bit rate of each frame. In general, the lower CRF value leads to higher quality of the encoded video and correspondingly larger file size; setting a higher CRF can save bandwidth and reduce the video size, but the cost is that the video frame will undergo significant distortion.

4.2 Reducing Overhead

As described in 3.2, we first make adjustments to the model configuration. For the shallow PGM module, the original prompt component scale (H, W) = (180, 320) is equal to the size of the input features. We consider such a configuration redundant and adjust it to (H', W') = (H/2, W/2), i.e. the length and width are each reduced to half of the original, and subsequently use bilinear interpolation to refill the intermediate result back to (H, W). For the deep transformer block, the original scale factor is 2.66, which we changed to 1.5 to reduce the number of parameters of the FFN layer. In addition to changing the model configuration, we enable half-precision training during the fine-tuning process and use fp16 instead of the default fp32 to further reduce the memory overhead.

4.3 Configurations

We choose L1 Loss for the stability during training; number of training epochs is set to 20, and validation is performed after each epoch; the maximum learning rate is set to 2e-4, and we adopt an update strategy of cosine decay with a warmup period of 2 epochs. We use AdamW optimizer during the process.

5 Results and Analysis

Figure 3. Frame compression and restoration results. The two sets of pictures from left to right are original frames, compressed frames, and restored frames.

Figure 3 shows the video frames before and after the model is restored. The model fine-tuned on the new task can restore the high-CRF compressed video frames. As shown in Figure 3, the restored video frames have less jaggedness and distortion, and the re-encoded video can achieve a smoother viewing experience.
To compare the performance of the fully re-trained model and the modified fine-tuned model, we sample some video frames for quantitative comparison. We choose the Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index Metric (SSIM) as metrics. The former is a pixel-based objective quality evaluation indicator used to measure the error between the reconstructed image and the reference image, calculated by the mean square error (MSE); the latter is a structure-based perceptual quality evaluation indicator that aims to simulate the sensitivity of the human eye to the structure, brightness, and contrast of image perception, calculated by the mean, variance, and other information of the two images. The results are shown in Table 2, where the optimal and suboptimal results are marked in bold and underlined, respectively. The performance of the fine-tuned model is slightly lower than that of the full-parameter retrained model, but the gap is relatively small. Moreover, with the light-weighting strategy, the memory overhead during training is reduced from 29.91 GB

for the full-parametric training to 11.26 GB for fine-tuning, which is meaningful for deploying the model on resource-constrained devices.

Table 2. Metrics on different frames

Model	**Frame 1**		**Frame 2**		**Frame 3**	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Compress	27.150	0.782	24.496	0.646	28.374	0.797
No Retrained	26.356	0.776	24.207	0.645	27.472	0.797
Full-Parameter Retrained	**28.577**	**0.835**	**25.418**	**0.695**	**29.364**	**0.822**
*Lightweight Finetuned	<u>28.100</u>	<u>0.822</u>	<u>25.108</u>	<u>0.681</u>	<u>29.051</u>	<u>0.814</u>

6 Conclusion and Future Work

6.1 Conclusion

In this article, we improve the existing PromptIR work by changing the model configuration to reduce the number of redundant parameters, adopting fine-tuning and half-precision training strategies to verify the capabilities of the PromptIR module itself while reducing training overheads, and migrating the task to video frame restoration to verify the generalization ability of PromptIR. The quantitative evaluation shows that the improvements we made significantly reduce training overhead while maintaining the precision of the model, and the module can be well adapted to tasks in different fields.

6.2 Future Work

Focusing on the lightweighting and generalization capabilities of the module, we plan to consider the following factors in future research:

(a) Further lightweighting: Our work has significantly reduced the memory overhead during model training by improving model configuration and training methods, but the current work still has room for further improvement, especially considering that the size of the PromptIR module is highly related to the dimensions of the input features (including channels, sizes), which means larger input features will lead to significantly increased parameter quantities and activations. We consider designing a general adaptability method to ensure that the additional overhead caused by PromptIR can be unified into a reasonable range and adapted to different locations in different backbone networks.

(b) Further exploration of generalization ability: Our work migrated PromptIR to the video field and achieved good performance. However, if the model still needs to have the ability to recover from image degradation caused by weather factors, it is necessary to take it into consideration during the training phase, and we need to design a reasonable training method so that the model can have the ability to recover from multiple degradation kernels at the same time.

References

[1] Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model. In *European Conference on Computer Vision*, 2024.

[2] Long Peng, Yang Cao, Yuejin Sun, and Yang Wang. Lightweight adaptive feature de-drifting for compressed image classification. *IEEE Transactions on Multimedia*, 26:6424–6436, 2024.

[3] Vaishnav Potlapalli, Syed Waqas Zamir, Salman H. Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one image restoration. In *Neural Information Processing Systems*, 2023.

[4] Xin Su, Zhuoran Zheng, and Chen Wu. Review learning: Advancing all-in-one ultra-high-definition image restoration training method. *ArXiv*, abs/2408.06709, 2024.

[5] Sheng Wan, Tung-Yu Wu, Heng-Wei Hsu, Wing Hung Wong, and Chen-Yi Lee. Feature consistency training with jpeg compressed images. *IEEE Transactions on Circuits and Systems for Video Technology*, 30:4769–4780, 2020.

[6] Syed Waqas Zamir, Aditya Arora, Salman Hameed Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5718–5729, 2021.