Intra and Inter Parser-Prompted Transformers for Effective Image Restoration

Cong Wang[123] **Jinshan Pan**[4], **Liyan Wang**[5], **Wei Wang**[1*]

[1]Shenzhen Campus of Sun Yat-sen University, China
[2]Centre for Advances in Reliability and Safety, Hong Kong
[3]The Hong Kong Polytechnic University, Hong Kong
[4]Nanjing University of Science and Technology, China
[5]Dalian University of Technology, China
supercong94@gmail.com, sdluran@gmail.com, wangwei29@mail.sysu.edu.cn

Abstract

We propose Intra and Inter Parser-Prompted Transformers (PPTformer) that explore useful features from visual foundation models for image restoration. Specifically, PPTformer contains two parts: an Image Restoration Network (IRNet) for restoring images from degraded observations and a Parser-Prompted Feature Generation Network (PPFGNet) for providing IRNet with reliable parser information to boost restoration. To enhance the integration of the parser within IRNet, we propose Intra Parser-Prompted Attention (IntraPPA) and Inter Parser-Prompted Attention (InterPPA) to implicitly and explicitly learn useful parser features to facilitate restoration. The IntraPPA re-considers cross attention between parser and restoration features, enabling implicit perception of the parser from a long-range and intra-layer perspective. Conversely, the InterPPA initially fuses restoration features with those of the parser, followed by formulating these fused features within an attention mechanism to explicitly perceive parser information. Further, we propose a parser-prompted feed-forward network to guide restoration within pixel-wise gating modulation. Experimental results show that PPTformer achieves state-of-the-art performance on image deraining, defocus deblurring, desnowing, and low-light enhancement.

Code — https://github.com/supersupercong/pptformer

Introduction

Image restoration aims to reconstruct high-quality images from their degraded counterparts. This task is inherently challenging because it relies solely on the degraded images themselves while the clear images and degradation factors are unknown. To effectively solve this problem, statistical observations are employed to transform it into a 'well-posed' one, as discussed in various studies (Pan et al. 2016). While traditional methods have achieved some success in restoration, they are hampered by complex optimization algorithms that struggle with issues of non-convexity and non-smoothness.

The emergence of convolutional neural networks (He et al. 2016) and Transformers (Dosovitskiy et al. 2021) has revolutionized image restoration tasks. These advanced models have excelled by implicitly learning from large-scale data. This learning-based strategy has taken precedence in contemporary image restoration, outperforming previous methods with notable success, as evidenced by a series of studies (Ren et al. 2019; Liu et al. 2018; Wang, Pan, and Wu 2022; Wang et al. 2020b,a, 2021; Guo et al. 2022; Peng et al. 2024b,a; Wang et al. 2023, 2024a,b,d,c,e; Xu et al. 2024b,a). However, we note that most of the existing state-of-the-art approaches learn the 'self' degraded knowledge without considering additional helpful knowledge from other domains into network designs, limiting model capacity to some extent.

To address this problem, some conditional modulation-driven networks are suggested (Wang et al. 2018a). These approaches usually contain a restoration branch for image reconstruction and a conditional branch for providing

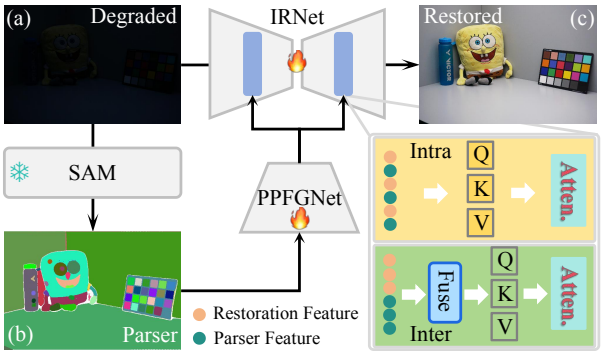


Figure 1: Illustration of our main idea. Our method is motivated by an interesting observation that SAM (Kirillov et al. 2023) can parse degraded images into useful hierarchical structures (b) although severely degraded inputs (a). While extreme degradation may not provide valuable information for restoration, the parser that benefits from the powerful ability of SAM can still describe reliable structures well to facilitate restoration. To better integrate the parser into the restoration process, we develop Intra Parser-Prompted Attention and Inter Parser-Prompted Attention to implicitly and explicitly learn valuable parser content to boost image recovery.

*Wei Wang is the corresponding author.

the restoration branch with useful conditional information. Among them, degraded images are usually severed as conditions to provide pixel-wise knowledge (Wang et al. 2018a). We notice that the degraded images usually contain unreliable pixels due to the degraded disruption. For example, in low-light conditions, extensive pixels approximate zero (as illustrated in Fig. 1(a)), which may not provide reliable content for the restoration branch, thus limiting recovery performance.

Recent large visual foundation models, e.g., SAM (Kirillov et al. 2023), have shown strong ability in visual understanding (Lu et al. 2023). As one of the powerful visual foundation models, the SAM model can effectively parse degraded images into hierarchical structure content although severe degradation (see Fig. 1(b)). Compared with degraded images (Fig. 1(a)) which almost cannot provide useful information, the parsed contents (Fig. 1(b)) contain semantic information with salient structures. Therefore, *a natural question is whether the parsed contents help image restoration. If so, how do we explore the parsed contents to better help image restoration?*

In this paper, we propose an Intra and Inter Parser-Prompted Transformer (PPTformer) to answer the above questions. Our PPTformer adequately considers the parsing content generated by SAM (Kirillov et al. 2023) into the restoration process within both long-range pixel dependency and pixel-wise modulation perspectives. Specifically, our PPTformer contains two parts: one is the Image Restoration Network (IRNet) for restoring images from the degraded observations while another is the Parser-Prompted Feature Generation Network (PPFGNet) for providing IRNet with reliable parsing features to boost restoration performance. To better integrate parsing content into IRNet, we propose the Intra Parser-Prompted Attention (IntraPPA) and the Inter Parser-Prompted Attention (InterPPA), which implicitly and explicitly learn useful parser features to guide restoration, respectively. The IntraPPA re-considers cross attention between parser features and restoration ones to implicitly perceive the parser content within the long-range intra-layer perspective, while the InterPPA first fuses the restoration features with parser ones and then conducts attention computation to explicitly perceive the parser features. Further, we propose a Parser-Prompted Feed-forward Network (PPFN) to guide restoration from a pixel-wise gating modulation perspective. Moreover, we propose a bidirectional parser-prompted fusion scheme to better fuse the parser features and restoration ones, which would be adopted in both InterPPA and PPFN. Fig. 1 illustrates the main idea of our PPTformer.

The main contributions are summarized below:

- We propose an intra and inter parser-prompted Transformer, which integrates the visual foundation models into the restoration process.
- We propose an intra parser-prompted attention and an inter parser-prompted attention to implicitly and explicitly explore useful parser features to facilitate restoration.
- We suggest a bidirectional parser-prompted fusion scheme to effectively fuse parser and restoration features.
- We show that the proposed method achieves favorable results on image restoration tasks including image deraining, single-image defocus deblurring, image desnowing, and low-light image enhancement.

Proposed Approach

Our goal aims to exploit SAM (Kirillov et al. 2023) to parse degraded images into hierarchical structures to prompt the restoration process with more useful information to facilitate restoration. To that end, we propose the PPTformer, an intra and inter Parser-Prompted Transformer. To better integrate parser information into restoration, we propose an intra parser-prompted attention, an inter parser-prompted attention, and a parser-prompted feed-forward network. We also suggest a bidirectional parser-prompted fusion scheme to fuse parser features with restoration ones.

Overall pipeline

Fig. 2 shows the overview of our PPTformer. It contains two parts: **(a)** Image Restoration Network (IRNet) and **(b)** Parser-Prompted Feature Generation Network (PPFGNet). The IRNet is used to restore images from given degraded observations while the PPFGNet is used to parse input degraded images into hierarchical structures to prompt the IRNet with useful information to facilitate restoration. To better fuse parser features into IRNet, we propose Intra Parser-Prompted Attention (see Fig. 3(a)), Inter Parser-Prompted Attention (see Fig. 3(b)) to implicitly and explicitly learn useful parser features from the long-range pixel dependency perspective, respectively. Furthermore, we propose a Parser-Prompted Feed-forward Network (see Fig. 3(c)) to integrate parser features into the restoration process within the pixel-wise gating modulation perspective. Moreover, we propose a Bidirectional Parser-Prompted Fusion scheme to effectively fuse parser features and restoration ones.

Image Restoration Network. Given a degraded input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, we first apply a 3×3 convolution as the feature extraction to obtain low-level embeddings $\mathbf{X}_0 \in \mathbb{R}^{H\times W\times C}$; where $H \times W$ denotes the spatial dimension and C is the number of channels. Next, the shallow features $\mathbf{X}_0$ gradually are hierarchically encoded into deep features $\mathbf{X}_l \in \mathbb{R}^{\frac{H}{l}\times\frac{W}{l}\times lC}$. After encoding the degraded input into low-resolution latent features $\mathbf{X}_3 \in \mathbb{R}^{\frac{H}{3}\times\frac{W}{3}\times 3C}$, the decoder progressively recovers the high-resolution representations. Finally, a reconstruction layer which contains a 3×3 convolution is applied to decoded features to generate residual image $\mathbf{S} \in \mathbb{R}^{H\times W\times 3}$ to which degraded image is added to obtain the restored output image: $\hat{\mathbf{H}} = \mathbf{I} + \mathbf{S}$.

Both encoder and decoder at l- level consists of one Inter and Intra Parser-Prompted Transformer (IN2PPT), as shown in Fig. 2(c), to perceive parser features, and multiple Parser-Prompted Transformers Blocks (PPTB) which consists of a Parser-Prompted Attention (PPA) and a Parser-Prompted Feed-forward Network (PPFN). To help better recovery, the encoder features are concatenated with decoder features via skip connections (Ronneberger, Fischer, and Brox 2015) by 1×1 convolutions.

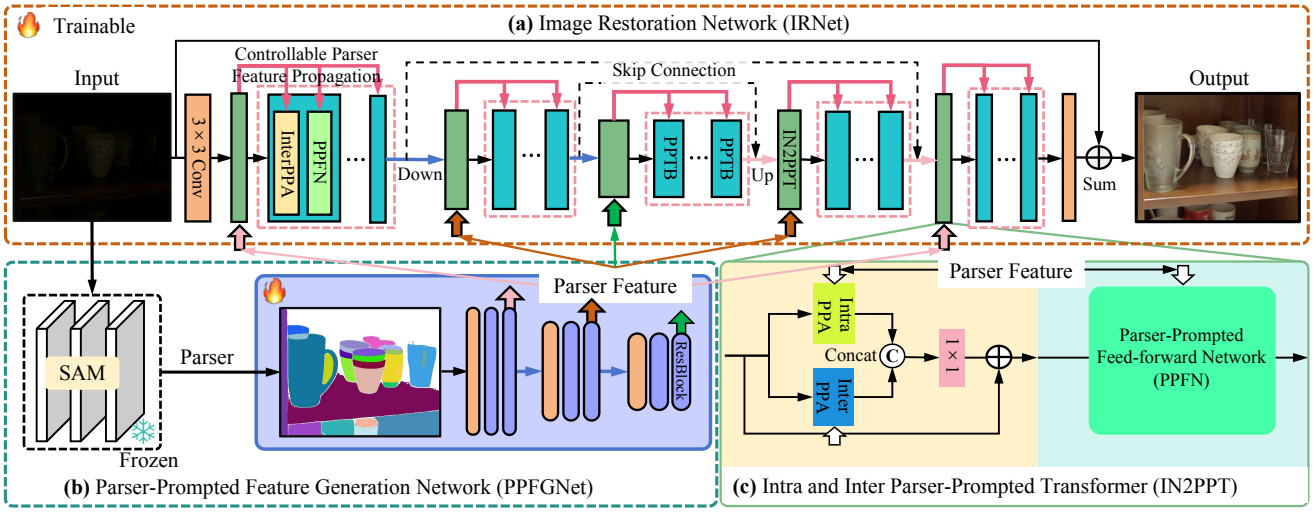


Figure 2: Overall framework of PPTformer. Our PPTformer consists of two parts: **(a)** Image Restoration Network (IRNet); **(b)** Parser-Prompted Feature Generation Network (PPFGNet). The IRNet is used to restore images, while PPFGNet is used to generate parser features to provide IRNet with useful information to facilitate restoration. To better utilize the parser features to guide IRNet, we propose the Intra and Inter Parser-Prompted Attention, which implicitly and explicitly explore the useful parser features in the restoration process. Further, we propose the Parser-Prompted Feed-forward Network to integrate parser features into the feed-forward encoding process, which allows parser features to effectively guide the restoration within the pixel-wise gating modulation perspective. Moreover, we introduce the Controllable Parser Feature Propagation scheme to control parser feature propagation in both attention and networks to allow useful information to be passed for better guide image restoration.

Parser-Prompted Feature Generation Network. The PPFGNet aims to parse degraded images into hierarchical structures and then generates multi-scale parser features to provide guidance for IRNet. Specifically, we first utilize SAM (Kirillov et al. 2023) to parse input images into hierarchical structures, which are then input to an autoencoder-like network to generate multi-scale parser features. The parser features would be utilized to prompt IRNet with useful information to facilitate better restoration.

Intra and Inter Parser-Prompted Transformer

To effectively perceive and exploit parser features, we propose the Intra and Inter Parser-Prompted Transformer (IN2PPT), which contains an Intra Parser-Prompted Attention (IntraPPA), an Inter Parser-Prompted Attention (InterPPA), and a Parser-Prompted Feed-forward Network (PPFN). The IntraPPA, shown in Fig. 3(a), implicitly learns parser features by building cross-attention between parser features and restoration ones, while the InterPPA, shown in Fig. 3(b), explicitly explores useful parser features by first fusing parser features via Bidirectional Parser-Prompted Fusion (see Fig. 4) and restoration ones before conducting attention computation, both to prompt restoration from the long-range pixel dependency perspective. The PPFN, shown in Fig. 3(c), integrates parser features into restoration by pixel-wise gating modulation. The IntraPPA and InterPPA adopt the parallel structures and are further concatenated and fused by one 1×1 convolution. The fused features are then sent to PPFN.

Intra Parser-Prompted Attention. The IntraPPA, shown in Fig. 3(a), fully considers the implicit global fusion between parser features and restoration ones. IntraPPA utilizes cross-attention to interactively learn more useful parser features for restoration, allowing restoration features to implicitly explore more reliable parser information to effectively guide the restoration process. Specifically, we first split parser features $\mathbf{M}$ into two tensors: Key ($\mathbf{K_M}$) and Value ($\mathbf{V_M}$), by a 1×1 point-wise convolution (W_p) and 3×3 depth-wise convolution (W_d). With similar manner, restoration features $\mathbf{X}$ are split into three tensors: Query ($\mathbf{Q_R}$), Key ($\mathbf{K_R}$), and Value ($\mathbf{V_R}$). Then, we respectively fuse the Keys and Values of restoration features and parser features by concatenation and 1×1 convolution to form a new Key and Value. Lastly, the Query, Key, and Value vectors are performed attention. The process of IntraPPA can be formulated by:

$$
\begin{aligned}
\mathbf{K_M}, \mathbf{V_M} &= \mathcal{S}\Big(W_d W_p(\mathbf{M})\Big), \\
\mathbf{Q}, \mathbf{K_R}, \mathbf{V_R} &= \mathcal{S}\Big(W_d W_p(\mathbf{X})\Big), \\
\mathbf{K} &= W_p\Big(\mathcal{C}\big[\mathbf{K_M}, \mathbf{K_R}\big]\Big), \\
\mathbf{V} &= W_p\Big(\mathcal{C}\big[\mathbf{V_M}, \mathbf{V_R}\big]\Big), \\
\mathbf{A}_{\text{Intra}} &= \mathbf{V} \otimes \text{Softmax}\Big(\mathbf{K} \otimes \mathbf{Q}/\alpha\Big),
\end{aligned}
\tag{1}
$$

where $\mathcal{S}(\cdot)$ means the split operation; $\mathcal{C}[\cdot,\cdot]$ denotes the concatenation operation at channel dimension; α is a learnable scaling parameter to control the magnitude of the dot prod-

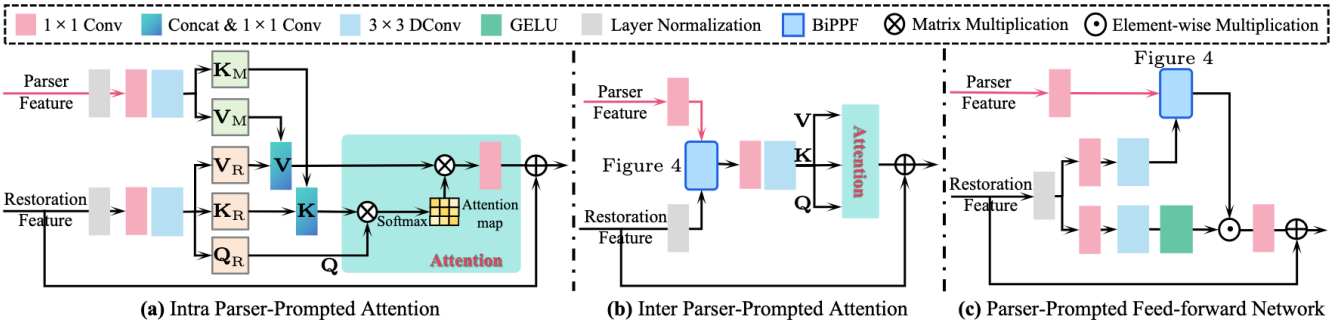


Figure 3: (a) Intra Parser-Prompted Attention (IntraPPA), (b) Inter Parser-Prompted Attention (InterPPA), and (c) Parser-Prompted Feed-forward Network (PPFN). Our IntraPPA exploits the cross-attention between parser features and restoration features to implicitly explore useful parser features. The InterPPA explicitly explores the aggregation, which first utilizes BiPPF (see Fig. 4) to fuse parser features with restoration ones and then conducts attention to explicitly learn beneficial parser features. The PPFN integrates parser features into one of the parallel paths by BiPPF, which allows parser features to effectively guide the feed-forward restoration process within a pixel-wise gating modulation mechanism.

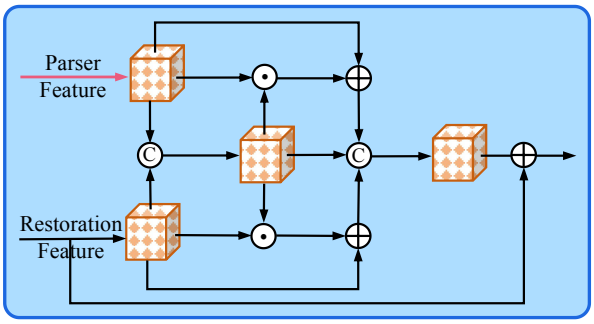


Figure 4: Bidirectional Parser-Prompted Fusion (BiPPF). Our BiPPF method effectively integrates a bidirectional flow scheme for feature fusion. It transforms parser features into restoration features and vice versa, allowing interactive integration of parser features into the restoration process. Notably, all convolutions in BiPPF are 1×1 for efficient design.

uct; $\otimes$ denotes the matrix multiplication (Zamir et al. 2022); $\mathbf{A}_{\text{Intra}}$ means the output of IntraPPA.

Inter Parser-Prompted Attention. The InterPPA, shown in Fig. 3(a), re-considers the explicit global fusion between parser features and restoration ones within attention. Different from IntraPPA, the InterPPA first fuses parser features with restoration ones by the introduced Bidirectional Parser-Prompted Fusion (see Fig. 4). Then, the fused features are computed via an attention module. This process of InterPPA can be formulated as:

$$\begin{aligned}
\mathbf{X}_{\text{BiPPF}} &= \mathcal{P}\Big(\mathbf{X}, \mathbf{M}\Big), \\
\mathbf{Q}, \mathbf{K}, \mathbf{V} &= \mathcal{S}\Big(W_d W_p(\mathbf{X}_{\text{BiPPF}})\Big), \\
\mathbf{A}_{\text{Inter}} &= \mathbf{V} \otimes \text{Softmax}\Big(\mathbf{K} \otimes \mathbf{Q}/\alpha\Big),
\end{aligned} \tag{2}$$

where $\mathcal{P}(\cdot,\cdot)$ denotes the operation of Bidirectional Parser-Prompted Fusion; $\mathbf{A}_{\text{Intra}}$ means the output of InterPPA.

Parser-Prompted Feed-Forward Network. Our PPFN is based on GDFN (Zamir et al. 2022), which contains two parallel paths with a gating mechanism to control information flow. We modify the GDFN by integrating parser features into it. Specifically, we integrate parser features into one of the parallel paths by BiPPF, which allows parser features to effectively guide the restoration process within a pixel-wise modulation perspective. Then, the integrated features are gated with the features of another path. The PPFN can be formulated as:

$$\begin{aligned}
\mathbf{X}_1, \mathbf{X}_2 &= \mathcal{S}\Big(W_d W_p(\mathbf{X})\Big), \\
\mathbf{X}_{\text{BiPPF}} &= \mathcal{P}\big(\mathbf{X}_1, \mathbf{M}\big), \\
\mathbf{X}_{\text{PPFN}} &= W_p\Big(\mathbf{X}_{\text{BiPPF}} \odot \phi(\mathbf{X}_2)\Big) + \mathbf{X},
\end{aligned} \tag{3}$$

where ϕ denotes the GELU activation; $\odot$ means the element-wise multiplication; $\mathbf{X}_{\text{PPFN}}$ refers to the output of PPFN. Following (Zamir et al. 2022), we use expanding channel capacity factor as 3 to intermediate feature in PPFN.

Controllable Parser Feature Propagation

To enhance the capability of each block in perceiving parser features for improved restoration, we introduce the Controllable Parser Feature Propagation (CPFP). CPFP effectively channels valuable parser information into the attention and feed-forward networks within our Parser-Prompted Transformer Block (PPTB). In this work, the InterPPA serves as the attention layer in PPTB, while the PPFN is utilized as its feed-forward network.

Bidirectional Parser-Prompted Fusion

To better fuse parser features with restoration ones, we develop a Bidirectional Parser-Prompted Fusion (BiPPF) module, as shown in Fig. 4. Unlike the previous widely-used fusion module SFT (Wang et al. 2018b), we explore the feature

Benchmark	Metrics	DerainNet	SEMI	UMRL	RESCAN	PreNet	MSPFN	DCSFN	MPRNet	SPAIR	Uformer	MAXIM-2S	SFNet	**PPTformer**
Test100	PSNR ↑	22.77	22.35	24.41	25.00	24.81	27.50	27.46	30.27	30.35	29.17	31.17	31.47	31.48
	SSIM ↑	0.810	0.788	0.829	0.835	0.851	0.876	0.887	0.897	0.909	0.880	0.922	0.919	0.922
Rain100H	PSNR ↑	14.92	16.56	26.01	26.36	26.77	28.66	28.98	30.41	30.95	30.06	30.81	31.90	31.77
	SSIM ↑	0.592	0.486	0.832	0.786	0.858	0.860	0.887	0.890	0.892	0.884	0.903	0.908	0.907
Test2800	PSNR ↑	24.31	24.43	29.97	31.29	31.75	32.82	30.96	33.64	33.34	33.36	33.80	33.69	34.01
	SSIM ↑	0.861	0.782	0.905	0.904	0.916	0.930	0.903	0.938	0.936	0.935	0.943	0.937	0.945
Average	PSNR ↑	20.67	21.11	26.80	27.55	27.78	29.66	29.13	31.44	31.55	30.86	31.93	*32.35*	**32.42**
	SSIM ↑	0.754	0.685	0.855	0.842	0.875	0.889	0.892	0.908	0.912	0.900	*0.923*	0.921	**0.925**

Table 1: Image deraining results. Our PPTformer advances recent 13 state-of-the-art approaches on average.

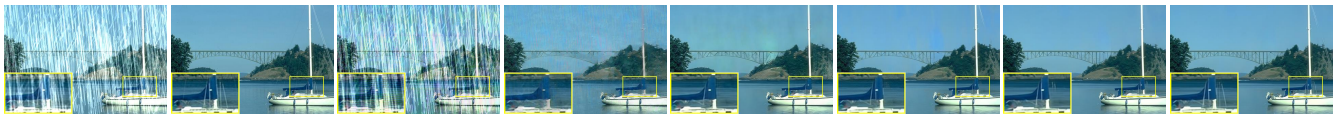


Figure 5: Image deraining example on Rain100H (Yang et al. 2017a). Our PPTformer recovers results with finer structures.

fusion manner within bidirectional perspectives to maximize fusion. Specifically, we first concatenate parser features and restoration ones and then use 1×1 convolution to fuse them. Then, the fused features are gated with parser features and restoration features by element-wise multiplication, which are respectively added to the original features. Lastly, we concatenate these added features and use 1×1 convolution to fuse them. The process can be formulated into:

$$\begin{aligned}
&\hat{\mathbf{X}} = W_p(\mathbf{X}); \hat{\mathbf{M}} = W_p(\mathbf{M}); \mathbf{X}_{\text{fusion}} = W_p\Big(\mathcal{C}\big[\hat{\mathbf{X}}, \hat{\mathbf{M}}\big]\Big), \\
&\tilde{\mathbf{X}} = \mathbf{X}_{\text{fusion}} \odot \hat{\mathbf{X}} + \hat{\mathbf{X}}; \tilde{\mathbf{M}} = \mathbf{X}_{\text{fusion}} \odot \hat{\mathbf{M}} + \hat{\mathbf{M}}, \\
&\mathbf{X}_{\text{BiMPF}} = W_p\Big(\mathcal{C}\big[\tilde{\mathbf{X}}, \tilde{\mathbf{M}}\big]\Big) + \mathbf{X},
\end{aligned} \tag{4}$$

where $\mathbf{X}_{\text{BiMPF}}$ denotes the output of BiPPF.

Experimental Results

We evaluate **PPTformer** on benchmarks for 4 image restoration tasks: **(a)** image deraining, **(b)** single-image defocus deblurring, **(c)** image desnowing, and **(d)** low-light image enhancement.

Implementation Details

We train PPTformer using the AdamW optimizer (Loshchilov and Hutter 2019) with the initial learning rate $5e^{-4}$ that is gradually reduced to $1e^{-7}$ with the cosine annealing (Loshchilov and Hutter 2017). The training patch size is set as 256×256 pixels. For downsampling and upsampling, we adopt pixel-unshuffle and pixel-shuffle (Shi et al. 2016), respectively. For the parser, we first employ the SAM (Kirillov et al. 2023) to generate the parser before training the PPTformer. Then, we treat the parser as an image to input the PPTformer for efficiency as loading the SAM consumes much more memory when training. To constrain the training of PPTformer, we use the same loss function (Kong et al. 2023) with default parameters.

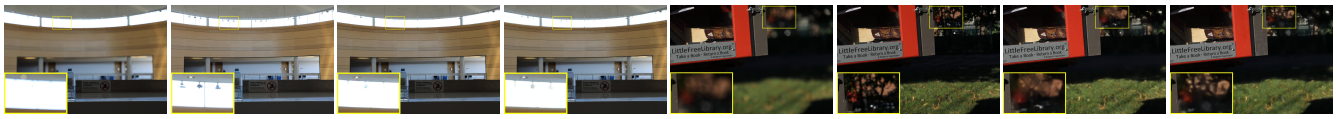
Main Results

Image Deraining Results. Similar to existing methods (Jiang et al. 2020; Zamir et al. 2021; Purohit et al. 2021), we report PSNR/SSIM scores using Y channel in YCbCr color by comparing with DerainNet (Fu et al. 2017a), SEMI (Wei et al. 2019), DIDMDN (Zhang and Patel 2018), UMRL (Yasarla and Patel 2019), RESCAN (Li et al. 2018), PreNet (Ren et al. 2019), MSPFN (Jiang et al. 2020),DCSFN (Wang et al. 2020b), MPRNet (Zamir et al. 2021), SPAIR (Purohit et al. 2021), Uformer (Wang et al. 2021), MAXIM-2S (Tu et al. 2022), SFNet (Cui et al. 2023). Tab. 1 shows that our PPTformer outperforms current state-of-the-art approaches when averaged across these three datasets, including Test100 (Zhang, Sindagi, and Patel 2019), Rain100H (Yang et al. 2017b), and Test2800 (Fu et al. 2017a). Compared to recent the best method SFNet (Cui et al. 2023), our PPTformer achieves 0.07 dB improvement on average. On individual datasets, the gain can be as large as 0.32 dB on Test2800 (Fu et al. 2017b). Fig. 5 further presents a challenging example on Rain100H (Yang et al. 2017b), where our PPTformer generates a clearer image with finer details.

Single-Image Defocus Deblurring Results. Tab. 2 summarises the single-image defocus deblurring results compared with conventional defocus deblurring methods (EBDB (Karaali and Jung 2017) and JNB (Shi, Xu, and Jia 2015)) as well as learning-based approaches (Abuolaim and Brown 2020; Son et al. 2021; Lee et al. 2021; Zamir et al. 2022) on the DPDD dataset (Abuolaim and Brown 2020). As one can see our PPTformer achieves comparable results with Restormer (Zamir et al. 2022) but outperforms other state-of-the-art approaches in terms of PSNR/SSIM/MAE/LPIPS metrics. In particular, PPTformer yields the best PSNR and MAE in all categories. Compared with Restormer (Zamir et al. 2022), our method improves 0.15dB PSNR gains on combined scenes. Fig. 6 presents two visual examples of indoor and outdoor scenes, where our PPTformer is more ef-

Method	Indoor Scenes				Outdoor Scenes				Combined			
	PSNR ↑	SSIM ↑	MAE ↓	LPIPS ↓	PSNR ↑	SSIM ↑	MAE ↓	LPIPS ↓	PSNR ↑	SSIM ↑	MAE ↓	LPIPS ↓
EBDB (Karaali and Jung 2017)	25.77	0.772	0.040	0.297	21.25	0.599	0.058	0.373	23.45	0.683	0.049	0.336
DMENet (Lee et al. 2019)	25.50	0.788	0.038	0.298	21.43	0.644	0.063	0.397	23.41	0.714	0.051	0.349
JNB (Shi, Xu, and Jia 2015)	26.73	0.828	0.031	0.273	21.10	0.608	0.064	0.355	23.84	0.715	0.048	0.315
DPDNet (Abuolaim and Brown 2020)	26.54	0.816	0.031	0.239	22.25	0.682	0.056	0.313	24.34	0.747	0.044	0.277
KPAC (Son et al. 2021)	27.97	0.852	0.026	0.182	22.62	0.701	0.053	0.269	25.22	0.774	0.040	0.227
IFAN (Lee et al. 2021)	28.11	0.861	0.026	0.179	22.76	0.720	0.052	0.254	25.37	0.789	0.039	0.217
Restormer (Zamir et al. 2022)	*28.87*	**0.882**	*0.025*	**0.145**	*23.24*	**0.743**	*0.050*	**0.209**	25.98	**0.811**	*0.038*	**0.178**
PPTformer	**29.16**	*0.880*	**0.024**	*0.156*	**23.26**	*0.737*	**0.050**	*0.227*	**26.13**	*0.807*	**0.037**	*0.193*

Table 2: Single-image defocus deblurring comparisons on the DPDD test set (Abuolaim and Brown 2020) (containing 37 indoor and 39 outdoor scenes). Our PPTformer achieves the best PSNR.



Input GT Restormer **Ours** Input GT Restormer **Ours**

Figure 6: Single-image defocus deblurring example in indoor and outdoor scenes on DPDD test set (Abuolaim and Brown 2020). Our PPTformer is able to recover clearer results with sharper structures. Best viewed with zoom-in.

fective in recovering image structures from complex blurry scenes than Restormer.

Image Desnowing Results. For the image desnowing task, we compare our PPTformer on the CSD (Chen et al. 2021), SRRS (Chen et al. 2020), and Snow100K (Liu et al. 2018) datasets with existing state-of-the-art competitors (Liu et al. 2018; Chen et al. 2020, 2021, 2022; Valanarasu, Yasarla, and Patel 2022). Except for the above methods with specific designs for image desnowing, we also compare with recent Transformer-based general image restoration approaches like Restormer (Zamir et al. 2022) and Uformer (Wang et al. 2021). Tab. 3 shows that our PPTformer respectively yields a 0.84 dB, 1.25 dB, and 1.47 dB PSNR improvement over the state-of-the-art approach (Zamir et al. 2022) on the CSD (Chen et al. 2021), SRRS (Chen et al. 2020), and Snow100K (Liu et al. 2018) benchmarks. The visual example in Fig. 7 demonstrates that our PPTformer is able to remove spatially varying snowflakes than state-of-the-art approaches.

Low-Light Image Enhancement Results. We perform the low-light image enhancement experiment on both LOL (Wei et al. 2018) and LOL-v2 (Yang et al. 2020). Tab. 4 summarizes the quantitative results, compared with Retinex-Net (Wei et al. 2018), Zero-DCE (Guo et al. 2020), AGLL-Net (Lv, Li, and Lu 2021), Zhao et al. (Zhao et al. 2021), RUAS (Liu et al. 2021), SCI (Ma et al. 2022), URetinex-Net (Wu et al. 2022), UHDFour (Li et al. 2023). Compared to recent works UHDFour (Li et al. 2023), our method receives 2.39 dB and 2.74 dB PSNR gains on LOL (Wei et al. 2018) and LOL-v2 (Yang et al. 2020), respectively. Fig. 8 shows our PPTformer is able to generate clearer results with more vivid colors.

Method	CSD (2000)		SRRS (2000)		Snow100K (2000)	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
DesnowNet	20.13	0.81	20.38	0.84	30.50	0.94
STASR	27.96	0.88	25.82	0.89	23.12	0.86
HDCW-Net	29.06	0.91	27.78	0.92	31.54	0.95
TransWeather	31.76	0.93	28.29	0.92	31.82	0.95
MSP-Former	33.75	0.96	30.76	0.95	33.43	0.96
Uformer	33.80	0.96	30.12	0.96	33.81	0.94
Restormer	35.43	0.97	32.24	0.96	34.67	0.95
PPTformer	**36.27**	**0.99**	**33.49**	**0.98**	**36.14**	**0.96**

Table 3: Image desnowing results on CSD (2000) (Chen et al. 2021), SRRS (2000) (Chen et al. 2020), and Snow100K (2000) (Liu et al. 2018).

Method	LOL		LOL-v2	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Retinex-Net	16.77	0.54	15.43	0.64
Zero-DCE	16.79	0.67	12.84	0.54
AGLLNet	17.52	0.77	20.69	0.78
Zhao et al.	21.67	0.87	18.84	0.78
RUAS	16.44	0.70	15.48	0.67
SCI	14.78	0.62	16.74	0.62
URetinex-Net	19.84	0.87	-	-
UHDFour	23.09	0.87	21.78	0.87
PPTformer	**25.48**	**0.93**	**23.52**	**0.91**

Table 4: Low-light image enhancement results on LOL (Wei et al. 2018) and LOL-v2 (Yang et al. 2020).

Ablation study

We conduct the ablation study using the low-light image enhancement model trained on LOL dateset (Nah, Hyun Kim, and Mu Lee 2017) for $100,000$ iterations only. Except for

Figure 7: Image desnowing example on CSD (2000) (Chen et al. 2021). Our PPTformer recovers cleaner results.

Figure 8: Low-light image enhancement example on LOL (Wei et al. 2018). Our PPTformer restores more vivid colors.

Experiment	SSIM ↑	LPIPS ↓	FLOPs (G)	Params (M)
(a) w/o Using Parser	0.9153	0.1211	48.42	9.27
(b) Parser→Degraded Image	0.9207	0.1151	168.91	20.48
(c) Ours	**0.9248**	**0.1138**	168.91	20.48

Table 5: Effect on using of parser. Compared with models without using the parser and with widely-used input degraded images as conditional modulation, our method that uses the parser to prompt the image restoration network performs the best.

Figure 9: Visual effect on parser. Using the parser helps produce more natural results (d) with vivid colors than the model without using the parser (b) or with the degraded image (c).

reporting SSIM and LPIPS metrics, we also provide the FLOPs and the number of parameters (Params) for reference, where FLOPs are computed on image size 256×256. Next, we describe the influence of each component individually.

Effect of Parser. The main design of our PPTformer is that we use the parser map generated by a large visual foundation model (Kirillov et al. 2023) to guide image restoration. One may wonder to know that if this strategy is more effective than a previous widely-used scheme that uses input images as the restoration guidance or without using any conditions. To answer this question, we conduct ablation experiments in Tab. 5. One can observe that when we do not integrate the parser to the restoration network, i.e., Tab. 5(a), or we replace the parser with degraded images, i.e., Tab. 5(b), the performance suffers from drop in terms of distortion metrics like SSIM and perceptual measurement like LPIPS, compared with our model with SAM parser guidance (Tab. 5(c)). Fig. 9 presents a visual example, where our method is able to produce results with more consistent colors (Fig. 9(d)) than the model without using parser (Fig. 9(b)) or with degraded images as conditions (Fig. 9(c)).

To better comprehend the impact of our parser on restoration, we visualize the parser features and the restoration features both before and after applying Intra and Inter Parser-Prompted Attention, as shown in Fig. 10. From Fig. 10(d), it is evident that the parser features prominently display salient

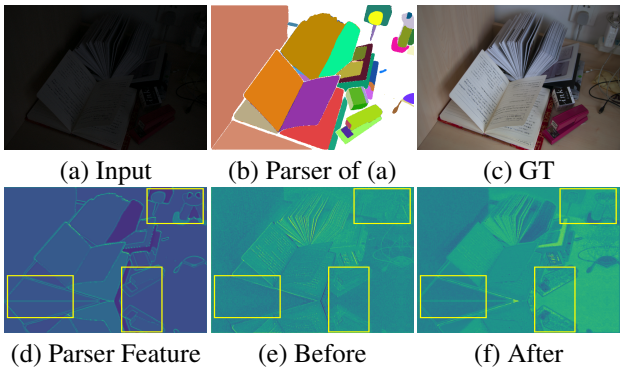
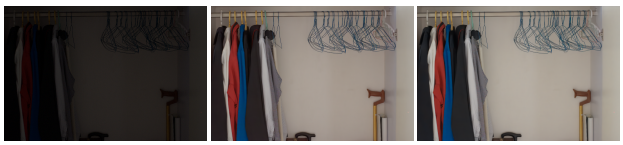


Figure 10: Visualization Understanding of the parser for guiding restoration in deep feature space. We present the averaged channel-wise features at the first IN2PPT, before and after the application of IntraPPA and InterPPA to restoration features. This comparison notably illustrates how the parser, by offering valuable structural features (d), enhances the restoration features (e), leading to significantly sharper content (f) after being prompted through the use of IntraPPA and InterPPA.

structures. However, prior to the use of IntraPPA and InterPPA, the restoration features appear notably indistinct, as seen in Fig. 10(e). Post-application of the parser's prompt-

Experiment	SSIM ↑	LPIPS ↓	FLOPs (G)	Params (M)
(a) w/o IntraPPA&InterPPA	0.9110	0.1383	129.05	17.37
(b) w/o IntraPPA	0.9152	0.1322	158.75	19.69
(c) w/o InterPPA	0.9157	0.1305	164.29	20.12
(d) Both IntraPPA	0.9176	0.1200	163.38	20.05
(e) Both InterPPA	0.9164	0.1230	174.45	20.90
(f) Ours	**0.9248**	**0.1138**	168.91	20.48

Table 6: Effect on intra and inter parser-prompted attention. Compared with models without intra and inter attention or with one of them or with both intra and inter attention, our method that utilizes intra and inter attention can implicitly and explicitly learn useful content from the parser, thus leading to the best results.



(a) Input (b) Tab. 6(a) (c) **Ours**

Figure 11: Visual effect on intra and inter parser-prompted attention. Using IntraPPA and InterPPA helps produce more natural results with vivid colors, i.e., (c), than the model without both IntraPPA and InterPPA, i.e., (b).

ing, the feature becomes significantly sharper (Fig. 10(f)), which greatly aids in restoration. These visualizations offer profound insights into our PPTformer. They demonstrate that the parser generated by the SAM (Kirillov et al. 2023) effectively provides valuable information, steering image restoration towards more refined results.

Effect of Intra and Inter Parser-Prompted Attention. We propose IntraPPA and InterPPA to implicitly and explicitly explore the parser to guide image restoration. Hence, analyzing this module is necessary. Tab. 6 summarises the ablation results. As can be seen, removing IntraPPA or InterPPA or both of them decreases the enhancement performance. On the other hand, we also use two IntraPPA or two InterPPA to replace the IntraPPA-InterPPA structure. However, we note that our default model achieves the best, adequately demonstrating the effectiveness of our proposed IntraPPA and InterPPA. The visual example presented in Fig. 11 further suggests that using IntraPPA-InterPPA helps produce more natural results.

Effect of Bidirectional Parser-Prompted Fusion. As we introduce the BiPPF module to fuse the parser features and restoration ones, we need to compare it with the widely-used feature fuse module SFT (Wang et al. 2018b). Tab. 7 shows that our proposed BiPPF significantly outperforms the SFT in terms of SSIM and LPIPS. Fig. 12 shows our BiPPF helps produce more vivid results than the model with SFT (Wang et al. 2018b).

Discussion and Limitations. To efficiently train the PPTformer, we initially generate the parser using SAM (Kirillov et al. 2023) offline. This involves first leveraging SAM to produce the parser, which is then input into the networks as an image. We acknowledge that using SAM's intermediate features might yield better restoration results than our current method. However, incorporating SAM directly into the training process would greatly increase computational demands, adversely affecting training efficiency. Additionally, while our ablation experiments confirm SAM's advantages in low-light image enhancement over models that either do not use the parser or use degraded images as input conditions, its effectiveness in other image restoration tasks like dynamic scene deblurring and image dehazing remains to be further investigated.

Experiment	SSIM ↑	LPIPS ↓	FLOPs (G)	Params (M)
(a) BiPPF→SFT	0.9154	0.1305	120.59	14.39
(b) Ours	**0.9248**	**0.1138**	168.91	20.48

Table 7: Effect on bidirectional parser-prompted fusion. Compared with the widely-used feature fusion module SFT (Wang et al. 2018b), our proposed BiPPF significantly outperforms it in terms of SSIM and LPIPS, demonstrating the effectiveness of BiPPF.

(a) Input (b) SFT (c) **Ours**

Figure 12: Visual effect on bidirectional parser-prompted fusion. Our BiPPF helps recover clearer results with more natural colors than the model with widely-used fusion module SFT (Wang et al. 2018b).

Concluding Remarks

We have proposed the PPTformer, an inter and intra Parser-Prompted Transformer, for image restoration. Our PPTformer nicely integrates the power of visual foundation models into the restoration process. To effectively fuse the parser generated by SAM (Kirillov et al. 2023) into restoration, we have proposed the inter and intra parser-prompted attention to implicitly and explicitly learn useful information to facilitate restoration. Moreover, we have suggested a bidirectional parser-prompted fusion scheme to better fuse parser features with restoration ones. Extensive experiments have demonstrated that our PPTformer outperforms state-of-the-art approaches on 4 restoration tasks, including image deraining, single-image defocus deblurring, image desnowing, and low-light image enhancement.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No.62306343), the China Postdoctoral Science Foundation (No.2024M753741), the Centre for Advances in Reliability and Safety (CAiRS) admitted under AiR@InnoHK Research Cluster.

References

Abuolaim, A.; and Brown, M. S. 2020. Defocus Deblurring Using Dual-Pixel Data. In *ECCV*.

Chen, S.; Ye, T.; Liu, Y.; Liao, T.; Ye, Y.; and Chen, E. 2022. MSP-Former: Multi-Scale Projection Transformer for Single Image Desnowing. *arXiv preprint arXiv:2207.05621*.

Chen, W.-T.; Fang, H.-Y.; Ding, J.-J.; Tsai, C.-C.; and Kuo, S.-Y. 2020. JSTASR: Joint size and transparency-aware snow removal algorithm based on modified partial convolution and veiling effect removal. In *ECCV*, 754–770.

Chen, W.-T.; Fang, H.-Y.; Hsieh, C.-L.; Tsai, C.-C.; Chen, I.; Ding, J.-J.; Kuo, S.-Y.; et al. 2021. All snow removed: Single image desnowing algorithm using hierarchical dual-tree complex wavelet representation and contradict channel loss. In *ICCV*, 4196–4205.

Cui, Y.; Tao, Y.; Bing, Z.; Ren, W.; Gao, X.; Cao, X.; Huang, K.; and Knoll, A. 2023. Selective Frequency Network for Image Restoration. In *ICLR*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

Fu, X.; Huang, J.; Ding, X.; Liao, Y.; and Paisley, J. 2017a. Clearing the skies: A deep network architecture for single-image rain removal. *TIP*.

Fu, X.; Huang, J.; Zeng, D.; Huang, Y.; Ding, X.; and Paisley, J. 2017b. Removing rain from single images via a deep detail network. In *CVPR*.

Guo, C.; Li, C.; Guo, J.; Loy, C. C.; Hou, J.; Kwong, S.; and Cong, R. 2020. Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement. In *CVPR*, 1777–1786.

Guo, C.-L.; Yan, Q.; Anwar, S.; Cong, R.; Ren, W.; and Li, C. 2022. Image Dehazing Transformer With Transmission-Aware 3D Position Embedding. In *CVPR*, 5812–5820.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Jiang, K.; Wang, Z.; Yi, P.; Huang, B.; Luo, Y.; Ma, J.; and Jiang, J. 2020. Multi-Scale Progressive Fusion Network for Single Image Deraining. In *CVPR*.

Karaali, A.; and Jung, C. R. 2017. Edge-based defocus blur estimation with adaptive scale selection. *TIP*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollár, P.; and Girshick, R. 2023. Segment Anything. In *ICCV*.

Kong, L.; Dong, J.; Ge, J.; Li, M.; and Pan, J. 2023. Efficient Frequency Domain-Based Transformers for High-Quality Image Deblurring. In *CVPR*, 5886–5895.

Lee, J.; Lee, S.; Cho, S.; and Lee, S. 2019. Deep defocus map estimation using domain adaptation. In *CVPR*.

Lee, J.; Son, H.; Rim, J.; Cho, S.; and Lee, S. 2021. Iterative Filter Adaptive Network for Single Image Defocus Deblurring. In *CVPR*.

Li, C.; Guo, C.-L.; Zhou, M.; Liang, Z.; Zhou, S.; Feng, R.; and Loy, C. C. 2023. Embedding Fourier for Ultra-High-Definition Low-Light Image Enhancement. In *ICLR*.

Li, X.; Wu, J.; Lin, Z.; Liu, H.; and Zha, H. 2018. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *ECCV*.

Liu, R.; Ma, L.; Zhang, J.; Fan, X.; and Luo, Z. 2021. Retinex-Inspired Unrolling With Cooperative Prior Architecture Search for Low-Light Image Enhancement. In *CVPR*, 10561–10570.

Liu, Y.-F.; Jaw, D.-W.; Huang, S.-C.; and Hwang, J.-N. 2018. DesnowNet: Context-aware deep network for snow removal. *TIP*, 27(6): 3064–3073.

Loshchilov, I.; and Hutter, F. 2017. SGDR: Stochastic gradient descent with warm restarts. In *ICLR*.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *ICLR*.

Lu, Z.; Xiao, Z.; Bai, J.; Xiong, Z.; and Wang, X. 2023. Can SAM Boost Video Super-Resolution? *arXiv preprint arXiv:2305.06524*.

Lv, F.; Li, Y.; and Lu, F. 2021. Attention guided low-light image enhancement with a large scale low-light simulation dataset. *IJCV*, 129(7): 2175–2193.

Ma, L.; Ma, T.; Liu, R.; Fan, X.; and Luo, Z. 2022. Toward fast, flexible, and robust low-light image enhancement. In *CVPR*, 5637–5646.

Nah, S.; Hyun Kim, T.; and Mu Lee, K. 2017. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*.

Pan, J.; Sun, D.; Pfister, H.; and Yang, M.-H. 2016. Blind image deblurring using dark channel prior. In *CVPR*.

Peng, L.; Cao, Y.; Sun, Y.; and Wang, Y. 2024a. Lightweight Adaptive Feature De-drifting for Compressed Image Classification. *IEEE TMM*.

Peng, L.; Li, W.; Pei, R.; Ren, J.; Wang, Y.; Cao, Y.; and Zha, Z.-J. 2024b. Towards Realistic Data Generation for Real-World Super-Resolution. *arXiv preprint arXiv:2406.07255*.

Purohit, K.; Suin, M.; Rajagopalan, A.; and Boddeti, V. N. 2021. Spatially-Adaptive Image Restoration using Distortion-Guided Networks. In *ICCV*.

Ren, D.; Zuo, W.; Hu, Q.; Zhu, P.; and Meng, D. 2019. Progressive image deraining networks: A better and simpler baseline. In *CVPR*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*.

Shi, J.; Xu, L.; and Jia, J. 2015. Just noticeable defocus blur detection and estimation. In *CVPR*.

Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A. P.; Bishop, R.; Rueckert, D.; and Wang, Z. 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*.

Son, H.; Lee, J.; Cho, S.; and Lee, S. 2021. Single Image Defocus Deblurring Using Kernel-Sharing Parallel Atrous Convolutions. In *ICCV*.

Tu, Z.; Talebi, H.; Zhang, H.; Yang, F.; Milanfar, P.; Bovik, A.; and Li, Y. 2022. MAXIM: Multi-Axis MLP for Image Processing. In *CVPR*, 5769–5780.

Valanarasu, J. M. J.; Yasarla, R.; and Patel, V. M. 2022. Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In *CVPR*.

Wang, C.; Pan, J.; Lin, W.; Dong, J.; Wang, W.; and Wu, X.-M. 2024a. Selfpromer: Self-prompt dehazing transformers with depth-consistency. In *AAAI*, volume 38, 5327–5335.

Wang, C.; Pan, J.; Wang, W.; Dong, J.; Wang, M.; Ju, Y.; and Chen, J. 2023. PromptRestorer: A Prompting Image Restoration Method with Degradation Perception. In *NeurIPS*.

Wang, C.; Pan, J.; Wang, W.; Fu, G.; Liang, S.; Wang, M.; Wu, X.-M.; and Liu, J. 2024b. Correlation Matching Transformation Transformers for UHD Image Restoration. In *AAAI*, volume 38, 5336–5344.

Wang, C.; Pan, J.; and Wu, X. 2022. Online-Updated High-Order Collaborative Networks for Single Image Deraining. In *AAAI*, 2406–2413.

Wang, C.; Wang, L.; Mu, J.; Yu, C.; and Wang, W. 2024c. Progressive Local and Non-Local Interactive Networks with Deeply Discriminative Training for Image Deraining. In *ACM MM*.

Wang, C.; Wang, W.; Yu, C.; and Mu, J. 2024d. Explore Internal and External Similarity for Single Image Deraining with Graph Neural Networks. In *IJCAI*.

Wang, C.; Wu, Y.; Su, Z.; and Chen, J. 2020a. Joint Self-Attention and Scale-Aggregation for Self-Calibrated Deraining Network. In *ACM MM*, 2517–2525.

Wang, C.; Xing, X.; Wu, Y.; Su, Z.; and Chen, J. 2020b. DCSFN: Deep Cross-scale Fusion Network for Single Image Rain Removal. In *ACM MM*, 1643–1651.

Wang, C.; Yu, C.; Mu, J.; and Wang, W. 2024e. PercepLIE: A New Path to Perceptual Low-Light Image Enhancement. In *ACM MM*.

Wang, X.; Yu, K.; Dong, C.; and Loy, C. C. 2018a. Recovering Realistic Texture in Image Super-Resolution by Deep Spatial Feature Transform. In *CVPR*.

Wang, X.; Yu, K.; Dong, C.; and Loy, C. C. 2018b. Recovering Realistic Texture in Image Super-Resolution by Deep Spatial Feature Transform. In *CVPR*.

Wang, Z.; Cun, X.; Bao, J.; and Liu, J. 2021. Uformer: A General U-Shaped Transformer for Image Restoration. *arXiv:2106.03106*.

Wei, C.; Wang, W.; Yang, W.; and Liu, J. 2018. Deep Retinex Decomposition for Low-Light Enhancement. In *BMVC*, 155.

Wei, W.; Meng, D.; Zhao, Q.; Xu, Z.; and Wu, Y. 2019. Semi-supervised transfer learning for image rain removal. In *CVPR*.

Wu, W.; Weng, J.; Zhang, P.; Wang, X.; Yang, W.; and Jiang, J. 2022. URetinex-Net: Retinex-Based Deep Unfolding Network for Low-Light Image Enhancement. In *CVPR*.

Xu, K.; Ma, Z.; Xu, L.; He, G.; Li, Y.; Yu, W.; Han, T.; and Yang, C. 2024a. An End-to-End Real-World Camera Imaging Pipeline. In *ACM MM*.

Xu, K.; Xu, L.; He, G.; Yu, W.; and Li, Y. 2024b. Beyond Alignment: Blind Video Face Restoration via Parsing-Guided Temporal-Coherent Transformer. In *IJCAI*.

Yang, W.; Tan, R. T.; Feng, J.; Liu, J.; Guo, Z.; and Yan, S. 2017a. Deep Joint Rain Detection and Removal from a Single Image. In *CVPR*.

Yang, W.; Tan, R. T.; Feng, J.; Liu, J.; Guo, Z.; and Yan, S. 2017b. Deep joint rain detection and removal from a single image. In *CVPR*.

Yang, W.; Wang, S.; Fang, Y.; Wang, Y.; and Liu, J. 2020. From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In *CVPR*.

Yasarla, R.; and Patel, V. M. 2019. Uncertainty guided multi-scale residual learning-using a cycle spinning cnn for single image de-raining. In *CVPR*.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2022. Restormer: Efficient Transformer for High-Resolution Image Restoration. In *CVPR*, 5718–5729.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; Yang, M.-H.; and Shao, L. 2021. Multi-Stage Progressive Image Restoration. In *CVPR*.

Zhang, H.; and Patel, V. M. 2018. Density-aware single image de-raining using a multi-stream dense network. In *CVPR*.

Zhang, H.; Sindagi, V.; and Patel, V. M. 2019. Image de-raining using a conditional generative adversarial network. *TCSVT*.

Zhao, L.; Lu, S.; Chen, T.; Yang, Z.; and Shamir, A. 2021. Deep Symmetric Network for Underexposed Image Enhancement with Recurrent Attentional Learning. In *ICCV*.