Context-aware Pretraining for Efficient Blind Image Decomposition

Chao Wang[1,2*], Zhedong Zheng[3], Ruijie Quan[1], Yifan Sun[2], Yi Yang[1†]
[1]ReLER, CCAI, Zhejiang University
[2]Baidu Inc.
[3]Sea-NExT Joint Lab, School of Computing, National University of Singapore

Abstract

*In this paper, we study Blind Image Decomposition (BID), which is to uniformly remove multiple types of degradation at once without foreknowing the noise type. There remain two practical challenges: (1) Existing methods typically require massive data supervision, making them infeasible to real-world scenarios. (2) The conventional paradigm usually focuses on mining the abnormal pattern of a superimposed image to separate the noise, which de facto conflicts with the primary image restoration task. Therefore, such a pipeline compromises repairing efficiency and authenticity. In an attempt to solve the two challenges in one go, we propose an efficient and simplified paradigm, called Context-aware Pretraining (**CP**), with two pretext tasks: mixed image separation and masked image reconstruction. Such a paradigm reduces the annotation demands and explicitly facilitates context-aware feature learning. Assuming the restoration process follows a structure-to-texture manner, we also introduce a Context-aware Pretrained network (**CPNet**). In particular, CPNet contains two transformer-based parallel encoders, one information fusion module, and one multi-head prediction module. The information fusion module explicitly utilizes the mutual correlation in the spatial-channel dimension, while the multi-head prediction module facilitates texture-guided appearance flow. Moreover, a new sampling loss along with an attribute label constraint is also deployed to make use of the spatial context, leading to high-fidelity image restoration. Extensive experiments on both real and synthetic benchmarks show that our method achieves competitive performance for various BID tasks.*

1. Introduction

Different from traditional image restoration, Blind Image Decomposition (BID) aims to remove arbitrary degradation combinations without knowing noise type and mixing mechanism. This task is challenging due to the huge gap between different noises and varying mixing patterns as amalgamated noises increase. Although many existing methods [30, 54, 61, 62] have been proposed as generic restoration networks, they are still fine-tuned on the individual datasets and do not use a single general model for all the noise removal tasks (Figure 1 (a)). All-in-one [28] further proposes a unified model across 3 datasets, but it still uses computationally complex separate encoders (Figure 1 (b)). To ameliorate this issue, TransWeather [49] introduces a single encoder-single decoder transformer network for multi-type adverse weather removal, yet this method is designed to restore one specific degradation at one time (Figure 1 (c)), which does not conform to the BID setting.

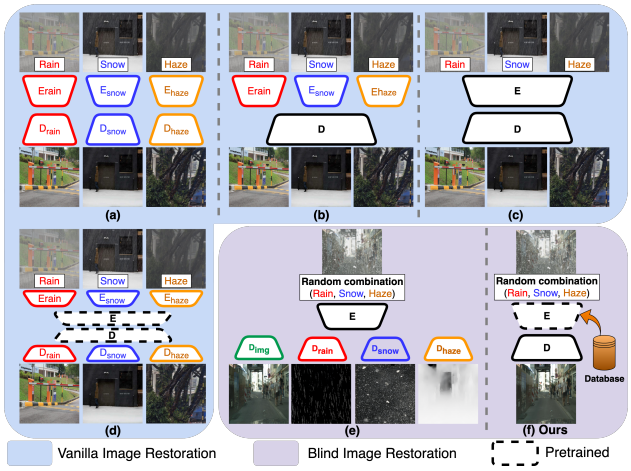

Figure 1. The prototype of restoration frameworks: (a) Traditional methods [61, 62] require task-specific network design and separate training. (b) All-in-one [28] relies on tedious one-by-one training of multiple heads. (c) TransWeather [49] is ad-hoc to remove one specific noise at a time. (d) IPT [4] extends (c) with a reusable pretrained middle-level transformer, which only works on specific tasks. (e) BIDeN [17] returns to the complex multiple decoders and demands dense supervision from noise labels. (f) The proposed method studies removing the general noise combinations by harnessing the prior knowledge learned during pretraining, which largely simplifies the pipeline. (Please zoom in to see the details.)

∗ Work done during an internship at Baidu.
† Corresponding author: Yi Yang.

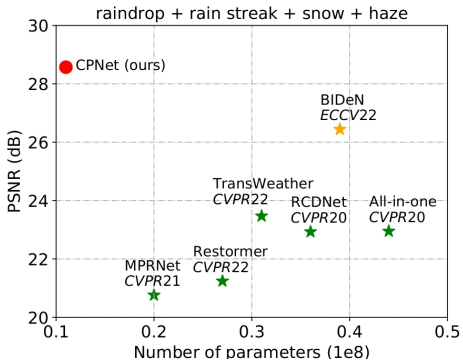

Figure 2. Comparison between the proposed method and existing approaches in terms of Peak Signal-to-Noise Ratio (PSNR) performance and the parameter numbers. We can observe that a single model instance of our method significantly outperforms both single-task and multi-task networks with much fewer parameters.

Recently, Han *et al*. [17] propose a blind image decomposition network (BIDeN) that first explores a feasible solution for BID task. This method intuitively considers a corrupted image to be composed of a series of superimposed images, and proposes a CNN-based network to separate these components. In particular, BIDeN designs a multi-scale encoder combined with separate decoders. However, this network still requires tedious training as there are multiple decoders for each component including the noise masks, which compromises the primary restoration task (Figure 1 (e)). Moreover, this deeply-learned model is data-hungry, considering the task-specific data can be limited under certain circumstances (*e.g*., medical and satellite images). Besides, various inconsistent factors (*e.g*., camera parameters, illumination, and weather) can further perturb the distribution of the captured data for training. In an attempt to address the data limitation, IPT [4] first explores the pretraining scheme on several image processing tasks. However, since the middle transformer does not learn the shared representative features, IPT is still limited to task-specific fine-tuning with complex multiple heads/tails and fails to BID mission (Figure 1 (d)).

This paper aims at addressing the aforementioned challenges as a step toward an efficient and robust real-world restoration framework. Inspired by successes in Masked AutoEncoders (MAE), where the pretraining model on ImageNet can efficiently adapt to the high-level representative vision benchmarks such as recognition and detection [18, 57], we argue that pretraining is still a potential solution for BID task. We also note that pretraining on MAE in low-level vision tasks is still under-explored. To fill this gap, we resort to model pretraining via self-supervised learning to acquire sufficient representative priors for BID.

In this paper, we propose a new Context-aware Pretraining (**CP**), containing separation and reconstruction for corrupted images. As shown in Figure 3, the pretext task is designed as a dual-branch pattern combining mixed image separation and masked image reconstruction. Our intuition underlying the proposed task is encouraging the network to mine context information (*i.e*., noise boundaries and types, local and non-local semantics), and such knowledge can be easily transferred to various restoration scenarios. We also develop a pretrained model for BID using the transformer architecture, namely, Context-aware Pretrained Network (**CPNet**). In contrast to previous methods, the proposed CPNet can (1) remove arbitrary types or combinations of noises at once, (2) avoid multi-head mask supervision of each source component (Figure 1 (f)), and (3) be efficiently employed for high-fidelity restoration after fine-tuning, as shown in Figure 2. To our knowledge, this work provides the first framework to apply a self-supervised pretraining learning strategy for BID task. Meanwhile, these two branches are hierarchically connected with an information fusion module that explicitly facilitates the feature interaction through multi-scale self-attention. Furthermore, we empirically subdue the learning difficulty by dividing the restoration process into structure reconstruction during pretraining and texture refinement during fine-tuning. Instead of simply learning the mixed pattern and proportionally scaling the pixel values in previous methods, our method intuitively gives the model more "imagination" and thus leads to a more compelling and robust performance under complex scenes. Moreover, a novel flow sampling loss combing with a conditional attribute loss is further introduced for precise and faithful blind image decomposition.

Overall, our contributions are summarized as follows:

- Different from existing BID works, we introduce a new self-supervised learning paradigm, called Context-aware Pretraining (**CP**) with two pretext tasks: mixed image separation and masked image reconstruction. To facilitate the feature learning, we also propose Context-aware Pretrained Network (**CPNet**), which is benefited from the proposed information fusion module and multi-head prediction module for texture-guided appearance flow and conditional attribute label.
- Extensive experiments on BID benchmarks substantiate that our method achieves competitive performance for blind image restoration. More importantly, our method consistently outperforms competitors by large margins in terms of efficiency, *e.g*., 3.4 $\times$ fewer FLOPs and 50 $\times$ faster inference time over BIDeN [17].

2. Related work

Blind image decomposition. Aiming at the single-task limitation, several restoration works [13, 17, 70] have discussed the emerging image decomposition task, regarding raindrops and other real-world corruptions as superimposing and separable to a clean image. Gandelsman

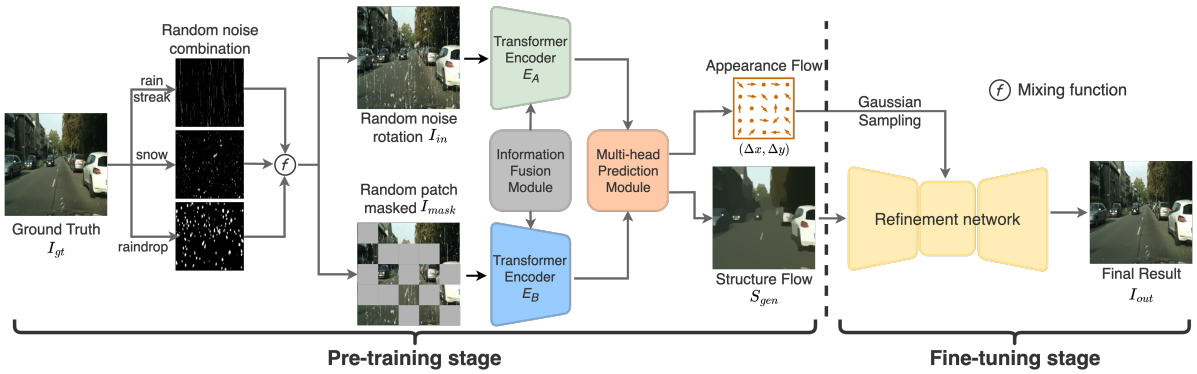

Figure 3. Overview of the proposed context-aware pretraining framework. (1) During the pretraining stage, we add an arbitrary degradation combination with random rotation to the clean image. Taking this corrupted image as input, two parallel transformer encoders are deployed to simultaneously perform masked image decomposition and reconstruction. In order to learn more context-aware prior knowledge, features of both encoders first interact through an information fusion module. Then a multi-flow prediction head is designed to separately produce the repaired structure and a texture-guided appearance map. (2) During the fine-tuning stage, we fix the pretrained two-branch encoder and only train a parameter-efficient refinement network from scratch on specific paired datasets.

et al. first propose a unified framework Double-DIP [13] for layer decomposition based on coupled DIP [48] networks. DAD [70] further introduces three discriminators and a crossroad L_1 loss for more complex mixture conditions. Based on blind source separation problem [15,16,27], Han *et al*. further propose the "Blind Image Decomposition" (BID) [17] setting, which treats degraded images as an arbitrary combination of individual components, and aims to solve multi-type degradation at once. The training paradigm of BID, however, stays heavily dependent on time-consuming end-to-end reconstruction. In comparison, our pretrained method only relies on simple texture fine-tuning and requires no extra auxiliary labels.

Self-supervised learning for image restoration. Supervised learning requires massive paired references that map input measurements to the clean image [20, 21, 41, 63, 65], which is hard to be satisfied when the noise model is unknown. To overcome this problem, several self-supervised approaches [6, 47, 50] have been proposed based on different pretext tasks. DIP [48] first proves that a simple generator can sufficiently reconstruct low-level image statistics priors, which leads to a diversity of models that avoid the modeling of the degradation process [2, 24, 26]. However, limited by the variations across different degradation, these methods can only repair a single task or require repetitive training. In contrast, our method divides this process into two pretext tasks, *i.e*., locate and generate, thus removing various noises at once with a unified model.

Pretrained vision transformers. Recently, transformer has been adapted to numerous vision tasks such as recognition [10, 60] and segmentation [56, 66]. Due to the impressive performance, it has also been introduced for low-level vision problems such as image restoration [30, 54, 61]. To further utilize the prior knowledge learned from transformer, IPT [4] presents a universal pretraining scheme, yet this method still requires complex multi-head training. Several works [18, 57] have also leveraged the masked image modeling (MIM) [1, 5, 69] paradigm and explored a generative pretrained framework for high-level representation learning. However, few related works focus on utilizing self-supervised pretrained transformer for low-level vision tasks. In contrast to the previous works, we aim for combining self-supervised image decomposition with masked reconstruction priors to facilitate more general and efficient blind image restoration.

3. Method

In this work, we present a dual-path pretraining framework that contains two parallel transformer encoders, an information fusion module, and a multi-head prediction module, as shown in Figure 3. In order to avoid early information leakage during the feature fusion, we perform random masking on the noisy image patches instead of the clean image as another pretext task. The details of the proposed modules and objective functions are expressed below.

3.1. Information Fusion Module

In this work, we encourage the separation branch to learn more spatial information by locating the superimposed noises, while the reconstruction branch exploits more generating priors by predicting the masked token with neighboring patches. Thus we deliberately facilitate the spatial sensitivity of E_A as well as the content generativity of E_B, building a sophisticated trade-off between two networks. To explicitly leverage and further enhance this balance, we proposed a multi-dimensional fusion module, which consists of the feature interaction block (FIBlock), refinement block (FRBlock) and enhancement block (FEBlock).

Feature interaction block. Although the two parallel encoders are supposed to focus on their respective tasks, they

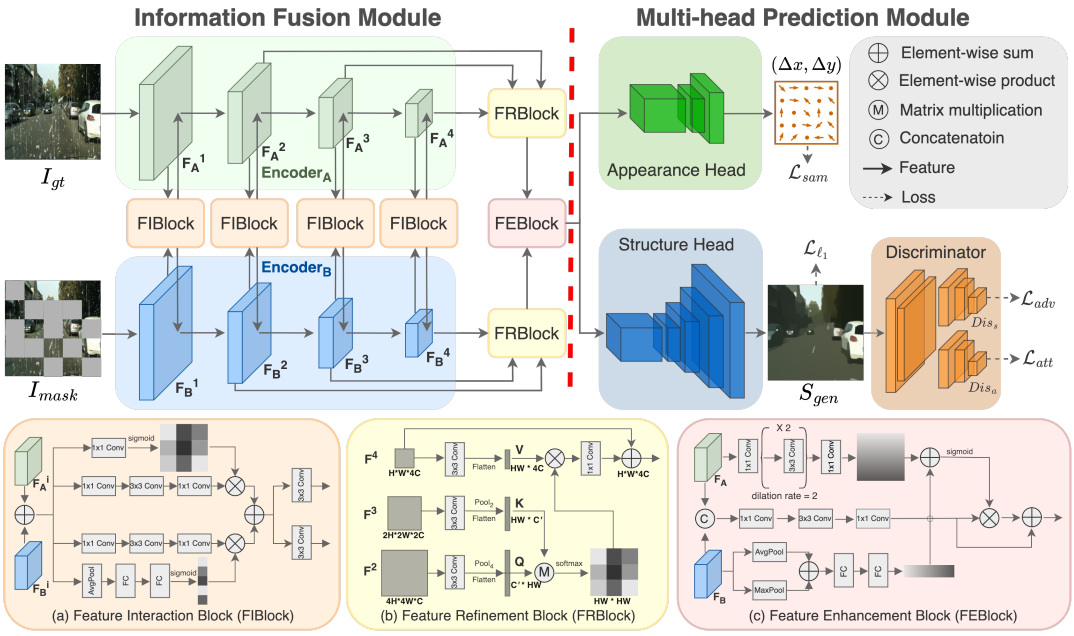

Figure 4. The architecture of our information fusion (left) and multi-head prediction (right) modules. F_A^i and F_B^i represent the features from the i-th layer of E_A and E_B respectively, which are the inputs of FIBlock. Then the features $F^i (i = 2, 3, 4)$ from several layers of each encoder are further sent to the FRBlock. After refinement, features from two encoders are sent to FEBlock for targeted enhancement.

share some mutual correlation during training. For example, the spatial features of noises learned from E_A can reversely indicate the uncorrupted content for E_B and vice versa. Therefore, we introduce the FIBlock with both channel and spatial attention to boost the feature interaction, as shown in Figure 4 (a). We also deploy the residual bottleneck [11, 19] structure considering efficiency.

Feature refinement block. To further refine the features learned from two encoders, the FRBlock is proposed based on a self-attention mechanism [53]. The detail of FRBlock is shown in Figure 4 (b), which can be formulated as:

$$\begin{aligned} \mathbf{Q} &= \mathrm{Pool}_4(\mathrm{Conv}(\boldsymbol{F}^2)), \\ \mathbf{K} &= \mathrm{Pool}_2(\mathrm{Conv}(\boldsymbol{F}^3)), \\ \mathbf{V} &= \mathrm{Conv}(\boldsymbol{F}^4)), \\ \mathrm{Att}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) &= \mathrm{SoftMax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}, \end{aligned} \tag{1}$$

where d is the dimension of the query set $\mathbf{Q}$, Pool_a is the average pooling operation with both kernel size and stride a, and $\boldsymbol{F}^i$ represents the feature map from the i-th layer of E_A or E_B. For efficient non-local computation while fusing the pyramid representations, feature maps from higher layers are accordingly pooled and then directly input into the FRBlock after an individual convolution layer, which significantly reduces the computation and memory intensity while improving the robustness.

Feature enhancement block. According to the inversion theory [3, 37], it can be known that different objects usually correspond to different channels (convolution kernels). Based on this observation, we further introduce the FEBlock to empirically enhance the spatial correlation of E_A features, as well as the channel features of E_B. We adopt the Bottleneck Attention Module [38] and further improve it with multiple pooling [55] and residual connection (Figure 4 (c)). Notably, we cat features instead of adding them since both encoders have interacted with each other.

Transformer encoder. Since the proposed method mainly focuses on the feature processing between two tasks, the encoder selection can be relatively trivial. In this paper, we adopt the same setting as Restormer [61], which contains many lightweight modules to improve efficiency. Notice that the transformer design in our framework could be any cutting-edge combinations such as swin structure [33]. More discussion can be found in the **suppl**.

3.2. Multi-head Prediction Module

To coordinate with the dual-path pretraining framework, we perform the multi-head prediction respectively on structure map and appearance flow, as shown in Figure 4.

Structure flow. In vision tasks, the decoder reconstructs pixels that have lower-level semantics compared to common recognition tasks, which means the network design is crucial for determining the semantic level of the learned latent representations [18]. Driven by this analysis, we design a simple yet effective structure head, which consists of a resblock [19] followed by several convolution layers. To further simplify the training objective, we adopt an edge-

preserved smooth method [58] to remove high-frequency textures of I_{gt} while retaining the global structures. The reconstruction loss is defined as the ℓ_1 distance between the predicted structures S_{gen} and the ground truth structures S_{gt} smoothed from I_{gt}:

$$\mathcal{L}_{\ell_1} = \|S_{gen} - S_{gt}\|_1 . \tag{2}$$

Meanwhile, to mimic the distributions of the target structures S_{gt}, we further apply the generative adversarial framework [14]. The adversarial loss can be written as:

$$\mathcal{L}_{adv} = \mathbb{E}\left[\log\left(1 - Dis_s\left(S_{gen}\right)\right)\right] + \mathbb{E}\left[\log Dis_s\left(S_{gt}\right)\right], \tag{3}$$

where Dis_s is the discriminator of the structure head.

Conditional learning. Inspired by the image translation works [7, 68], BID can also be regarded as an attribute editing task with an initial random one-hot attribute label. Thus we design a discriminator branch Dis_a for the conditional attribute classification task:

$$\mathcal{L}_{att} = -\sum_{i=1}^{N} \log P_i\left(S_{gen} \mid \theta_{Dis_a}\right). \tag{4}$$

Here $P_i(x)$ represents the probability that x belongs to the i-th attribute (noise type), which is predicted by Dis_a with the parameter θ_{Dis_a}. In contrast to previous multi-head methods [4, 17, 28], this conditional discriminator implicitly enables multiple noise restoration with a unified structure, leading to higher flexibility and training efficiency. Meanwhile, we also show that the unified attribute mechanism can selectively remove arbitrary degradation types by simply specifying the attribute label.

Appearance flow. While obtaining the reconstructed structure S_{gen}, an appearance flow head is further deployed to warp the extracted features of the inputs, as shown in Figure 4. Based on the appearance flow in [43], we further introduce a new sampling loss to simultaneously facilitate the local texture propagation and global structure calibration, which can be formulated as follow:

$$\mathcal{L}_{sam} = \frac{1}{N} \sum_{(x,y)\in\Omega} \exp\left(-\frac{\mu\left(\mathbf{\Phi}_{I_{gt}}^{x,y}, \mathbf{\Phi}_{I_{in}}^{x+\Delta x, y+\Delta y}\right)}{\alpha\left\|\mathbf{\Phi}_{S_{gt}} - \mathbf{\Phi}_{S_{gen}}\right\|_1 + \epsilon}\right), \tag{5}$$

where $\mathbf{\Phi}_{I_{gt}}$ and $\mathbf{\Phi}_{I_{in}}$ are the features generated by a specific layer of VGG19 on the ground truth image I_{gt} and the input corrupted image I_{in}. $(\Delta x, \Delta y)$ represents the predicted coordinate offset from the appearance head. $\mu(*)$ denotes the cosine similarity and Ω denotes a coordinate set containing all valid coordinates in $\mathbf{\Phi}_{I_{in}}$. N is the number of elements in set Ω. Since the positions of the random noises are unknown, our flow sampling loss uniformly calculates the relative cosine similarity between the ground truth features and the sampled features in each region. However, the appearance flow training without the mask supervision may struggle to capture global dependency and stuck in a bad local minima [34, 42]. To tackle this problem, we further impose a global constraint between structure S_{gt} and S_{gen} as an additional normalization term. α represents the scaling parameter fixed as 10 and ϵ is a constant term. We also adopt Gaussian sampling [43] to expand the receptive field.

In this way, the appearance flow map is enforced to determine whether the current sampled region is uncorrupted, and which damaged block matches this region best in texture. Combing with learnable relative positional embedding [46], the position calibrations of noises are further ensured and features containing vivid textures can "flow" to the corrupted regions. The texture head shares similar structures with the structure head. More details on model structure and sampling operation are provided in the **suppl**.

3.3. Fine-tuning and Optimization

During pretraining, we jointly optimize the parallel encoders $\mathbf{E}$, multi-head decoders $\mathbf{H}$ and multi-head discriminators $\mathbf{Dis}$ to the objective, which is a weighted sum of the following losses:

$$\mathcal{L}_{total}(\mathbf{E}, \mathbf{H}, \mathbf{Dis}) = \lambda_{\ell_1}\mathcal{L}_{\ell_1} + \mathcal{L}_{adv} + \mathcal{L}_{att} + \lambda_{sam}\mathcal{L}_{sam}, \tag{6}$$

where λ_{ℓ_1} and λ_{sam} are regularization parameters. In our experiments, we set λ_{ℓ_1} = 4 and λ_{sam} = 0.25. After pretraining, we can simply fine-tune an autoencoder-like network to generate more detailed textures. The loss functions during fine-tuning are composed of a standard ℓ_2 reconstruction loss and a perceptual loss [22]. We also test to unlock the multi-head prediction part for better performance. More details on finetuning are given in the **suppl**.

4. Experiment

We conduct extensive experiments to show the effectiveness of our proposed method. In what follows, we explain the experimental settings, implementation details, comparison with state-of-the-art methods and ablation studies.

4.1. BID Tasks and Datasets

Without loss of generality, we adopt the large-scale ImageNet dataset [44] as the pretraining set, which contains over 1M images with 1K scenes. We generate the corrupted images with random mixed combinations of total 7 types of degradation: rain-streak [29, 59], raindrop [39], snow [32], haze [45], shadow [40, 51], reflection [64] and watermark [31]. Following the similar settings in BIDeN [17], we evaluate the performance under three most common noise combinations: **I:** Joint raindrop/rainstreak/snow/haze removal, **II:** Real-world bad weather removal, and **III:** Joint shadow/reflection/watermark removal.

Table 1. Quantitative results of Task I in driving scenario. We evaluate the performance in Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) under 6 BID cases, which are (1): rain streak, (2): rain streak + snow, (3): rain streak + light haze, (4): rain streak + heavy haze, (5): rain streak + moderate haze + raindrop, (6) rain streak + snow + moderate haze + raindrop. The best performance under each case is marked in **bold** with the second performance underlined.

Case	Input		MPRNet [62]		Restormer [61]		All-in-one [28]		BIDeN [17]		Ours	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
(1)	25.69	0.786	33.39	0.945	**34.29**	**0.951**	32.38	0.937	30.89	0.932	33.95	0.948
(2)	18.64	0.564	30.52	0.909	30.60	0.917	28.45	0.892	29.34	0.899	**33.42**	**0.937**
(3)	17.45	0.712	23.98	0.900	23.74	0.905	27.14	0.911	28.62	0.919	**32.99**	**0.932**
(4)	11.12	0.571	18.54	0.829	20.33	0.853	19.67	0.865	26.77	0.891	**29.02**	**0.908**
(5)	14.05	0.616	21.18	0.846	22.17	0.859	24.23	0.889	27.11	0.898	**30.07**	**0.925**
(6)	12.38	0.461	20.76	0.812	21.24	0.821	22.93	0.846	26.44	0.870	**29.57**	**0.914**

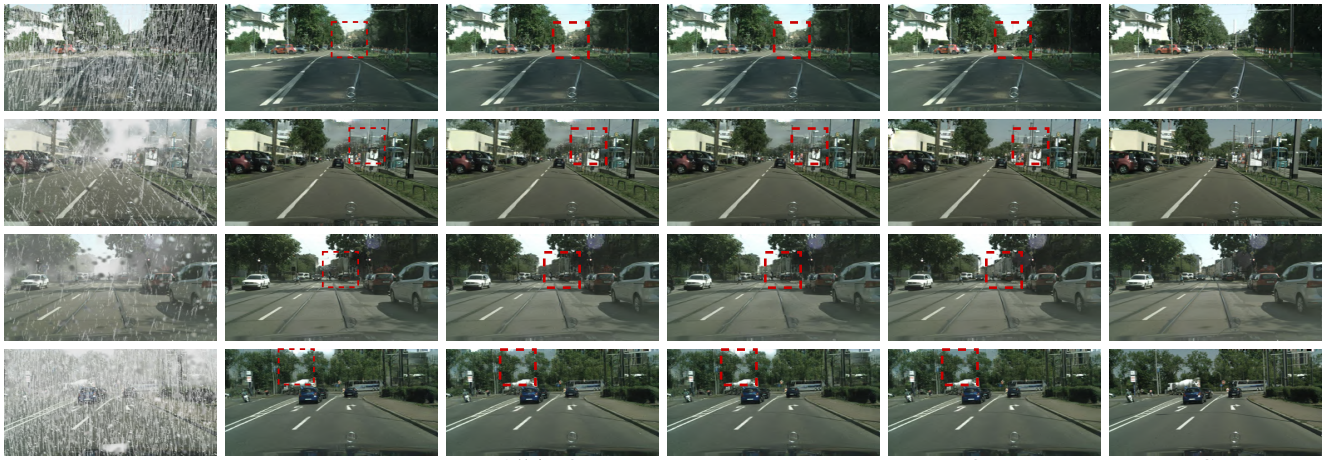

Figure 5. Qualitative results of Task I in driving scenario under several mixed cases. Row 1-4 represents the cases (3)-(6) respectively as presented in Table 1. For all cases, our model can produce more precise and faithful images. (Please zoom in to see the details.)

4.2. Implementation Details

Throughout the pretraining stage on ImageNet, we use 4 NVIDIA Tesla V100 GPUs with the conventional Adam optimizer [23] with $\beta_1 = 0.9$, $\beta_2 = 0.999$ for both E_A and E_B. The model is pretrained for 80 epochs on ImageNet [44] with an initial learning rate of $5e^{-4}$ and decayed to $2e^{-4}$ after 50 epochs with 32 batch size. After pretraining, we fine-tune the refinement model on the BID dataset [17] for 30 epochs with a learning rate of $3e^{-4}$. Random crop and horizontal flip are randomly applied as data augmentation. For a fair comparison, we deploy the same evaluation settings as in BIDeN [17]. More details on noise construction, dataset and experimental settings are given in **suppl**.

Reproducibility. We will release both Pytorch and Paddlepaddle version implementation at [1].

4.3. Comparison with SOTA

Task I: We first perform the BID task on CityScape [8] for the driving scenario. Experiments are conducted with both single-task and multi-task methods under the same BID settings. Table 1 and Figure 5 show the comparison between our method and baselines. It can be seen that the task-specific method MPRNet [62] and Restormer [61] perform well with single-type noise in cases (1) and (2), but the performance drops rapidly in more complex situations. All-in-one [28] performs slightly better since the multi-head encoder learns more universal features during the BID training. BIDeN [17] is able to handle more complex cases, while the performance in simple scenarios such as cases (1) and (2) is limited by the BID training setting and insufficient feature learning. In contrast, our proposed method consistently achieves competitive performance in all cases benefiting from the context-aware learning of parallel encoders. More importantly, our method maintains higher generality under complex conditions, *e.g.*, 29.57 PSNR for case (6) is still higher than 29.34 PSNR of BIDeN in case (2). More results under each case are provided in the **suppl.**

Task II: To further verify the generalization performance of the proposed method, we also conduct experiments on real-world nature images with the identical rainstreak, raindrop and snow masks from Task I. We separately train the BIDeN [17] under 3 cases, which are (1): task-specific mask, (2) rainstreak + raindrop, (3) rainstreak + raindrop + snow. Then we test these trained models on each single-

[1] https://github.com/Oliiveralien/CPNet

Table 2. Quantitative results of Task II.B. (1)-(3) indicate that models are trained under different settings. The best performances are marked in **bold** with the second performance underlined. Performance variations between cases (1) and (3) are marked in **blue**.

Method		Input	MPRNet	BIDeN (1)	BIDeN (2)	BIDeN (3)	Ours (1)	Ours (2)	Ours (3)
Rainstreak	NIQE ↓	4.87	**4.10**	4.15	4.28	4.33 **(+0.18)**	4.12	4.12	4.13 **(+0.01)**
	BRISQUE ↓	27.82	28.66	25.76	26.19	26.57 **(+0.81)**	**25.53**	25.57	25.58 **(+0.05)**
Raindrop	NIQE ↓	5.63	4.87	4.55	4.67	4.72 **(+0.17)**	**4.48**	4.59	4.50 **(+0.02)**
	BRISQUE ↓	24.88	29.17	20.29	20.82	21.22 **(+0.93)**	**20.08**	20.11	20.16 **(+0.08)**
Snow	NIQE ↓	4.75	4.48	4.21	4.25	4.31 **(+0.10)**	**4.14**	4.15	4.16 **(+0.02)**
	BRISQUE ↓	22.68	25.78	21.99	22.25	22.42 **(+0.43)**	**21.83**	21.85	21.88 **(+0.05)**

Table 3. Quantitative results of Task III. We evaluate the Root Mean Square Error (RMSE ↓) in LAB color space. The best performances are marked in **bold** with the second performance underlined. Performance variations between cases (1) and (3) are marked in **blue**.

RMSE	DHAN	Auto-Exposure	BIDeN (1)	BIDeN (2)	BIDeN (3)	Ours (1)	Ours (2)	Ours (3)
Shadow	8.94	**8.56**	12.01	14.15	15.49 **(+2.14)**	8.65	8.70	8.76 **(+0.11)**
Non-Shadow	**4.80**	5.75	7.52	8.21	8.93 **(+2.46)**	4.98	4.98	4.99 **(+0.01)**
All	**5.67**	6.51	8.77	9.85	10.69 **(+3.34)**	5.97	5.99	6.03 **(+0.06)**

Figure 6. Visual comparisons for real-world rainstreak removal on SPAData [52] for task II. BIDeN(1) means a two-head network trained with a single-type rainstreak mask and BIDeN(3) means the four-head structure trained with rainstreak + raindrop + snow.

type degradation, respectively. Both training and test sets are obtained from BIDeN [17]. Performance is evaluated with no-reference metrics NIQE [36] and BRISQUE [35]. A task-specific method MPRNet [62] is also deployed as the baseline. As shown in Table 2 and Figure 6, both BIDeN and our method perform well on single-task restoration. However, a significant performance drop can be observed for BIDeN when applying the multi-head trained model for one specific task, *e.g.*, performance drops 0.93 (from 20.29 to 21.22) in terms of BRISQUE for raindrop removal. This is mainly because the multiple objective regression during training stems from more representative learning of the encoder. In contrast, the proposed method remains stable on each task, indicating the training robustness under various noise combinations. Please refer to **suppl** for more results.

Task III: We also conducted experiments on other degradation following a similar setting in Task II. We compare the shadow removal results to BIDeN and several task-specific shadow removal baselines [9, 12] on the SRD [40] dataset, as shown in Table 3 and Figure 7. Similar performance drop on multi-head BIDeN also appears in single-type shadow removal task, while our method consistently remains comparable even with task-specific methods. We also visualize the multi-head outputs in Figure 7. It can be observed that our pretrained model can produce a restored structure. Meanwhile, the appearance flow can not only find the noise location but also indicate the sampling direction of textural features. More visual comparisons are given in **suppl**.

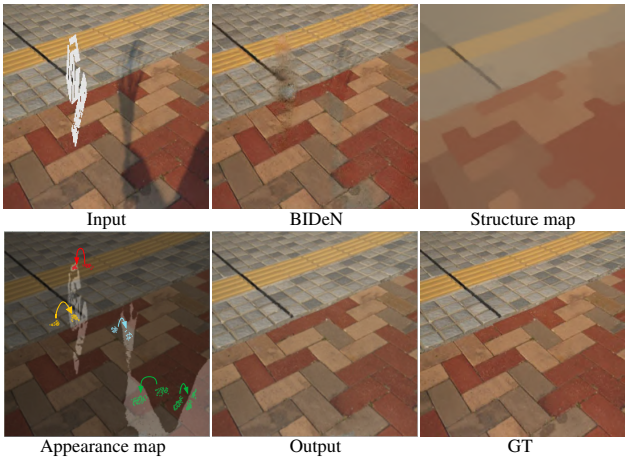

Figure 7. Visualizations of the outputs during pretraining and fine-tuning. To visualize the appearance flow fields, we plot part of the sample points of typical missing regions. The arrows show the direction of the appearance flow. Please zoom in to see the details.

4.4. Discussions

To explore the role of two encoders in joint learning, we also visualize the activation maps of features from both encoders as in Figure 8, it can be clearly observed that E_A mainly focus on the mask regions for better repairing integrity, while E_B pays more attention on the global context for better repairing authenticity. More importantly, benefiting from the deployment of the attribute label on mask type, our proposed method can remove a specified type of degradation while leaving other noises unchanged.

Ablation study. We conduct an ablation study to understand the contributions of individual components proposed in this paper. The quantitative results can be found in Ta-

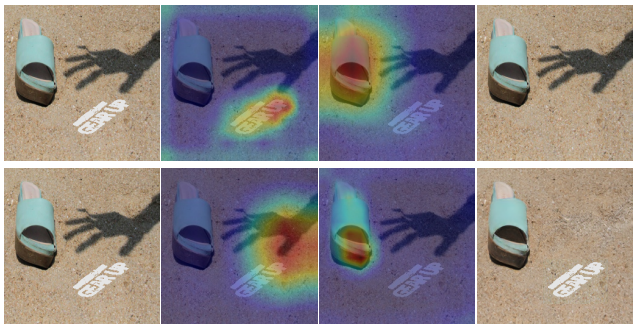

Figure 8. Attention visualizations. Mask attention represents the feature activation map in E_A, while the context attention comes from E_B . The top row shows the outputs for watermark removal and the bottom row shows shadow removal results.

Table 4. Ablation study on the rainstreak + raindrop + snow masks. PSNR is computed on the same dataset in Task II.

Method	$\mathcal{L}^t_{sam}$	$\mathcal{L}^t_{con}$	FIBlock	FRBlock	FEBlock	PSNR
without E_A	✓	✓		✓	✓	18.25
without E_B	✓	✓		✓	✓	21.36
E_A + E_B		✓	✓	✓	✓	25.84
E_A + E_B	✓		✓	✓	✓	27.23
E_A + E_B	✓	✓	✓			27.47
E_A + E_B	✓	✓	✓	✓		28.29
Full model	✓	✓	✓	✓	✓	28.57

ble 4. We start with the joint pretraining scheme by disabling each encoder and the FIBlock. It can be concluded from Table 4 that high-quality bind image decomposition requires both the parallel encoders to jointly learn complementary spatial semantic information. The restored images tend to retain some missing noises without E_A while producing more artifacts without E_B. We also conduct ablations on information fusion modules and loss functions, it can be observed that each individual contribution to this work helps in improving the performance.

Pre-training dataset. we also explore the possibility of further accelerating pretraining with customized datasets. First, we pre-train and fine-tune the model on the same BID dataset. As shown in Table 5, we find that pretraining on a larger and diverse dataset like ImageNet can achieve better performance. Second, we randomly selected 10% images from each class of ImageNet for pretraining on Task I. It can be observed that our pretraining method still arrives at a competitive performance. Finally, We selected 100 kinds of scene images in ImageNet for pretraining such as cars and buildings. We also construct another pretraining dataset with 100 kinds of irrelevant object images such as sport and food. We find that if the data distribution of the pretraining dataset is close to the target finetuning dataset, the performance will be significantly better. This phenomenon further brings up an interesting potential direction. We can collect a smaller but general dataset with reasonable classes for saving total training time costs in the future.

Table 5. Ablation on the pre-training dataset for Task I case (5).

Dataset	BIDeN	only BID	10% ImageNet	Scene class	Object class	Full ImageNet
PSNR↑	27.11	26.80	28.95	28.43	24.89	**30.07**

Table 6. Discussion on the model efficiency. All models are tested under the same environment for fair comparisons.

Method	MPRNet	All-in-one	BIDeN	Ours
Param (M)	21.15	44.26	38.61	11.30
FLOPs (G)	135	350	344	102
Inference time (s)	0.21	0.34	13.21	0.26

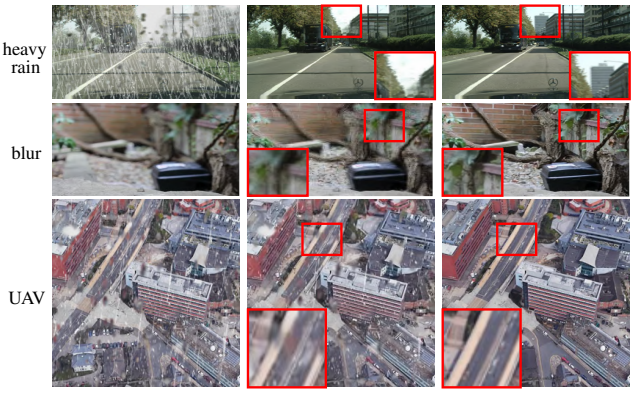

Figure 9. Failure cases. Please zoom in to see the details.

Efficiency. In Figure 2, we compare the number of parameters. We also compare the FLOPs and inference time efficiency in Table 6. The reported time corresponds to the average time each model takes with test images of dimensions 256 × 256 during the inference stage. We note that our method is much faster (over 50× than BIDeN) than the contemporary SOTA method BIDeN.

Limitations. It should be noted that the proposed method mainly focuses on common superimposing scenes with additive noises, which may lead to semantic distortion in complex background-attached scenes (*e.g.*, heavy rain, defocus deblurring [25] or UAV images [67]), as shown in Figure 9.

5. Conclusion

In this paper, we propose a new context-aware pretraining paradigm (CP) for the BID task. Different from previous methods, we shed light on the possibilities of self-supervised pretraining to remove multiple general noises in one go. During pretraining, the CPNet model is designed with two entangled encoders serving different image processing tasks, *i.e.*, mixed image separation and masked image reconstruction, for joint context-aware learning. Experiments on seven representative restoration tasks and three BID tasks demonstrate that CPNet consistently facilitates state-of-the-art performance in terms of both image restoration quality and efficiency.

Acknowledgements. This work was supported by the Fundamental Research Funds for the Central Universities (No. 226-2022-00051).

References

[1] H. Bao, L. Dong, and F. Wei. Beit: Bert pre-training of image transformers. *arXiv:2106.08254*, 2021. 3

[2] J. Batson and L. Royer. Noise2self: Blind denoising by self-supervision. In *ICML*, 2019. 3

[3] D. Bau, J. Y. Zhu, H. Strobelt, B. Zhou, J. B. Tenenbaum, W. T. Freeman, and A. Torralba. Gan dissection: Visualizing and understanding generative adversarial networks. *arXiv:1811.10597*, 2018. 4

[4] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao. Pre-trained image processing transformer. In *CVPR*, 2021. 1, 2, 3, 5

[5] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever. Generative pretraining from pixels. In *ICML*, 2020. 3

[6] M. Chen, Z. Zheng, Y. Yang, and T. S. Chua. Pipa: Pixel- and patch-wise self-supervised learning for domain adaptative semantic segmentation. *arXiv:2211.07609*, 2022. 3

[7] Y. Choi, M. Choi, M. Kim, J. W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018. 5

[8] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 6

[9] X. Cun, C. Pun, and C. Shi. Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting gan. In *AAAI*, 2020. 7

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3

[11] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu. Dual attention network for scene segmentation. In *CVPR*, 2019. 4

[12] L. Fu, C. Zhou, Q. Guo, F. Juefei-Xu, H. Yu, W. Feng, Y. Liu, and S. Wang. Auto-exposure fusion for single-image shadow removal. In *CVPR*, 2021. 7

[13] Y. Gandelsman, A. Shocher, and M. Irani. " double-dip": Unsupervised image decomposition via coupled deep-image-priors. In *CVPR*, 2019. 2, 3

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 5

[15] S. Gu, D. Meng, W. Zuo, and L. Zhang. Joint convolutional analysis and synthesis sparse representation for single image layer separation. In *ICCV*, 2017. 3

[16] R. Guo, Q. Dai, and D. Hoiem. Paired regions for shadow detection and removal. *TPAMI*, 35(12):2956–2967, 2012. 3

[17] J. Han, W. Li, P. Fang, C. Sun, J. Hong, M. A. Armin, L. Petersson, and H. Li. Blind image decomposition. In *ECCV*, 2022. 1, 2, 3, 5, 6, 7

[18] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 2, 3, 4

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[20] Z. Hu, Y. Sun, and Y. Yang. Switch to generalize: Domain-switch learning for cross-domain few-shot classification. In *ICLR*, 2022. 3

[21] Z. Hu, Y. Sun, and Y. Yang. Suppressing the heterogeneity: A strong feature extractor for few-shot segmentation. In *ICLR*, 2023. 3

[22] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 5

[23] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014. 6

[24] A. Krull, T. O. Buchholz, and F. Jug. Noise2void-learning denoising from single noisy images. In *CVPR*, 2019. 3

[25] J. Lee, H. Son, J. Rim, S. Cho, and S. Lee. Iterative filter adaptive network for single image defocus deblurring. In *CVPR*, 2021. 8

[26] J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila. Noise2noise: Learning image restoration without clean data. In *ICML*, 2018. 3

[27] A. Levin and Y. Weiss. User assisted separation of reflections from a single image using a sparsity prior. *TPAMI*, 29(9):1647–1654, 2007. 3

[28] R. Li, R. T. Tan, and L. F. Cheong. All in one bad weather removal using architectural search. In *CVPR*, 2020. 1, 5, 6

[29] S. Li, I. B. Araujo, W. Ren, Z. Wang, E. K. Tokuda, R. H. Junior, R. Cesar-Junior, J. Zhang, X. Guo, and X. Cao. Single image deraining: A comprehensive benchmark analysis. In *CVPR*, 2019. 5

[30] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte. Swinir: Image restoration using swin transformer. In *ICCV*, 2021. 1, 3

[31] Y. Liu, Z. Zhu, and X. Bai. Wdnet: Watermark-decomposition network for visible watermark removal. In *WACV*, 2021. 5

[32] Y. F. Liu, D. W. Jaw, S. C. Huang, and J. N. Hwang. Desnownet: Context-aware deep network for snow removal. *TIP*, 27(6):3064–3073, 2018. 5

[33] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 4

[34] Z. Liu, R. A. Yeh, X. Tang, Y. Liu, and A. Agarwala. Video frame synthesis using deep voxel flow. In *ICCV*, 2017. 5

[35] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *TIP*, 21(12):4695–4708, 2012. 7

[36] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 7

[37] X. Pan, X. Zhan, B. Dai, D. Lin, C. C. Loy, and P. Luo. Exploiting deep generative prior for versatile image restoration and manipulation. *TPAMI*, 44(11):7474–7489, 2021. 4

[38] J. Park, S. Woo, J. Y. Lee, and I. S. Kweon. Bam: Bottleneck attention module. In *BMVC*, 2018. 4

[39] R. Qian, R. T. Tan, W. Yang, J. Su, and J. Liu. Attentive generative adversarial network for raindrop removal from a single image. In *CVPR*, 2018. 5

[40] L. Qu, J. Tian, S. He, Y. Tang, and R. W. Lau. Deshadownet: A multi-context embedding deep network for shadow removal. In *CVPR*, 2017. 5, 7

[41] R. Quan, X. Yu, Y. Liang, and Y. Yang. Removing raindrops and rain streaks in one go. In *CVPR*, 2021. 3

[42] A. Ranjan and M. J. Black. Optical flow estimation using a spatial pyramid network. In *CVPR*, 2017. 5

[43] Y. Ren, X. Yu, R. Zhang, T. H. Li, S. Liu, and G. Li. Structureflow: Image inpainting via structure-aware appearance flow. In *ICCV*, 2019. 5

[44] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3):211–252, 2015. 5, 6

[45] C. Sakaridis, D. Dai, and L. Van Gool. Semantic foggy scene understanding with synthetic data. *IJCV*, 126(9):973–992, 2018. 5

[46] P. Shaw, J. Uszkoreit, and A. Vaswani. Self-attention with relative position representations. *arXiv:1803.02155*, 2018. 5

[47] C. Sun, Z. Zheng, X. Wang, M. Xu, and Y. Yang. Self-supervised point cloud representation learning via separating mixed shapes. *TMM*, 2022. 3

[48] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Deep image prior. In *CVPR*, 2018. 3

[49] J. M. J. Valanarasu, R. Yasarla, and V. M. Patel. Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In *CVPR*, 2022. 1

[50] F. A. Vasluianu, A. Romero, L. Van Gool, and R. Timofte. Shadow removal with paired and unpaired learning. In *CVPR*, 2021. 3

[51] J. Wang, X. Li, and J. Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *CVPR*, 2018. 5

[52] T. Wang, X. Yang, K. Xu, S. Chen, Q. Zhang, and R. W. Lau. Spatial attentive single-image deraining with a high quality real rain dataset. In *CVPR*, 2019. 7

[53] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *CVPR*, 2018. 4

[54] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li. Uformer: A general u-shaped transformer for image restoration. In *CVPR*, 2022. 1, 3

[55] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon. Cbam: Convolutional block attention module. In *ECCV*, 2018. 4

[56] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021. 3

[57] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu. Simmim: A simple framework for masked image modeling. In *CVPR*, 2022. 2, 3

[58] L. Xu, Q. Yan, Y. Xia, and J. Jia. Structure extraction from texture via relative total variation. *TOG*, 31(6):1–10, 2012. 5

[59] W. Yang, R. T. Tan, J. Feng, J. Liu, Z. Guo, and S. Yan. Deep joint rain detection and removal from a single image. In *CVPR*, 2017. 5

[60] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. H. Jiang, F. E. Tay, J. Feng, and S. Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *ICCV*, 2021. 3

[61] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M. H. Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 2022. 1, 3, 4, 6

[62] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M. H. Yang, and L. Shao. Multi-stage progressive image restoration. In *CVPR*, 2021. 1, 6, 7

[63] H. Zhang and V. M. Patel. Densely connected pyramid dehazing network. In *CVPR*, 2018. 3

[64] X. Zhang, R. Ng, and Q. Chen. Single image reflection separation with perceptual losses. In *CVPR*, 2018. 5

[65] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu. Residual non-local attention networks for image restoration. In *ICLR*, 2019. 3

[66] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr, and Zhang L. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, 2021. 3

[67] Z. Zheng, Y. Wei, and Y. Yang. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In *ACM MM*, 2020. 8

[68] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz. Joint discriminative and generative learning for person re-identification. In *CVPR*, 2019. 5

[69] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random erasing data augmentation. In *AAAI*, 2020. 3

[70] Z. Zou, S. Lei, T. Shi, Z. Shi, and J. Ye. Deep adversarial decomposition: A unified framework for separating superimposed images. In *CVPR*, 2020. 2, 3