Prior and Prediction Inverse Kernel Transformer for Single Image Defocus Deblurring

Peng Tang[1,6], Zhiqiang Xu[1*], Chunlai Zhou[2], Pengfei Wei[3], Peng Han[4], Xin Cao[5], Tobias Lasser [6]

[1] Machine Learning, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[2] Computer Science, Renmin University of China, Beijing, China
[3] Speech and Audio Team, Bytedance AI Lab, Singapore
[4] Computer Science, University of Electronic Science and Technology of China, Chengdu, China
[5] School of Computer Science and Engineering, Unvviversity of New South Wales, Sydney, Australia
[6] Schoold of CIT, CIIP group in Munich Institute of Bioengineering, Tecinical University of Munich, Garching, Germany
tangp@in.tum.de, zhiqiang.xu@mbzuai.ac.ae, czhou@ruc.edu.cn, wpf89928@gmail.com
penghan_study@foxmail.com, xin.cao@unsw.edu.au, lasser@in.tum.de

Abstract

Defocus blur, due to spatially-varying sizes and shapes, is hard to remove. Existing methods either are unable to effectively handle irregular defocus blur or fail to generalize well on other datasets. In this work, we propose a divide-and-conquer approach to tackling this issue, which gives rise to a novel end-to-end deep learning method, called prior-and-prediction inverse kernel transformer (P^2IKT), for single image defocus deblurring. Since most defocus blur can be approximated as Gaussian blur or its variants, we construct an inverse Gaussian kernel module in our method to enhance its generalization ability. At the same time, an inverse kernel prediction module is introduced in order to flexibly address the irregular blur that cannot be approximated by Gaussian blur. We further design a scale recurrent transformer, which estimates mixing coefficients for adaptively combining the results from the two modules and runs the scale recurrent "coarse-to-fine" procedure for progressive defocus deblurring. Extensive experimental results demonstrate that our P^2IKT outperforms previous methods in terms of PSNR on multiple defocus deblurring datasets.

Introduction

In an image captured by a camera, objects in the focal plane will appear sharp, otherwise blurry. The further away the objects are from the focal plane, the more blurry they are. This phenomenon is the so-called defocus blur (Quan, Wu, and Ji 2021). In photography, sometimes, defocus blur is an intentional artistic effect. However, in many computer vision tasks such as face recognition (Hua et al. 2012), biomedical imaging (Lefkimmiatis, Bourquard, and Unser 2011), and object detection (Dai et al. 2016), defocus blur is undesired as it affects the image quality and results in degrading performance. In such cases, single image defocus deblurring (SIDD) is crucial in many related high-level vision tasks (Campisi and Egiazarian 2017).

Two-stage approaches (Cho and Lee 2017; D'Andrès et al. 2016; Karaali and Jung 2017; Park et al. 2017; Shi, Xu, and Jia 2015; Liu et al. 2020) usually approximate the blur kernel with the prior kernel, either Gaussian (Shi, Xu, and Jia 2015; Park et al. 2017; Karaali and Jung 2017; Liu et al. 2020; Quan, Wu, and Ji 2021; Lee et al. 2019) or disc (D'Andrès et al. 2016; Bando and Nishita 2007) kernels, to reduce the complexity. Under the approximation, these methods first estimate a defocus map to derive the blur kernel. During the estimation, they only need to focus on optimizing the kernel size. Given the estimated defocus map, non-blind deconvolutions (Fish et al. 1995) are done to restore a sharp image. However, the kernel shapes in a real-world defocused image could be more complex than a prior kernel, which will cause an inaccurate defocus map estimation and consequently affect the deblurring quality.

Recently, learning-based approaches using deep neural networks (DNNs) (Abuolaim and Brown 2020; Son et al. 2021; Lee et al. 2021; Ruan et al. 2022; Quan, Wu, and Ji 2021) were proposed, which significantly improve the performance of the SIDD task compared to traditional two-stage approaches. Most learning-based approaches, such as DPDNet (Abuolaim and Brown 2020), IFAN (Lee et al. 2021), KPAC (Son et al. 2021), and DRBNet (Ruan et al. 2022), adopt an end-to-end training scheme which learns a mapping directly from a blurry image to a sharp image. However, the mapping learned by the end-to-end scheme is specific to images within the training dataset. It thus makes such mapping methods not robust enough to defocused images outside the training set. DMENet (Lee et al. 2019) and GKMNet (Quan, Wu, and Ji 2021) integrate the prior kernels approximation into the learning pipeline to reduce the complexity of the defocus deblurring. However, they still suffer from issues similar to traditional two-stage approaches, i.e., they are unable to effectively handle irregular blur whose shape cannot be approximated by prior kernels.

In this paper, we propose a novel deep learning approach, i.e., Prior-and-Prediction Inverse Kernel Transformer (P^2IKT) consisting of prior and prediction inverse kernel block (P^2IKB) and scale recurrent transformer (SRT), to tackle the above issues for single image defocus deblurring. We consider that defocus blur is either amenable

*Corresponding author

to Gaussian kernel approximation or irregular, and follow a divide-and-conquer approach to handle both types of blur. On the one hand, to effectively manage Gaussian-approximated blur with spatially-varying sizes, the P^{2}IKB block uses an Inverse Gaussian kernel module (IGKM) to perform the wiener deconvolutions (Wiener et al. 1949) based on approximate multi-size Gaussian blur kernels. IGKM improves generalization over previous direct mapping methods. On the other hand, to flexibly deal with the irregular blur, the P^{2}IKB block adopts an inverse kernel prediction module (IKPM), inspired by recent kernel prediction networks (KPNs), to predict the corresponding inverse irregular kernel (Mildenhall et al. 2018; Cho, Son, and Kim 2021; Ren et al. 2020). In contrast to the previous method based solely on the Gaussian kernel approximation, it complements IGKM to handle irregular blur. We then design SRT that maps a blurry image into a coefficient map for adaptively integrating deconvolutional feature maps from the IGKM and IKPM. SRT leverages the strong feature mapping of the latest transformer (Vaswani et al. 2017; Tsai et al. 2022) to produce accurate coefficient maps and uses the scale recurrent scheme to combine the results of the P^{2}IKB block from coarse to fine for progressive defocus deblurring.

To verify the effectiveness and generalization ability of our method, we conduct experiments on multiple defocus blur benchmarking datasets captured by different cameras, such as DPDD (Abuolaim and Brown 2020), RealDoF (Lee et al. 2021), LF-DOF (Ruan et al. 2021), DED (Ma et al. 2021), RTF (D'Andrès et al. 2016), PixelDP (Abuolaim and Brown 2020), and CUHK (Shi, Xu, and Jia 2014). The quantitative and qualitative comparisons with previous methods on these datasets demonstrate the effectiveness of our proposed method.

Related Works

Defocus Deblurring Most two-stage SIDD approaches (Cho and Lee 2017; D'Andrès et al. 2016; Karaali and Jung 2017; Park et al. 2017; Shi, Xu, and Jia 2015; Liu et al. 2020) focus on the optimization of the first stage, i.e., defocus map estimation (DME), and have the second stage addressed by existing non-blind deconvolution methods (Wiener et al. 1949; Krishnan and Fergus 2009). Various methods were proposed for DME, leveraging hand-crafted features (Karaali and Jung 2017; Shi, Xu, and Jia 2015; D'Andrès et al. 2016), deep features (Lee et al. 2019), or both (Park et al. 2017). However, these approaches still need help to handle irregular defocus blur due to the restrictive blur kernels, such as Gaussian and disc kernels.

Given the success of deep neural networks in computer vision, learning-based approaches using DNN have been proposed for SIDD. Abuolaim and Brown (Abuolaim and Brown 2020) proposed the first end-to-end defocus deblurring model, i.e., DPDNet, which directly maps blurry images to sharp images. It significantly outperforms the two-stage approaches but still can't handle spatially-varying blur. For further improving deblurring performance, various approaches such as kernel-sharing parallel atrous convolution (Son et al. 2021), iterative filter adaptive network (Lee et al. 2021), and dynamic residual network (Ruan et al. 2022), were put forward. These approaches improve the deblurring performance by enhancing the capability of handling spatially varying defocus blur to a certain extent. However, the downside is that the end-to-end mappings they learn only work for specific datasets. Some other methods, such as DMENet (Lee et al. 2019) and GKMNet (Quan, Wu, and Ji 2021), integrate the Gaussian kernel approximation into the learning process. As a result, it enhances the generalization ability, as most defocus blur can be approximated by the Gaussian kernel or its variants. However, they ignore the potentially present irregular blur.

Kernel Prediction Network Kernel prediction network methods have been applied for low-level computer vision tasks (Cho, Son, and Kim 2021; Mildenhall et al. 2018; Xia et al. 2020; Fan et al. 2021; Ren et al. 2020). For instance, Cho et al. (Cho, Son, and Kim 2021) proposed a weighted multi-kernel prediction network that considers the inter-dependencies of multi-scale kernels for efficient burst image super-resolution. Ren et al. (Ren et al. 2020) designed an unconstrained non-blind deconvolution model that predicts blur kernel and generates latent clean image simultaneously for adaptive motion deblurring. However, directly utilizing such methods may cause problems similar to end-to-end learning-based ones, i.e., estimated blur kernels may only work on data generated by specific cameras and thus affect their robustness on real-world defocused images.

Scale Recurrent Scheme Scale recurrent scheme has firstly been proposed for image deblurring task in (Tao et al. 2018), which extracts multi-scale information progressively and performs a coarse-to-fine dynamic deblurring for improving performance. Quan et al. (Quan, Wu, and Ji 2021) introduced attention modules (Zhong et al. 2020; Xu et al. 2021) into scale recurrent module (SRM) (Tao et al. 2018) to construct scale recurrent attention module (SRAM), aiming to enhance the feature representation ability and thus increase the defocus deblurring accuracy. Recently, the self-attention-based model, i.e., transformer (Vaswani et al. 2017), has proven its effectiveness in computer vision tasks. It inspires us, in this paper, to explore the combination of the most-advanced transformer (Tsai et al. 2022) with SRM for defocus deblurring.

Method

We now introduce how to design our Prior and Prediction Inverse Kernel Transformer (P^{2}IKT) for SIDD and elaborate on the network structure.

Main Idea

Our P^{2}IKT consists of the prior and prediction inverse kernel block (P^{2}IKB) and scale recurrent transformer (SRT). Before proceeding to the first component P^{2}IKB which aims to simulate inverse kernel deconvolutions, we show how to derive the inverse kernel deconvolution in what follows. Generally, a blur model is defined as follows:

$$I_B = k \otimes I_S, \tag{1}$$

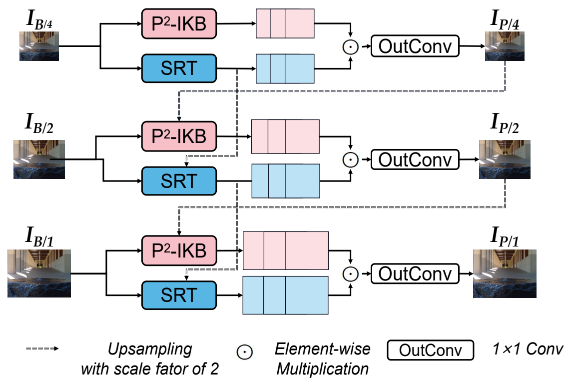


Figure 1: The SIDD framework of our method P^{2}IKT.

where I_B, I_S, k and $\otimes$ denote a blurry image, a sharp image, a blur kernel, and a convolution operation, respectively.

Gaussian kernel approximation has been proven effective in many defocus blur approximation methods (Shi, Xu, and Jia 2015; Park et al. 2017; Karaali and Jung 2017; Liu et al. 2020; Quan, Wu, and Ji 2021; Lee et al. 2019). However, in the real world, the blur kernel k of the defocused image is complex, and some irregular blurs are hard to be estimated by the Gaussian kernel or its variants. To reduce the complexity, we take a divide-and-conquer approach. We first consider that k is either a Gaussian-approximated kernel k_{ga} or irregular kernel k_{ir} (see Eq. (2)). Then we design the corresponding inverse kernel module to handle the defocus blur produced by k_{ga} and k_{ir}, respectively.

$$I_B = \begin{cases} k_{ga} \otimes I_S, k \text{ approximated as Gaussian} \\ k_{ir} \otimes I_S, \text{otherwise} \end{cases}, \quad (2)$$

Since defocus blurs are spatially-varying, it is unknown where the blur with k_{ga} or k_{ir} is in a defocused image. This makes piece-wise deblurring infeasible. Instead, we perform both inverse kernel-based deconvolutions on each defocused image. Then, we introduce a transformer to generate coefficient maps for combining the deconvolution results from these two inverse kernel deconvolutions, which implicitly differentiates between k_{ga} and k_{ir} and adaptively select and fuse these two deconvolution results for SIDD.

Inverse Kernel of k_{ga} For prior knowledge, we adopt the most recent approximation method, which uses a linear combination of multi-size Gaussian blur kernels (Quan, Wu, and Ji 2021) to represent the Gaussian-approximated defocus blur k_{ga}. Based on the above approximation, it is natural to use a linear combination of corresponding multi-size inverse Gaussian kernels to process the Gaussian-approximated defocus blur. Then, given the multi-size Gaussian kernels $g(\sigma_j)$, we can get the corresponding inverse kernel of $g(\sigma_j)$, based on the wiener deconvolution (Wiener et al. 1949; Gonzalez and Woods 2018), and thus we get the final form of Eq. (3) to restore the defocused image with k_{ga}.

$$I_S = \sum_{j=1}^{J} \alpha_j \odot \left(f^{-1}\left(\frac{\bar{f}(g(\sigma_j))}{f^2(g(\sigma_j)) + n}\right) \otimes I_B \right), \quad (3)$$

where $g(\sigma)$ is the Gaussian kernel of variance σ^2, and α_j represents the coefficient matrix for the j-th inverse Gaussian kernel, $\bar{f}(\cdot)$ denotes the conjugate of $f(\cdot)$, $f(\cdot)$ and $f^{-1}(\cdot)$ denote the discrete Fourier transform and inverse discrete Fourier transform, respectively, and n is the input noise. Since noises in the real world are unknown, they are empirically set to a constant (Gonzalez and Woods 2018) (0.01 in our experiments).

Inverse Kernel of k_{ir} To address the defocus blur with irregular kernel k_{ir}, we use the kernel prediction network (KPN) to implicitly predict the inverse kernel $k_{ir}^{\dagger}$ with an end-to-end learning scheme under the constraint as follows:

$$I_S = f^{-1}\left(\frac{1}{f(k_{ir})}\right) \otimes I_B = k_{ir}^{\dagger} \otimes I_B, \quad (4)$$

where $k_{ir}^{\dagger}$ denotes the inverse kernel of k_{ir} in Eq. (1).

Network Strcuture

Overall Structure We first give the overall structure of our method and then elaborate on two of its components, P^{2}IKB and SRT. As shown in Fig. 1, the network structure is divided into three recurrent deblurring stages on different scales for progressive defocus deblurring with multi-scale information. The deblurring process is performed from coarse (low-resolution blurry image $I_{B/4}$) to fine (high-resolution blurry image $I_{B/1}$), where both deblurred results (i.e., $I_{P/4}$ and $I_{P/2}$) and hidden states resulting from the first and second stages will be passed to P^{2}IKB and SRT in their next stages. The deblurred result $I_{P/1}$ in the third stage is the final output of our model P^{2}IKT.

We take the second stage as an example to illustrate the deblurring process of each stage, where $I_{B/S}$ represents that I_B is downscaled by scale factor S, and $I_{B/1}$ equals I_B. First, the deconvolutional feature maps (DFMs) are generated by feeding $I_{B/2}$ and upscaled deblurred results $I_{P/4}$ into the P^{2}IKB block, and coefficient maps (CMs) are generated by feeding $I_{B/2}$ and upscaled hidden states from the last stage into the SRT. Then, DFMs and CMs are combined via element-wise product and convolution to adaptively restore the defocused image $I_{B/2}$ to the deblurred image $I_{P/2}$.

Prior and Prediction Inverse Kernel Block As shown in Fig. 2, our P^{2}IKB block contains two modules, i.e., inverse Gaussian kernel module (IGKM) and inverse kernel prediction module (IKPM), which are designed based on Eqs. (3)-(4) to handle kernels k_{ga} and k_{ir} (see Eq. (2)), respectively.

The IGKM module is constructed as a group convolutional layer that uses a series of predefined inverse Gaussian kernels as the layer weight and is applied to R, G, and B channels of an input image to get corresponding results. During the construction, we first follow (Quan, Wu, and Ji 2021) to define J Gaussian kernels with different α (let k_s be the maximum Gaussian kernel size) and then compute the inverse Gaussian kernels based on the wiener deconvolution as shown in Eq. (3). The attention maps (AMs) generated by SRT represent the coefficient matrix α_j in Eq. (3).

The IKPM module, as shown in Fig. 2, consists of three parts: kernel prediction network (KPN), Trancov, and OutConv. The first part KPN adopts three convolution blocks

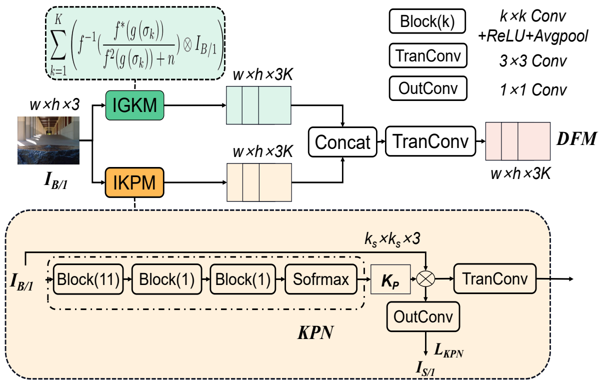


Figure 2: The whole network strcuture of the proposed P^{2}IKB.

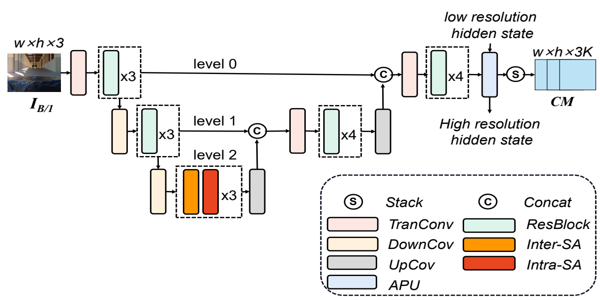


Figure 3: The diagram of our base model, i.e., SRT.

and one Softmax function to map a blurry image to a predicted kernel K_p of size $k_s \times k_s \times 3$. The third part of Outconv is to introduce an auxiliary loss L_{KPN} to simulate Eq. (4). By optimizing

$$L_{KPN} = L_2(I_{S/1}, \text{OutConv}(K_p \otimes I_{B/1})), \tag{5}$$

K_p tends to become the inverse kernel $k_{ir}^{\dagger}$ due to the implicit end-to-end training scheme. The second part TransConv is to make the number of channels of the deconvolutional feature maps generated by $K_p \otimes I_{B/1}$ equal to that of the feature maps generated by the IGKM, which enables the DFMs from two modules IGKM and IKPM to be treated equally during the learning process. Finally, the P^{2}IKB block uses the concatenation operation and then TransConv on the DFMs resulting from the two modules to generate the final DFMs, since it gives rise to better performance in our experiments with fewer parameters than directly outputting the DFMs from the two modules. More analysis of the concatenation operation and TransConv can be seen in the supplementary material.

Scale Recurrent Transformer. In our method, SRT is used to generate the coefficient maps to adaptively fuse DFMs from the P^{2}IKB block for the defocus deblurring. As these maps can be viewed as attention maps, we leverage the most advanced self-attention-based models, i.e., transformer-based models (Vaswani et al. 2017; Tsai et al. 2022).

Model	DPDD PSNR	DPDD SSIM	DPDD LPIPs	# Params (M)
Blurry Input	23.89	0.725	0.349	-
JNB	23.69	0.707	0.442	-
EBDB	23.94	0.723	0.402	-
DMENet	23.90	0.720	0.410	26.94
DPDNet-S	24.03	0.735	0.279	35.25
KPAC	25.22	0.774	0.227	1.58
IFAN	25.37	0.789	0.217	10.48
GKMNet	25.36	0.774	0.276	1.41
DRBNet	25.47	0.787	0.246	-
Restormer	<u>25.98</u>	**0.811**	**0.178**	26.10
P^{2}IKT (Ours)	**26.29**	<u>0.807</u>	<u>0.191</u>	3.32

Table 1: The quantitative results between our model and other state-of-the-art models on the DPDD test set (Abuolaim and Brown 2020). Lower value for LPIPs is better.

As shown in Fig. 3, SRT is built as an encoder-decoder structure, which is convenient for utilizing multi-scale features. It first adopts the Resblocks and DownConvs to encode the blurry image I_B into multi-scale features, and then embeds the inter-strip and intra-strip attention modules (Tsai et al. 2022) in the level 2 branch to improve the blur pattern adaptivity from different orientations for better prediction. Then, in the level 0 branch, multiple Resblocks are used to refine the feature maps combined with the encoded and decoded features. Finally, a recurrent unit, i.e., APU (Tao et al. 2018; Quan, Wu, and Ji 2021) block, combines the hidden state from the last stage (lower resolution stage) to predict the coefficient maps and generate the hidden state used for the prediction of SRT in the next stage. We follow the IFAN (Lee et al. 2021) and Restormer (Zamir et al. 2022) that add more basic blocks in the decoder/reconstruction part for more refined predictions.

Learning Objective We use the mean square error (L_2) as the main loss, and the frequency-domain loss L_{freq} and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018) loss L_{LPIPS} as auxiliary losses, to co-train our model P^{2}IKT, since these auxiliary losses have successfully improved performance in image restoration tasks (Zhao et al. 2016; Jiang et al. 2021; Cho et al. 2021; Lee et al. 2021). The overall loss function between the network output $I_{P/1}$ and the corresponding ground truth $I_{S/1}$ is given as follows

$$L = L_2 + \lambda_1 L_{freq} + \lambda_2 L_{LPIPS} + \lambda_3 L_{KPN}, \tag{6}$$

where λ_1, λ_2, and λ_3 are empirically set to 0.2, 0.2, and 0.05, respectively.

Experiments

Experimental Configuration In all experiments, we set the maximum number J of Gaussian kernels to 5 and the maximum kernel size k_s of Gaussian kernels to 5 too, as they are equal in the multi-size Gaussian kernel approximation (Quan, Wu, and Ji 2021). We use the Adam optimizer (Kingma and Ba 2014) with batch size 4 to train our model for 1600 epochs. The stochastic weights averaging scheme

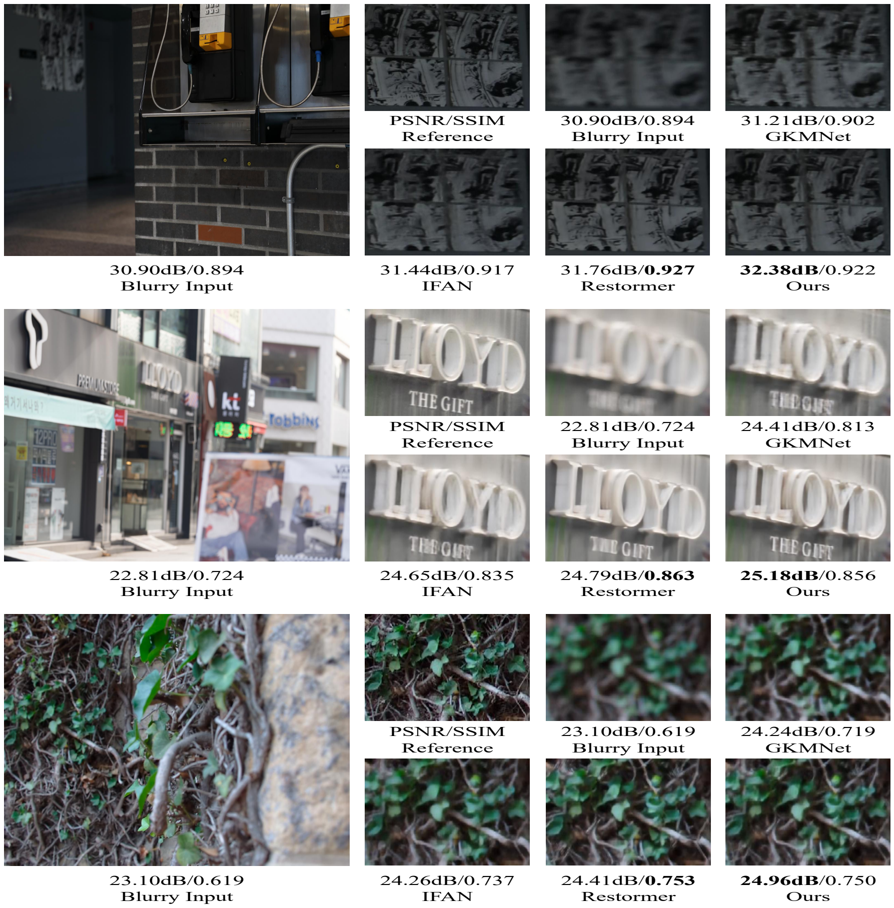


Figure 4: Qualitative results on DPDD (1th row), RealDoF (2nd row) and LF-DOF (3rd row) datasets by GKMNet (Quan, Wu, and Ji 2021), IFAN (Lee et al. 2021) , Restormer (Zamir et al. 2022) and our method P^2IKT.

(SWA) (Izmailov et al. 2018) is used in the last 100 epochs to generate the final weight for evaluations. The initial learning rate is set to 2e-4 and then reduced to 1e-4 at the 1000th epoch and 2.5e-5 at the 1500th epoch, as the SWA often uses a small learning rate to generate the final weight. The input images are randomly cropped to 384×384 in the first 1500 epochs and to 512×512 in the last 100 epochs. The data augmentation, including vertical and horizontal flipping and rotation, is executed during our training. All the experiments in this work were conducted with a GPU of NVIDIA A100. Code is available at https://github.com/TPZZZ/P2IKT.

Three commonly used evaluation metrics are used for the SIDD task (Son et al. 2021; Lee et al. 2021; Quan, Wu, and Ji 2021; Ruan et al. 2022), including the main metric PSNR (Peak Signal to Noise Ratio) and other two auxiliary metrics, SSIM (Structural Similarity Index Measure) (Wang et al. 2004) and LPIPS (Learned Perceptual Image Patch Similarity) (Zhao et al. 2016).

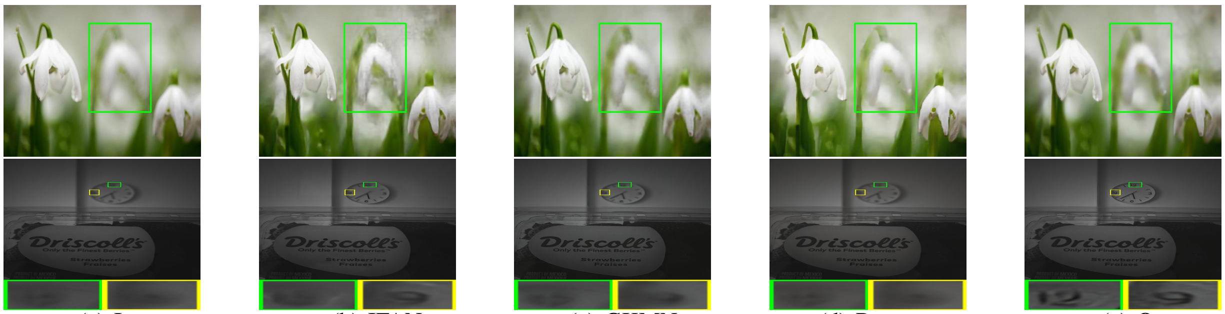
In our experiments, the DPDD training set (Abuolaim and Brown 2020) is used for training and other DPDD (Abuolaim and Brown 2020), RealDoF (Lee et al. 2021), LF-DOF (Ruan et al. 2021), DED (Ma et al. 2021), and RTF (D'Andrès et al. 2016) test sets are used for quantitative and qualitative evaluations. Note that the images in the RealDoF dataset were downscaled[1] to 1120×1680 for evaluation. The images of the DED and RTF test sets were cropped from 409×613 and 360×360 to 400×608 and 352×352, respectively, as the original image will cause the size problem during the evaluations. The DPDD training and test sets were captured by the same camera, while the other four datasets are not and used to simulate complex real-world images that are not from a specific camera. So, the experiments on the last four datasets were mainly used to evaluate model's generalization ability. Furthermore, the CUHK (Shi, Xu, and Jia 2014), and PixelDP (Abuolaim and Brown 2020) datasets were used to evaluate the generalization ability of models qualitatively.

Deblurring Performance Comparisons In this experiment, we compare our P^2IKT with currently advanced SIDD methods, including traditional two-stage methods

[1]It was downscaled to the same image size as the DPDD dataset. Due to our computational resource limit, the Restormer cannot be evaluated in the original image of the RealDoF dataset.

Model	RealDoF			LF-DOF			DED			RTF		
	PSNR	SSIM	LPIPs	PSNR	SSIM	LPIPs	PSNR	SSIM	LPIPs	PSNR	SSIM	LPIPs
Blurry Input	22.54	0.636	0.498	25.87	0.779	0.316	**28.58**	0.884	0.164	24.18	0.739	0.364
IFAN	25.01	0.770	0.250	26.11	0.817	0.220	27.85	**0.890**	**0.111**	24.92	0.821	0.215
GKMNet	24.58	0.735	0.337	25.96	0.802	0.271	27.93	0.883	0.144	25.10	0.826	0.274
Restormer	25.43	**0.801**	**0.218**	26.44	**0.824**	**0.207**	26.95	0.884	0.113	24.21	0.822	0.204
P^2IKT (Ours)	**25.78**	0.787	0.235	**26.90**	0.821	0.220	28.29	0.888	0.123	**25.85**	**0.839**	**0.207**

Table 2: The quantitative results between our model and other state-of-the-art models on the RealDoF (Lee et al. 2021), LF-DOF (Ruan et al. 2021), DED (Ma et al. 2021) and RTF (D'Andrès et al. 2016) test sets.



(a) Input (b) IFAN (c) GKMNet (d) Restormer (e) Ours

Figure 5: Qualitative results on CUHK (1th row) (Shi, Xu, and Jia 2014) and PixelDP (2nd row) (Abuolaim and Brown 2020) datasets by IFAN (Lee et al. 2021) , GKMNet (Quan, Wu, and Ji 2021), Restormer (Zamir et al. 2022) and our method P^2IKT.

(JNB (Shi, Xu, and Jia 2015), EBDB (Karaali and Jung 2017), DMENet (Lee et al. 2019)) and end-to-end learning SIDD methods (DPDNet (Abuolaim and Brown 2020), KPAC (Son et al. 2021), IFAN (Lee et al. 2021), GKMNet (Quan, Wu, and Ji 2021), DRBNet (Ruan et al. 2022), and the most advanced generalized image restoration method, Restormer (Zamir et al. 2022)). The experimental results of these methods are quoted from their papers or obtained from their released pre-trained weights and codes. All the models were only trained using the single image in the DPDD training set, while IFAN needs the extra dual-pixel data for its training (Lee et al. 2021).

Table 1 shows the quantitative results of different SIDD approaches on the benchmarking dataset of DPDD in terms of PSNR, SSIM, and LPIPs. As we can see in Table 1, the traditional two-stage approaches (3rd-5th rows) achieve a worse deblurring performance than other learning-based approaches (6th-10th rows), and one of the two-stage approaches (2nd row, JNB) outputs images that are even more blurry than input images (1st row) (23.69dB vs. 23.89dB). Among the learning-based approaches, the models, with more advanced modules (7th-10th rows for the KPAC, IFAN, GKMNet, and DRBNet, respectively) that are designed to handle spatially varying defocus blur, significantly improve the deblurring quality compared to the DPDNet (6th row). Specifically, KPAC, IFAN, GKMNet, and DRBNet outperform DPDNet by more than 1dB in PSNR. Furthermore, compared with previous methods, Restormer and our P^2IKT show the superiority in all the evaluation metics. Compared to Restormer, P^2IKT achieves a higher value (26.29dB vs. 25.98dB) in the main metric PSNR, but worse SSIM (0.807 vs. 0.811) and LPIPs (0.191 vs. 0.178) values, with much less model parameters (3.32M vs. 26.10M).

Generalization Ability Analyses To evaluate the generalization ability, we further conduct the quantitative comparison between IFAN, GKMNet, Restormer, and our method on RealDoF, LF-DOF, DED, and RTF test sets, and report results in Table 2. These methods were trained on the DPDD training set and then tested on the above four datasets.

As shown in Table 2, regarding comparisons on the RealDoF and LF-DOF test sets, IFAN obtains more accurate deblurring quality than GKMNet on both test sets. It is probably because the defocus disparity estimation trained on extra dual-pixel data helps IFAN to handle some defocus blur patterns. P^2IKT and Restormer outperform IFAN and GKMNet regarding the PSNR value on both datasets. In particular, P^2IKT obtains the highest value in PSNR, and Restormer obtains the best values in SSIM and LPIPs. On the DED test set, in terms of PSNR, all the methods fail to deblur input images, but our method works better than the other three methods. On the RTF test set, our method performs best in all the metrics, i.e., 25.85dB in PSNR, 0.839 in SSIM, and 0.207 in LPIPs. These experimental results show that on the datasets with high-resolution images, such as DPDD (1120×1680), RealDoF (1120×1680), and LF-DOF (688×1008), our method achieves the comparable performance with Restormer and outperforms other methods. On the datasets of relatively low-resolution images, such as DED (400×608) and RTF (352×352), the performance of Restormer is even worse than IFAN and GKMNet, and our method outperforms all other methods. These results demonstrate that P^2IKT can generalize to defocused images

P²IKT IGKM	IKPM	SRT	SRAM	DPDD			RealDoF			LF-DOF			# Params (M)
			✓	24.27	0.766	0.229	22.97	0.739	0.290	25.25	0.768	0.244	3.58
✓	✓		✓	25.91	0.791	0.208	25.22	0.766	0.252	26.40	0.809	0.222	3.64
		✓		26.03	0.798	0.199	25.46	0.782	0.240	26.01	0.799	0.220	3.26
✓		✓		26.13	0.800	0.200	25.72	0.783	0.238	26.73	0.807	0.220	3.26
	✓	✓		26.12	0.799	0.203	25.67	0.782	0.249	26.69	0.806	0.227	3.32
✓	✓	✓		**26.29**	**0.807**	**0.191**	**25.78**	**0.787**	**0.235**	**26.90**	**0.821**	**0.220**	3.32

Table 3: Qualitative results of ablation study. The IGKM, IKPM, SRT and SRAM stand for Inverse Gaussian Kernel Module, Inverse Kernel Prediction Module, Scale Recurrent Transformer, and Scale Recurrent Attention Model, respectively.

on multiple datasets better than other methods.

Figure 4 shows the qualitative comparisons between IFAN, GKMNet, Restormer, and our method P^{2}IKT on DPDD, RealDOF, and LF-DOF test sets. As we can see from Figure 4, our method and Restormer remove the remaining blur that GKMNet and IFAN cannot remove. Furthermore, our method achieves the highest PSNR value, while Restormer achieves the best SSIM value[2]. The visual results in Figure 4 indicate that our method tends to restore more details and Restormer tends to remove more ringing artifacts. For example, in the second row of the figure, our method restores the details of “The GIFT” but fails to remove the ringing artifacts around “LLOYD”, while the Restormer does the opposite.

We also extend the experiments to CUHK (Shi, Xu, and Jia 2014), and PixelDP (Abuolaim and Brown 2020) datasets. As the defocused images of these two datasets do not have corresponding ground truth, we only show the deblurred results by different methods for visual inspection (see Fig. 5). When dealing with the case from the CUHK dataset, the deblurred results of IFAN contain ringing artifacts, and those of GKMNet and Restormer remain blurred, while our P^{2}IKT produces more clear results without ringing artifacts. For the cases from the PixelDP dataset, GKMNet and Restormer cannot restore the details of ”12” (green box) and ”9” (yellow box) in Fig. 5, IFAN can only roughly restore digit ”9”, and our method is capable of clearly restoring the details of both digits (as well as ”3” and ”6” in the clock). These results qualitatively imply the generalization ability of our P^{2}IKT, and more visualization results can be found in the supplementary materials.

Ablation study To quantitatively validate the effect of each component, i.e., P^{2}IKB (IGKM+IKPM) and SRT, in our model P^{2}IKT, we conduct an ablation study and report the results on DPDD and RealDof datasets in Table 3. We first build two baseline models by replacing SRT with a scale recurrent attention model (SRAM) in IGKM+IKPM+SRT and SRT. SRAM is a combination of squeeze attention module (Xu et al. 2021; Zhong et al. 2020) and scale recurrent module (Tao et al. 2018), which has been applied in (Quan, Wu, and Ji 2021) to perform scale attention mechanism for SIDD. In our experiments, we add a Resblock in each block of the SRAM to ensure SRAM has similar parameters to SRT. Also, as most SIDD methods did, we removed all the BacthNormalization operations in SRAM to achieve better performance.

As shown in Table 3, both P^{2}IKB and SRT in our model can improve the deblurring performance. Adding IGKM or IKPM alone, the performance did not improve too much. However, adding the combination of IGKM and IKPM, i.e., P^2-IKB, boosts the deblurring performance of both SRT and SRAM backbones significantly on all the datasets. Specifically, IGKM+IKPM+SRT improves the PSNR value from 26.03dB to 26.29dB on the DPDD dataset, from 25.46dB to 25.78dB on the RealDOF dataset and from 26.01dB to 26.90dB on the LF-DOF dataset compared to SRT, while IGKM+IKPM+SRAM increases the PSNR value 24.37dB to 25.91dB on the DPDD dataset, from 22.97dB to 25.22dB on the RealDoF dataset and 25.25dB to 26.40dB on the LF-DoF dataset compared to SRAM. It is worth noting that these improvements only need less than1/40 parameters (0.05M) compared to SRT and SRAM. In addition, SRT and IGKM+IKPM+SRT significantly outperform SRAM and IGKM+IKPM+SRAM on three datasets with fewer models parameters, respectively.

Conclusion

We proposed a single image defocus deblurring method based on a novel prior and prediction inverse kernel transformer (P^{2}IKT). Inspired by the idea of “divide and conquer”, our P^{2}IKT considers the defocus blur to be either Gaussian-approximated or irregular and then builds an inverse Gaussian kernel module (IGKM) for the Gaussian approximated defocus blur and an Inverse Kernel Prediction module (IKPM) for irregular defocus blur. It is equipped with a scale recurrent transformer (SRT) which provides the scale recurrent mechanism for progressive defocus deblurring. Also, SRT generates coefficient maps to combine the deconvolution results from the two modules adaptively to achieve better defocus deblurring performance than each module alone. We experimentally verified the effect of each component in our model, and the comparisons on seven datasets with previous methods showed that our method generalizes well and outperforms existing methods in terms of PSNR. Despite achieving better performance than previous methods, our method still needs to deblur the cases in the DED dataset (see Table. 2). In our future work, we may investigate the domain adaption technique for further enhancing the generalization ability.

[2]This is a quantitative comparison between these two methods on the three datasets.

References

Abuolaim, A.; and Brown, M. S. 2020. Defocus deblurring using dual-pixel data. In *European Conference on Computer Vision*, 111–126. Springer.

Bando, Y.; and Nishita, T. 2007. Towards digital refocusing from a single photograph. In *15th Pacific Conference on Computer Graphics and Applications (PG'07)*, 363–372. IEEE.

Campisi, P.; and Egiazarian, K. 2017. *Blind image deconvolution: theory and applications*. CRC press.

Cho, S.; and Lee, S. 2017. Convergence analysis of MAP based blur kernel estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, 4808–4816.

Cho, S.-J.; Ji, S.-W.; Hong, J.-P.; Jung, S.-W.; and Ko, S.-J. 2021. Rethinking coarse-to-fine approach in single image deblurring. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4641–4650.

Cho, W.; Son, S.; and Kim, D.-S. 2021. Weighted multi-kernel prediction network for burst image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 404–413.

Dai, J.; Li, Y.; He, K.; and Sun, J. 2016. R-fcn: Object detection via region-based fully convolutional networks. *Advances in neural information processing systems*, 29.

D'Andrès, L.; Salvador, J.; Kochale, A.; and Süsstrunk, S. 2016. Non-parametric blur map regression for depth of field extension. *IEEE Transactions on Image Processing*, 25(4): 1660–1673.

Fan, H.; Wang, R.; Huo, Y.; and Bao, H. 2021. Real-time Monte Carlo Denoising with Weight Sharing Kernel Prediction Network. In *Computer Graphics Forum*, volume 40, 15–27. Wiley Online Library.

Fish, D.; Brinicombe, A.; Pike, E.; and Walker, J. 1995. Blind deconvolution by means of the Richardson–Lucy algorithm. *JOSA A*, 12(1): 58–65.

Gonzalez, R. C.; and Woods, R. E. 2018. Digital image processing, global edition. *Digital Image Processing, Global Edition*, 19.

Hua, F.; Johnson, P.; Sazonova, N.; Lopez-Meyer, P.; and Schuckers, S. 2012. Impact of out-of-focus blur on face recognition performance based on modular transfer function. In *2012 5th IAPR International Conference on Biometrics (ICB)*, 85–90. IEEE.

Izmailov, P.; Podoprikhin, D.; Garipov, T.; Vetrov, D.; and Wilson, A. G. 2018. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*.

Jiang, L.; Dai, B.; Wu, W.; and Loy, C. C. 2021. Focal frequency loss for image reconstruction and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13919–13929.

Karaali, A.; and Jung, C. R. 2017. Edge-based defocus blur estimation with adaptive scale selection. *IEEE Transactions on Image Processing*, 27(3): 1126–1137.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krishnan, D.; and Fergus, R. 2009. Fast image deconvolution using hyper-Laplacian priors. *Advances in neural information processing systems*, 22.

Lee, J.; Lee, S.; Cho, S.; and Lee, S. 2019. Deep defocus map estimation using domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12222–12230.

Lee, J.; Son, H.; Rim, J.; Cho, S.; and Lee, S. 2021. Iterative filter adaptive network for single image defocus deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2034–2042.

Lefkimmiatis, S.; Bourquard, A.; and Unser, M. 2011. Hessian-based norm regularization for image restoration with biomedical applications. *IEEE Transactions on Image Processing*, 21(3): 983–995.

Liu, Y.-Q.; Du, X.; Shen, H.-L.; and Chen, S.-J. 2020. Estimating generalized Gaussian blur kernels for out-of-focus image deblurring. *IEEE Transactions on circuits and systems for video technology*, 31(3): 829–843.

Ma, H.; Liu, S.; Liao, Q.; Zhang, J.; and Xue, J.-H. 2021. Defocus Image Deblurring Network With Defocus Map Estimation as Auxiliary Task. *IEEE Transactions on Image Processing*, 31: 216–226.

Mildenhall, B.; Barron, J. T.; Chen, J.; Sharlet, D.; Ng, R.; and Carroll, R. 2018. Burst denoising with kernel prediction networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2502–2510.

Park, J.; Tai, Y.-W.; Cho, D.; and So Kweon, I. 2017. A unified approach of multi-scale deep and hand-crafted features for defocus estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1736–1745.

Quan, Y.; Wu, Z.; and Ji, H. 2021. Gaussian Kernel Mixture Network for Single Image Defocus Deblurring. *Advances in Neural Information Processing Systems*, 34: 20812–20824.

Ren, D.; Zhang, K.; Wang, Q.; Hu, Q.; and Zuo, W. 2020. Neural blind deconvolution using deep priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3341–3350.

Ruan, L.; Chen, B.; Li, J.; and Lam, M. 2022. Learning to Deblur using Light Field Generated and Real Defocus Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16304–16313.

Ruan, L.; Chen, B.; Li, J.; and Lam, M.-L. 2021. Aifnet: All-in-focus image restoration network using a light field-based dataset. *IEEE Transactions on Computational Imaging*, 7: 675–688.

Shi, J.; Xu, L.; and Jia, J. 2014. Discriminative blur detection features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2965–2972.

Shi, J.; Xu, L.; and Jia, J. 2015. Just noticeable defocus blur detection and estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 657–665.

Son, H.; Lee, J.; Cho, S.; and Lee, S. 2021. Single image defocus deblurring using kernel-sharing parallel atrous convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2642–2650.

Tao, X.; Gao, H.; Shen, X.; Wang, J.; and Jia, J. 2018. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8174–8182.

Tsai, F.-J.; Peng, Y.-T.; Lin, Y.-Y.; Tsai, C.-C.; and Lin, C.-W. 2022. Stripformer: Strip Transformer for Fast Image Deblurring. In *ECCV*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wiener, N.; Wiener, N.; Mathematician, C.; Wiener, N.; Wiener, N.; and Mathématicien, C. 1949. *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*, volume 113. MIT press Cambridge, MA.

Xia, Z.; Perazzi, F.; Gharbi, M.; Sunkavalli, K.; and Chakrabarti, A. 2020. Basis prediction networks for effective burst denoising with large kernels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11844–11853.

Xu, Y.; Zhu, Y.; Quan, Y.; and Ji, H. 2021. Attentive deep network for blind motion deblurring on dynamic scenes. *Computer Vision and Image Understanding*, 205: 103169.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5728–5739.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhao, H.; Gallo, O.; Frosio, I.; and Kautz, J. 2016. Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging*, 3(1): 47–57.

Zhong, Z.; Lin, Z. Q.; Bidart, R.; Hu, X.; Daya, I. B.; Li, Z.; Zheng, W.-S.; Li, J.; and Wong, A. 2020. Squeeze-and-attention networks for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13065–13074.