Reblurring-Guided Single Image Defocus Deblurring: A Learning Framework with Misaligned Training Pairs

Dongwei Ren[✉1†], Xinya Shu[2†], Yu Li[2], Xiaohe Wu[2], Jin Li[3], Wangmeng Zuo[2]

[1]College of Intelligence and Computing, Tianjin University.
[2]Faculty of Computing, Harbin Institute of Technology.
[3]School of Electrical and Information Engineering, Tianjin University.

[†]These authors contributed equally to this work.

Abstract

For single image defocus deblurring, acquiring well-aligned training pairs (or training triplets), *i.e.*, a defocus blurry image, an all-in-focus sharp image (and a defocus blur map), is a challenging task for developing effective deblurring models. Existing image defocus deblurring methods typically rely on training data collected by specialized imaging equipment, with the assumption that these pairs or triplets are perfectly aligned. However, in practical scenarios involving the collection of real-world data, direct acquisition of training triplets is infeasible, and training pairs inevitably encounter spatial misalignment issues. In this work, we introduce a reblurring-guided learning framework for single image defocus deblurring, enabling the learning of a deblurring network even with misaligned training pairs. By reconstructing spatially variant isotropic blur kernels, our reblurring module ensures spatial consistency between the deblurred image, the reblurred image and the input blurry image, thereby addressing the misalignment issue while effectively extracting sharp textures from the all-in-focus sharp image. Moreover, spatially variant blur can be derived from the reblurring module, and serve as pseudo supervision for defocus blur map during training, interestingly transforming training pairs into training triplets. To leverage this pseudo supervision, we propose a lightweight defocus blur estimator coupled with a fusion block, which enhances deblurring performance through seamless integration with state-of-the-art deblurring networks. Additionally, we have collected a new dataset for single image defocus deblurring (SDD) with typical misalignments, which not only validates our proposed method but also serves as a benchmark for future research. The effectiveness of our method is validated by notable improvements in both quantitative metrics and visual quality across several datasets with real-world defocus blurry images, including DPDD, RealDOF, DED, and our SDD. The source code and dataset are available at https://github.com/ssscrystal/Reblurring-guided-JDRL.

Keywords: Defocus deblurring, image deblurring, reblurring model, isotropic blur kernels

1 Introduction

Defocus blur typically occurs when scene objects fall outside the camera's Depth of Field (DOF). This phenomenon can create a visually pleasing effect in certain photographic contexts. However, defocus blur often compromises the clarity of image details, thereby negatively impacting image quality and hindering research on high-level tasks such as object detection [3–5] and

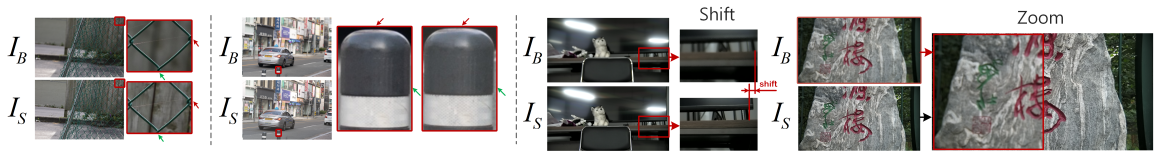


Fig. 1: Misalignment issues in the pairs of ground-truth sharp image $\boldsymbol{I}_S$ and defocus blurry image $\boldsymbol{I}_B$ in DPDD [1], RealDOF [2] and our SDD datasets. Although ground-truth and blurry image pairs in these datasets are designed to be aligned, spatial misalignment still exists.

segmentation [6–8]. To address these challenges, image defocus deblurring is required to handle various and complex blurred areas produced during the photographing process. Traditionally, the image defocus deblurring process adopts a two-step approach. It initially computes a defocus blur map [9–11], which delineates the amount of blur per pixel within a defocused blurry image. This map is then used to perform non-blind deconvolution [12, 13] on the image. The effectiveness of this strategy heavily depends on the precision of the defocus blur map. However, this approach often overlooks the nonlinear dynamics of real-world blurring, and tends to rely on simplistic blur models such as disk or Gaussian kernels. Recently, dual-pixel cameras have been employed to address defocus blur [1, 14] through the utilization of two-view images. However, it is worth noting that the majority of consumer cameras still produce single images for user observation. Therefore, this paper primarily focuses on the domain of single image defocus deblurring.

The emergence of deep learning techniques, particularly convolutional neural networks based models [1, 14] and Transformer based models [15–18], has significantly propelled the field forward by providing solutions to address defocus blur. These models demonstrate effectiveness through the acquisition of intricate mappings from extensive training data. However, the effectiveness of these learning-based approaches is intricately related to the quality and alignment of training samples. Often, these deblurring methods depended on precisely aligned pairs of images. In DED dataset [19], well-aligned defocus blurry images and their corresponding all-in-focus counterparts can be obtained using a light field camera like Lytro [20]. From such pairs, a defocus blur map can be estimated, resulting in training datasets comprising aligned triplets: a defocus blurry image, an all-in-focus ground-truth image, and a defocus blur map. The spatial alignment plays a pivotal role in effectively training and validating deep learning models for image defocus deblurring task.

However, for other consumer cameras, *e.g.*, digital single lens reflex or smartphone cameras, the availability of training triplets is impractical. In widely used datasets for image defocus deblurring, such as DPDD [1] and RealDOF [2] datasets, meticulous control is exercised over the capturing camera to ensure consistent acquisition of training pairs comprising a defocus blurry image and a ground-truth sharp image. Nevertheless, as illustrated in Fig. 1, inevitable misalignments still occur. We also note that different imaging sensors introduce various types of defocus blur, posing challenges for learned deblurring models based on specific cameras when handling cases involving other sensor types. Therefore, effective deployment on target devices requires learning a device-specific image defocus deblurring model while relaxing the requirement for perfect alignment in training pairs.

In this work, we propose a novel learning framework for single image defocus deblurring (Fig. 2), specifically focusing on effectively learning a single image defocus deblurring model from misaligned training pairs that can be easily obtained for a given imaging camera.

We first develop a reblurring-based learning framework to address the challenge of misaligned training pairs. To ensure the consistent spatial alignment between the deblurred image, the reblurred image and the input blurry image, our reblurring module consists of a Kernel Prediction Network (KPN) and a Weight Prediction Network (WPN) to reconstruct spatially variant isotropic blur kernels. Moreover, spatially variant blur can be derived from the reblurring module, and serve

as pseudo supervision for defocus blur map during training, interestingly expanding the training pairs to training triplets. To employ the pseudo supervision, we design a lightweight defocus blur estimator coupled with a fusion block, which enhances deblurring performance through seamless integration with state-of-the-art deblurring networks.

Furthermore, we introduce a new dataset named SDD for single image defocus deblurring. The image pairs are captured using a HUAWEI X2381-VG camera, which can be adjusted to capture pairs of blurry and sharp images by manipulating the camera motor or aperture size. Despite our efforts to maintain alignment during the collection process of the SDD dataset, some misalignment persists due to variations in consumer-grade cameras and collection settings. The misalignment primarily manifests in two forms: zoom misalignment and shift misalignment, as illustrated in Fig. 1. Importantly, the degree of misalignment within the SDD dataset tends to be more severe compared to that observed in the DPDD [1] dataset, making it be a testbed for evaluating our reblurring-guided image defocus deblurring techniques and serve as a benchmark for future research in this field.

This work is previously presented as a conference paper with oral presentation [21], upon which this manuscript has made three major improvements: (*i*) We have enhanced the learning framework by deriving pseudo defocus blur maps from the reblurring module and constructing triplets in training datasets. (*ii*) Correspondingly, we have improved the deblurring network architecture by incorporating a lightweight defocus blur map estimator coupled with a fusion block. This design not only seamlessly integrates with existing deblurring models but also significantly enhances deblurring performance. (*iii*) To provide a comprehensive comparison, we have included state-of-the-art methods based on convolutional neural networks (CNN) and Transformer for evaluation, including UformerT [16], Restormer [15], DID-ANet [19] and Loformer [22]. Furthermore, we have incorporated the DED dataset [19] to evaluate competing methods.

In summary, the contributions of this paper can be summarized as

- A novel reblurring-guided learning framework is proposed for image defocus deblurring that effectively exploits misaligned training pairs for learning deblurring models.
- A lightweight blur map estimator and a fusion block are designed to integrate with state-of-the-art deblurring networks, where the pseudo supervision on spatially variant blur map can be derived from our reblurring module.
- A new dataset named SDD comprising high-resolution image pairs with diverse contents is introduced for evaluating image defocus deblurring models and facilitating future research in this field. On benchmark datasets including DPDD [1], RealDOF [2], DED [19] and our SDD, our method is significantly superior to existing methods.

The remainder of this paper is organized as follows: Section 2 provides a comprehensive review of the relevant literature, Section 3 introduces our proposed method along with the new dataset, Section 4 experimentally validates the effectiveness of the proposed approach, and finally Section 5 concludes this paper by summarizing key findings.

2 Related Work

In this section, we briefly review relevant works including methods and datasets for defocus deblurring, and reblurring strategies.

2.1 Defocus Deblurring Methods

Traditional defocus deblurring approaches typically focus on estimating a defocus map [9, 23] by leveraging predefined models. These methods then apply non-blind deconvolution techniques [12, 13] to restore a sharp image. However, their performance is often limited by the inherent constraints of these blur models.

Recent approaches predominantly utilize deep learning to overcome the limitations of the traditional methods by restoring the image directly from the blurred image. Abuolaim et al. [1] pioneered the first end-to-end learning-based method DPDNet. This approach achieved significantly better results than traditional two-stage methods, establishing a new benchmark in the field. Subsequently, Lee et al. [2] designed a network

featuring an iterative filter adaptive module to address spatially varying defocus blur. Son et al. [24] proposed a kernel prediction adaptive convolution technique that further refines the capacity to address complex defocus patterns. Despite these advancements, it is crucial to note that the effectiveness of these deep learning-based defcus deblurring methods heavily depends on the quality of the training data, which is primarily derived from the DPDD [1] dataset. The dependence on high-quality, well-aligned training pairs has been a persistent challenge, motivating our research to focus on single image defocus deblurring with misaligned training data. Although some weakly supervised or unsupervised methods [25–27] can partially address the reliance on large-scale data, these methods often do not perform as well as supervised approaches.

Recently, All-in-One restoration methods have emerged, aiming to tackle multiple complex and unknown image degradations with a unified model. Park et al. [28] proposed an adaptive discriminative filter-based model to restore images with unknown degradations. Some advanced approaches [29–33] utilized large language model or diffusion model to effectively restore images from various types and levels of degradation, such as PromptIR [29]. Ai et al. [30] harnessed Stable Diffusion priors to further enhance the restoration process which has also shown promising results in the field of defocus deblurring. While these approaches have demonstrated commendable performance in defocus deblurring, they heavily rely on the capabilities of large models and sometimes require fine-tuning, which is a relatively labor-intensive process. Specifically targeting defocus blur, several methods [34,35] that employ specially designed defocus blur kernels have been proposed. These methods address the blur issue with greater precision. We propose a novel framework designed to overcome the limitations of existing datasets by accommodating misaligned training pairs, thus broadening the applicability of defocus deblurring techniques in real-world scenarios.

2.2 Defocus Deblurring Datasets

The availability of high-quality datasets plays a pivotal role in training and evaluating defocus deblurring algorithms. While several datasets have been proposed for single image deblurring, most synthetic datasets [36, 37] used for network training primarily focus on specific types of defocus blur, such as Gaussian or disk blur. Consequently, they often overlook other forms of blur that are commonly encountered in real-world scenarios.

To address this gap, various datasets have been introduced, such as the DPDD [1] dataset, which was collected using remote control mechanisms to acquire aligned data pairs, exhibiting varying degrees of defocus blur. Lee et al. [2] developed the RealDOF dataset using a dual-camera system which is capable of capturing both blurry and sharp images concurrently, offering a novel approach to dataset creation in this field. However, one common challenge with these datasets is the requirement for precisely aligned image pairs, which limits their applicability in real-world scenarios. Ma et al. [19] proposed DED dataset, which used a Lytro Illum light field camera [20] to collect a dataset with strictly aligned ground truth and input images. However, such cameras are not commonly used in everyday life. Additionally, the images in the DED dataset [19] have relatively lower resolution compared to existing datasets. In our research, instead of striving for the construction of a perfectly aligned dataset, we focus on incorporating and addressing misalignments within our network.

2.3 Reblurring Process

Compared with the end-to-end defocus deblurring methods, reblurring process is less explored in learning-based approaches, yet it holds significant promise for enhancing image restoration tasks, including defocus deblurring. Zhang et al. [38] introduced a reblurring network designed to generate additional blurry training images using GAN model, while Chen et al. [39] enhanced the video deblurring process by utilizing separately operated reblurring and deblurring networks. Furthermore, Lee et al. [2] introduced an additional network that inverts predicted deblurring filters to reblurring filters, and reblurred an all-in-focus image. These studies show that the reblurring process can beneficially contribute to the deblurring process. Our reblur module leverages the concept of spatially variant reblurring in defocus deblurring field, where the reblurring process can adapt to different regions of the image, thus better simulating real-world blur phenomena. Specifically,

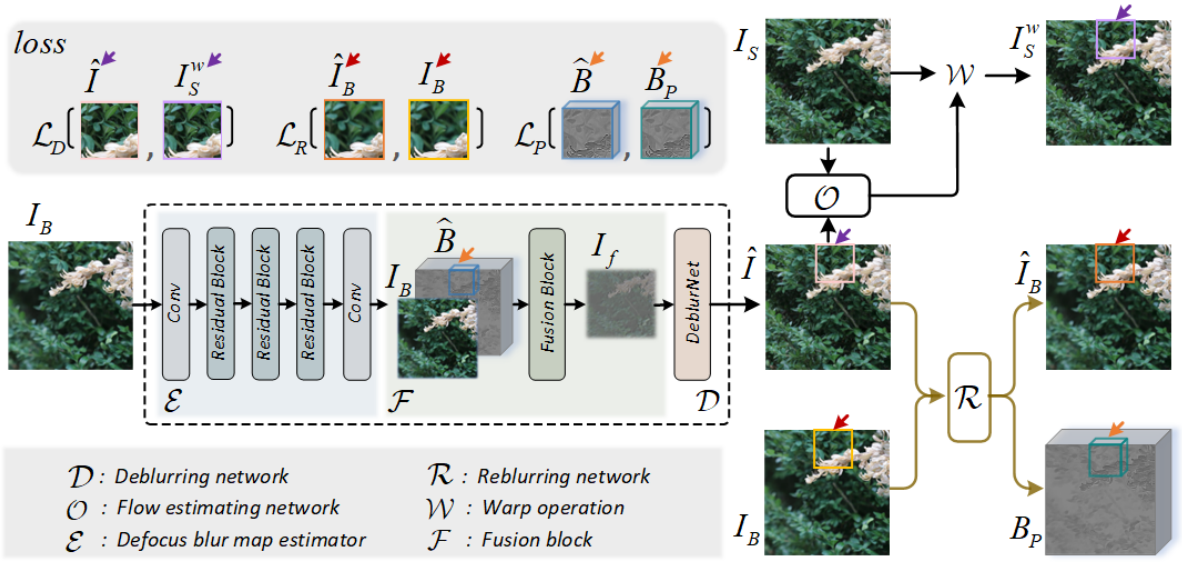


Fig. 2: Overview of our reblurring-guided learning framework for image defocus deblurring. It consists of deblurring module and reblurring module. In deblurring module, our introduced blur estimator $\mathcal{E}$ and fusion block $\mathcal{F}$ provide spatially variant degradation priors for enhancing deblurring performance, and can be seamlessly integrated with existing deblurring network $\mathcal{D}$. To tackle spatial misalignment of training pairs, optical flow-based deformation $\mathcal{W}$ is adopted to accommodate misalignment, while reblurring network $\mathcal{R}$ ensures spatial consistency of the deblurred image, the reblurred image and the input blurry image. Pseudo supervision $\boldsymbol{B}_P$ of defocus blur map $\hat{\boldsymbol{B}}$ can be derived from reblurring network.

the reblur module predicts a set of concentric isotropic blur kernels along with corresponding weight maps, which can more accurately simulate the spatial variations of blur found in real-world scenes. Acknowledging the standalone utility of the reblurring process in previous studies, our research posits that reblurring can also offer valuable prior cues to the deblurring network. This integration not only enriches the deblurring process but also utilizes the reblurring stage as a means to provide the deblurring network with insights that guide more effective image restoration. By jointly training the JDRL network, we aim to recover sharp images using misaligned training data pairs.

3 Proposed Method

Recent advancements in image defocus deblurring have primarily focused on the development of learning-based models using training pairs, denoted as $\{\boldsymbol{I}_B^n, \boldsymbol{I}_S^n\}_{n=1}^N$, where $\boldsymbol{I}_B$ represents a defocus blurry image and $\boldsymbol{I}_S$ denotes a ground-truth sharp image. However, even with careful alignment during data collection and post-processing techniques, spatial misalignment remains an inevitable issue as depicted in Fig. 1. Furthermore, when considering practical applications with new sensors, the severity of spatial misalignment issues may increase compared to those encountered in DPDD [1]. Additionally, it is important to note that different imaging sensors exhibit distinct patterns of defocus blur which limits the generalization ability of learned deblurring models to real-world cases captured by other devices. Therefore, for effective deployment on target devices, it is necessary to learn a image defocus deblurring model specifically tailored for each device while relaxing the requirement for perfect alignment in training pairs.

For instance, we acquire training pairs using a HUAWEI camera, and the models trained on DPDD dataset and DED dataset have limited performance. If we employ a pixel-wise loss function, such as Chamober loss [40], to train a UNet [41] based on the misaligned dataset, it may introduce

Input UNet UNet† GT

Fig. 3: The direct utilization of pixel-wise loss function for training a UNet model based on misaligned training pairs often leads to deformation artifacts, such as distorted brick lines. These artifacts can be effectively alleviated by our UNet†, where our reblurring-based learning framework Eq. (2) is adopted for training.

deformation artifacts in the deblurred results, as depicted in Fig. 3, where distortions like those highlighted by yellow boxes significantly impact image restoration quality.

To sum up, learning deblurring models based on misaligned training pairs is both challenging and meaningful. In the following, we first provide an overview of our proposed reblurring-based learning framework, which effectively leverages misaligned training pairs to learn image defocus deblurring models. Subsequently, we offer detailed explanations on the reblurring module with derivation of pseudo defocus blur map, the deblurring model equipped with a defocus blur map estimator and a fusion block, and finally introduce our newly established dataset.

3.1 The Overall Framework

Given a training set with N pairs $\{\boldsymbol{I}_B^n, \boldsymbol{I}_S^n\}_{n=1}^N$, where the blurry image $\boldsymbol{I}_B$ is not perfectly aligned with the ground-truth sharp image $\boldsymbol{I}_S$, our objective is to learn a deblurring model that effectively addresses misalignment issues while minimizing deformation distortions. In this work, we propose a novel reblurring-guided learning framework, wherein the misalignment problem can be resolved through the integration of a reblurring module.

As illustrated in Fig. 2, our proposed framework comprises two main components: a deblurring network, and a reblurring network. The deblurring netwok utilizes a dedicated network $\mathcal{D}$ to process the input blurry image $\boldsymbol{I}_B$ and generate an estimated deblurred image $\hat{\boldsymbol{I}}$.

To address misalignment between the ground-truth sharp image $\boldsymbol{I}_S$ and its corresponding estimate $\hat{\boldsymbol{I}}$, we introduce an optical flow-based deformation strategy integrated in our approach instead of relying solely on direct pixel-wise loss. In this manner, the deblurred result $\hat{\boldsymbol{I}}$ can adaptively learn sharp textures from the sharp image $\boldsymbol{I}_S$, thereby liberating it from the constraints of pixel-level precision. To address potential artifacts caused by optical flow deformation, we introduce a calibration mask and cycle deformation, which are further elaborated in Section 3.3. Subsequently, the reblurring module ensures spatial consistency between $\hat{\boldsymbol{I}}$ and $\boldsymbol{I}_B$. For this purpose, we propose a reblurring network $\mathcal{R}$ tasked with generating a reblurred image $\hat{\boldsymbol{I}}_B$ that closely approximates $\boldsymbol{I}_B$. Meanwhile, to ensure spatial coherence between $\hat{\boldsymbol{I}}$ and $\boldsymbol{I}_B$, the reblurring network $\mathcal{R}$ can also predict the isotropic blur kernels in polar coordinates.

The training loss for learning the parameters in deblurring network $\mathcal{D}$ and reblurring network $\mathcal{R}$ can be formally expressed as

$$\mathcal{L} = \mathcal{L}_D(\hat{\boldsymbol{I}}, \boldsymbol{I}_S) + \alpha \mathcal{L}_R(\hat{\boldsymbol{I}}_B, \boldsymbol{I}_B), \tag{1}$$

where $\mathcal{L}_D$ (Eq. (14)) is deblurring loss, $\mathcal{L}_R$ (Eq. (7)) is reblurring loss, and α is a trade-off parameter. Benefiting from the reblurring-guided training strategy, a Unet can be trained to be free from deformation artifacts, as shown in Fig. 3.

In this work, we further suggest that incorporating a defocus blur map can enhance the deblurring performance of deblurring networks. Specifically, an existing deblurring network $\mathcal{D}$ can be incorporated with a defocus blur map estimator $\mathcal{E}$ coupled with a fusion block $\mathcal{F}$ to enable the utilization of predicted degradation prior for enhancement of deblurring performance. The estimator $\mathcal{E}$ predicts the defocus blur map $\hat{\boldsymbol{B}}$, while $\mathcal{F}$ incorporates the degradation prior to enhance deblurring performance using a deformable attention mechanism. The utilization of degradation-related priors as input has been validated in various tasks [19, 42], as estimating the degradation is comparatively easier than performing deblurring itself. To enable the training of our baseline deblurring model, we need to prepare training triples, denoted as $\{\boldsymbol{I}_B^n, \boldsymbol{I}_S^n, \boldsymbol{B}^n\}_{n=1}^N$, where $\boldsymbol{B}$ represents the ground-truth spatially variant defocus blur map. However, obtaining such ground-truth maps

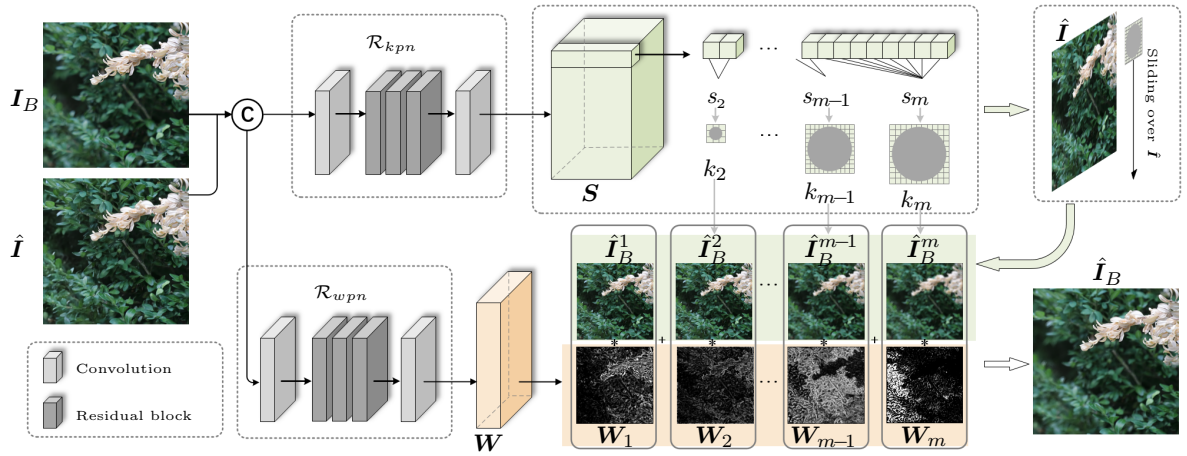


Fig. 4: The structure of reblurring network $\mathcal{R}$. There are two branches in $\mathcal{R}$: Kernel Prediction Network $\mathcal{R}_{kpn}$ predicts isotropic defocus blur kernels that are then used to generate blurred images with different blur levels, and Weight Prediction Network $\mathcal{R}_{wpn}$ predicts weight maps for integrating reblurred images.

is impractical due to difficulties in capturing spatially variant blur accurately. While previous work [19] estimated defocus blur maps from light field data captured by a Lytro camera, this approach is not suitable for popular consumer cameras. Fortunately, our proposed reblurring module provides a means to obtain pseudo ground-truth defocus maps $\boldsymbol{B}_P$ easily for each input blurry image and serves as supervision for the defocus blur map estimator $\mathcal{E}$ within our deblurring model. Therefore, based on the training triplets $\{\boldsymbol{I}_B^n, \boldsymbol{I}_S^n, \boldsymbol{B}_P^n\}_{n=1}^N$, the final training loss can be defined as

$$\mathcal{L}=\mathcal{L}_D(\hat{\boldsymbol{I}}, \boldsymbol{I}_S)+\alpha\mathcal{L}_R(\hat{\boldsymbol{I}}_B, \boldsymbol{I}_B)+\beta\mathcal{L}_P(\hat{\boldsymbol{B}}, \boldsymbol{B}_P), \quad (2)$$

where $\hat{\boldsymbol{B}} = \mathcal{E}(\boldsymbol{I}_B)$ denotes the estimated defocus blur map, $\mathcal{L}_P$ (Eq. (15)) is a loss function for defocus blur map, and β is a positive trade-off parameter.

In this way, we have developed a robust framework that enables the deblurring network to effectively leverage spatially adaptive sharp textures from the ground-truth sharp image while maintaining spatial consistency with the blurry input. During inference, our baseline deblurring model can generate latent sharp images while discarding the reblurring module.

3.2 Reblurring Module

In this section, We first introduce our reblurring network along with its loss function $\mathcal{L}_R$, and then present the acquisition of pseudo defocus blur map $\boldsymbol{B}_P$.

3.2.1 Reblurring Network

As illustrated in Fig. 4, reblurring network $\mathcal{R}$ consists of a Kernel Prediction Network $\mathcal{R}_{kpn}$ and a Weight Prediction Network $\mathcal{R}_{wpn}$.

Kernel Prediction Network $\mathcal{R}_{kpn}$: The objective of $\mathcal{R}_{kpn}$ is to predict the blur kernel for each pixel. Defocus blur typically arises from a circular aperture, resulting in symmetric and uniformly distributed blur spots in all directions. When defocusing occurs, the optical characteristics of the lens usually cause light to scatter uniformly, indicating that the defocus blur kernel can be considered isotropic in nature. Initially, $\mathcal{R}_{kpn}$ predicts kernel seeds. The functionality of $\mathcal{R}_{kpn}$ can be described as

$$\boldsymbol{S} = \mathcal{R}_{kpn}(\hat{\boldsymbol{I}}, \boldsymbol{I}_B), \quad (3)$$

where the input consists of the concatenation of $\hat{\boldsymbol{I}}$ and $\boldsymbol{I}_B$, both having dimensions $H \times W \times 3$, while the output, denoted as $\boldsymbol{S}$, is a feature volume with dimensions $H \times W \times M$. The feature vector of size $1 \times 1 \times M$ corresponding to each position (u, v) is partitioned into a set of kernel seeds $\{s_i^{u,v} \mid i = 2, 3, ..., m,\ M = \sum_{i=2}^{m} i\}$ for further processing. These seeds are utilized to generate a collection of isotropic kernels. Each kernel $k_i^{u,v}$ represents a single-channel map with

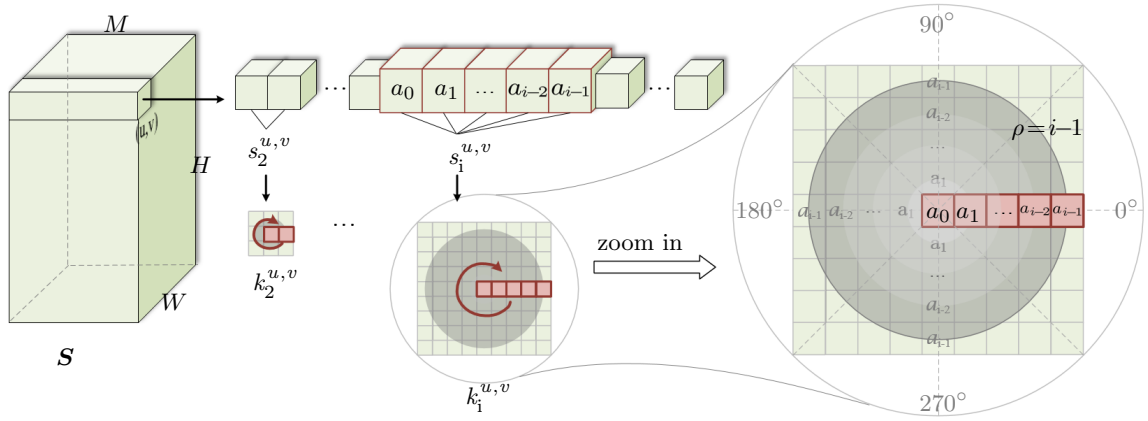


Fig. 5: Illustration of how isotropic blur kernels are generated. For a feature vector located at coordinate (u, v), it is first split into a set of kernel seeds $\{s_i^{u,v}\}$ and then converted to blur kernels $\{k_i^{u,v}\}$. Referring to Eq. (4), the value of each element in $k_i^{u,v}$ is interpolated in polar coordinates by considering the distance between this element and the center of $k_i^{u,v}$.

dimensions $(2i-1)\times(2i-1)$. The process of generating the kernels is illustrated in Fig. 5. Let $s_i^{u,v} = [a_0, a_1, ..., a_{i-1}]^{\mathrm{T}}$ denote the kernel seed of position (u, v), we explain in detail how the corresponding kernel $k_i^{u,v}$ is generated.

For a single element of $k_i^{u,v}$, its value is determined by its distance from the center of $k_i^{u,v}$ using interpolation in polar coordinates. Specifically, we first represent the elements of $k_i^{u,v}$ with the form (ρ, θ) in polar coordinates. To formalize this process, we represent the elements of $k_i^{u,v}$ as (ρ, θ) in polar coordinates

$$k_i^{u,v}(\rho,\theta)=\begin{cases} a_\rho \ , \ \text{if } \rho \le i-1 \text{ and } \rho \text{ is integer} \\ 0 \ , \ \text{if } \rho > i-1 \\ \dfrac{\rho-\lceil\rho\rceil}{\lfloor\rho\rfloor-\lceil\rho\rceil}a_{\lfloor\rho\rfloor}+\dfrac{\rho-\lfloor\rho\rfloor}{\lceil\rho\rceil-\lfloor\rho\rfloor}a_{\lceil\rho\rceil}, \ \text{else} \end{cases} \tag{4}$$

where $\lfloor\rfloor$ and $\lceil\rceil$ denote the floor and ceiling operations. The calculated kernel values are then normalized using a `Softmax` function, ensuring that $\sum_{\rho,\theta} k_i^{u,v}(\rho,\theta) = 1$ and $k_i^{u,v}(\rho,\theta) > 0$. Through these operations, we obtain $m-1$ pixel-wise isotropic defocus blur kernels of sizes $\{3\times 3, 5\times 5, ..., (2m-1)\times(2m-1)\}$ to account for different positions of $\hat{\boldsymbol{I}}$. Subsequently, $m-1$ blurry images $\{\hat{\boldsymbol{I}}_B^2, ..., \hat{\boldsymbol{I}}_B^m\}$ at various levels of blur can be generated by convolving $\hat{\boldsymbol{I}}$ with the corresponding blur kernels. It is important to note that the zero-blur-level image $\hat{\boldsymbol{I}}_B^1$ corresponds to the non-blurred image $\hat{\boldsymbol{I}}$.

Weight Prediction Network $\mathcal{R}_{wpn}$: The generation of weight maps to integrate the estimated blurry images is crucial due to the spatial variability of blur across a blurry image. The operation $\mathcal{R}_{wpn}$ can be mathematically expressed as

$$\boldsymbol{W} = \mathcal{R}_{wpn}(\hat{\boldsymbol{I}}, \boldsymbol{I}_B), \tag{5}$$

where the feature volume $\boldsymbol{W}$ is of size $H\times W\times m$ and is subsequently divided into m weight maps, each with a single channel $(\boldsymbol{W}_1, \boldsymbol{W}_2, ..., \boldsymbol{W}_m)$. These weight maps are then normalized using the `Softmax` function to ensure that the values at each position (u, v) sum up to 1, *i.e.*, $\sum_i \boldsymbol{W}_i^{u,v} = 1$, where $\boldsymbol{W}_i^{u,v} \ge 0$.

3.2.2 Reblurring Loss $\mathcal{L}_R$

The reblurred image can be reconstructed as

$$\hat{\boldsymbol{I}}_B = \sum_{i=1}^{m} \boldsymbol{W}_i * \hat{\boldsymbol{I}}_B^i. \tag{6}$$

The reblurring loss is specified as

$$\mathcal{L}_R = \sqrt{\| \boldsymbol{I}_B - \hat{\boldsymbol{I}}_B \|^2 + \varepsilon^2}, \tag{7}$$

where $\varepsilon = 1\times 10^{-3}$ is set in experiments. It is observed that the spatial consistency among $\hat{\boldsymbol{I}}$, $\hat{\boldsymbol{I}}_B$, and $\boldsymbol{I}_B$ is guaranteed due to the isotropic characteristics of the predicted blur kernels. In terms of

network architectures, simple networks consisting of fundamental convolutional layers and residual blocks are utilized for both $\mathcal{R}_{kpn}$ and $\mathcal{R}_{wpn}$, as depicted in Fig. 4.

The final issue lies in the acquisition of pseudo defocus map supervision $\boldsymbol{B}_P$. As shown in Eq. (6), a reblurred image can be considered as a weighted summation of blurred images generated by isotropic blur kernels. Consequently, the blur amount measurement at a pixel coordinate in the blurred image can be obtained through the utilization of weighted kernel seeds. Therefore, obtaining $\boldsymbol{B}_P$ with size $H \times W \times M$ involves assigning weights $\boldsymbol{W}$ with size $H \times W \times m$ to the kernel seeds $\boldsymbol{S}$ with size $H \times W \times M$.

3.2.3 Derivation of Pseudo Defocus Map

The final issue lies in the acquisition of pseudo defocus map supervision $\boldsymbol{B}_P$. As shown in Eq. (6), a reblurred image can be considered as a weighted summation of blurred images generated by isotropic blur kernels. Consequently, the blur amount measurement at a pixel coordinate in the blurred image can be obtained through the utilization of weighted kernel seeds. Therefore, obtaining $\boldsymbol{B}_P$ with size $H \times W \times M$ involves assigning weights $\boldsymbol{W}$ with size $H \times W \times m$ to the kernel seeds $\boldsymbol{S}$ with size $H \times W \times M$. We note that i-th kernel seed at a pixel coordinate has i values, and the summation $M = \sum_{i=2}^{m} i$ satisfies. To ensure dimension compatibility, each channel $\boldsymbol{W}_i$ with size $H \times W \times 1$ needs to be expanded to match the dimension $H \times W \times i$ of the corresponding i-th kernel seed in $\boldsymbol{S}$. Formally, the pseudo defocus map $\boldsymbol{B}_P$ can be obtained by

$$\boldsymbol{B}_P = \mathit{Repmat}(\boldsymbol{W}) \odot \boldsymbol{S}, \tag{8}$$

where $\odot$ is element-wise multiplication, and *Repmat* duplicates matrix $\boldsymbol{W}_i$ for i times to ensure that kernel seeds $\boldsymbol{s}_i$ and their corresponding weighting matrices $\boldsymbol{W}_i$ have matching dimensions. This process is shown in Fig. 6. Subsequently, pseudo defocus map $\boldsymbol{B}_P$ will be utilized to train the deblurring module as presented in Sec. 3.3.2, supervising the estimation of defocus blur map.

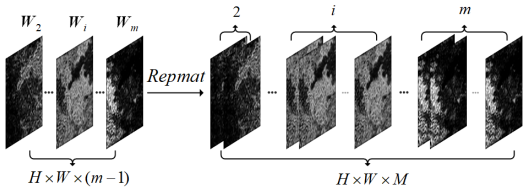


Fig. 6: Visualization of *Repmat*: The matrix $\boldsymbol{W}_i$ is replicated i times, enabling feasible element-wise multiplication with $\boldsymbol{S} \in \mathbb{R}^{H \times W \times M}$, where $M = \sum_{i=2}^{m} i$. Note that $\boldsymbol{W}_1$, which corresponds to the zero-blur-level image $\hat{\boldsymbol{I}}_B^1$, does not contribute to the blur amount, and thus is excluded before *Repmat* operation.

3.3 Deblurring Module

We present the deblurring network, and the training loss functions $\mathcal{L}_D$ and $\mathcal{L}_P$.

3.3.1 Deblurring Model

Upon existing deblurring network $\mathcal{D}$, *e.g.*, the CNN-based or Transformer-based deblurring networks, we introduce a blur map estimator $\mathcal{E}$ and a fusion block $\mathcal{F}$, by which the estimated blur map $\hat{\boldsymbol{B}} = \mathcal{E}(\boldsymbol{I}_B)$ and fused image $\boldsymbol{I}_f = \mathcal{F}(\boldsymbol{I}_B, \hat{\boldsymbol{B}})$. Taking the blurry image $\boldsymbol{I}_B$ as input, the deblurring module can be formulated as

$$\hat{\boldsymbol{I}} = \mathcal{D}(\boldsymbol{I}_f) = \mathcal{D}\left(\mathcal{F}\left(\boldsymbol{I}_B, \mathcal{E}(\boldsymbol{I}_B)\right)\right). \tag{9}$$

Since we do not modify the architecture of $\mathcal{D}$, we in the following only focus on blur map estimator $\mathcal{E}$ and fussion block $\mathcal{F}$.

Defocus Blur Map Estimator $\mathcal{E}$: The defocus blur map estimator is designed to predict the defocus blur map, denoted as $\hat{\boldsymbol{B}} = \mathcal{E}(\boldsymbol{I}_B)$. This lightweight network consists of 2 convolutional layers and 3 residual layers. However, since no direct supervision is available for the estimated defocus blur map $\hat{\boldsymbol{B}}$, it may result in significant deviation from the true defocus blur map. To address this issue, we utilize the pseudo ground-truth defocus maps generated by our reblurring module, as described in Sec. 3.2.3, and introduce a dedicated loss term $\mathcal{L}_P$ to optimize the estimator.

Fusion Block $\mathcal{F}$: Empirically, we observe that the defocus blur map and blurry image also have misalignment issues, especially for early training epochs. Therefore, we explore the fusion strategy by employing deformable attention mechanism as illustrated in Fig. 7 to effectively incorporate the degradation-related prior for image deblurring. The deformable attention mechanism allows

the model to focus on relevant spatial regions of the pseudo defocus blur maps while dynamically adjusting the attention weights based on the local features of the blurred image.

Given the estimated defocus blur map $\hat{\boldsymbol{B}} \in \mathbb{R}^{H\times W\times M}$ and the blurry input image $\boldsymbol{I}_B \in \mathbb{R}^{H\times W\times 3}$, we introduce a deformable cross-attention mechanism to adaptively fuse information across different feature representations. For each pixel coordinate (x, y) in the spatial domain, the fusion process of the deformable cross-attention is formulated as follows:

$$\boldsymbol{I}_f(x,y) = \boldsymbol{I}_B(x,y) + \boldsymbol{W}_o \sum_{n=1}^{N_p} \left(\boldsymbol{W}_{q(n)}\boldsymbol{I}_B(x,y)\right) \odot \left(\boldsymbol{W}_{v(n)}\hat{\boldsymbol{B}}(x+\Delta x_n,\ y+\Delta y_n)\right), \tag{10}$$

where $\odot$ is element-wise multiplication, and N_p is the number of deformable sampling points, which controls the spatial scope of attention. Specifically, $\boldsymbol{I}_B(x,y) \in \mathbb{R}^{3\times 1}$ denotes the feature vector of the input image $\boldsymbol{I}_B$ at position (x, y), while $\hat{\boldsymbol{B}}(x,y) \in \mathbb{R}^{M\times 1}$ represents the corresponding feature from the blur map $\hat{\boldsymbol{B}}$ sampled at position (x, y). The learnable weights $\boldsymbol{W}_{q(n)} \in \mathbb{R}^{M\times 3}$ and $\boldsymbol{W}_{v(n)} \in \mathbb{R}^{M\times M}$ are used to project features in blurry image and defocus blur map, respectively. Δx_n and Δy_n are the spatial offsets. $\boldsymbol{W}_o \in \mathbb{R}^{3\times M}$ is the output projection matrix that maps the aggregated features back to the original image dimension. The fused image $\boldsymbol{I}_f \in \mathbb{R}^{H\times W\times 3}$ can then be fed to existing deblurring network without modifying on architecture. This strategy overcomes the limitations of traditional fixed attention mechanisms by enabling more flexible integration of information from defocus blur maps, leading to better alignment of the blur prior with the image content.

The architecture of $\mathcal{F}$ employs 5 attention heads, with each head using 4 sampling points. Each attention head independently computes its attention weights and feature fusion process, and the outputs from all heads are subsequently merged that enables the model to focus on relevant spatial regions of the pseudo defocus blur maps. The parameters of $\mathcal{D}$, $\mathcal{E}$ and $\mathcal{F}$ can be learned by optimizing a single deblurring loss $\mathcal{L}_D$. Moreover,

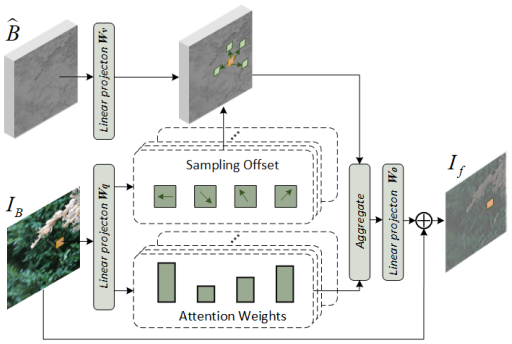


Fig. 7: Overview of our fusion block for integrating defocus map and input blurry image, which incorporates a deformable attention mechanism that better aligns the blur prior with the image content. The fused image $\boldsymbol{I}_f$ is of size $H \times W \times 3$ that can be fed to existing deblurring networks without architecture modification.

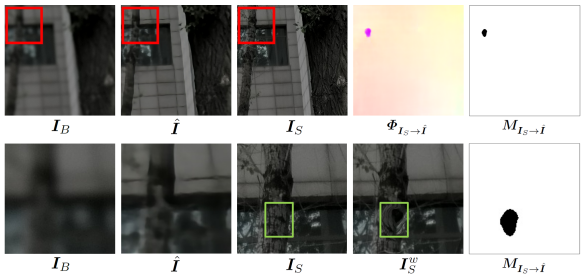


Fig. 8: Errors in optical flow estimation can incorrectly deform the sharp ground-truth image, *e.g.*, in areas marked by green rectangles. Calibration masks can help by filtering out this adverse region.

to supervise the estimated defocus blur map $\hat{\boldsymbol{B}}$, we also introduce the dedicated loss $\mathcal{L}_P$.

3.3.2 Deblurring Losses $\mathcal{L}_D$ and $\mathcal{L}_P$

To account for the spatial misalignment inherent in the training pairs, we incorporate an optical flow-based deformation. This approach allows the framework to accommodate potential misalignment between $\boldsymbol{I}_S$ and $\hat{\boldsymbol{I}}$. Specifically, we employ an optical flow estimation network $\mathcal{F}_{flow}$ [43] to estimate the optical flow $\boldsymbol{\Phi}_{\boldsymbol{I}_S\to\hat{\boldsymbol{I}}}$ from $\boldsymbol{I}_S$ to $\hat{\boldsymbol{I}}$

$$\boldsymbol{\Phi}_{\boldsymbol{I}_S\to\hat{\boldsymbol{I}}} = \mathcal{F}_{flow}(\boldsymbol{I}_S, \hat{\boldsymbol{I}}). \tag{11}$$

Then, $\boldsymbol{I}_S$ is deformed towards $\hat{\boldsymbol{I}}$ using the estimated optical flow

$$\boldsymbol{I}_S^w = \mathcal{W}(\boldsymbol{I}_S, \boldsymbol{\Phi}_{\boldsymbol{I}_S\to\hat{\boldsymbol{I}}}), \tag{12}$$

where $\mathcal{W}$ denotes linear warping operation [43].

We utilize a calibration mask, denoted as $\boldsymbol{M}$, to identify and exclude regions with inaccurate optical flow estimations. The process begins by calculating the average optical flow value, denoted by $\overline{\boldsymbol{\Phi}}$. Based on this, the calibration mask is defined as follows

$$\begin{aligned}\boldsymbol{M}_{\boldsymbol{I}_S\to\hat{\boldsymbol{I}}} = [(1-\lambda)\times\overline{\boldsymbol{\Phi}}_{\boldsymbol{I}_S\to\hat{\boldsymbol{I}}} < \boldsymbol{\Phi}_{\boldsymbol{I}_S\to\hat{\boldsymbol{I}}} \\ < (1+\lambda)\times\overline{\boldsymbol{\Phi}}_{\boldsymbol{I}_S\to\hat{\boldsymbol{I}}}],\end{aligned} \tag{13}$$

where the value of $\boldsymbol{M}$ is 1 if the condition in $[\cdot]$ is satisfied, and otherwise 0.

The effectiveness of Eq. (13) can be attributed to the slight variation in the misalignment between $\boldsymbol{I}_S$ and $\hat{\boldsymbol{I}}$ across different spatial positions, enabling the detection of inaccurate optical flow estimation through anomalies in its magnitude. This process is illustrated in Fig. 8. Additionally, we incorporate a cycle deformation strategy by calculating the optical flow $\boldsymbol{\Phi}_{\hat{\boldsymbol{I}}\to\boldsymbol{I}_S}$ in reverse order and applying reverse deformation, significantly enhancing the robustness of the deformation process. Finally, the adaptive deblurring loss can be calculated based on Charbonnier loss [40]

$$\begin{aligned}\mathcal{L}_D = &\sqrt{\| \boldsymbol{M}_{\boldsymbol{I}_S\to\hat{\boldsymbol{I}}} * (\boldsymbol{I}_S^w - \hat{\boldsymbol{I}}) \|^2 + \varepsilon^2} \\ &+ \sqrt{\| \boldsymbol{M}_{\hat{\boldsymbol{I}}\to\boldsymbol{I}_S} * (\hat{\boldsymbol{I}}^w - \boldsymbol{I}_S) \|^2 + \varepsilon^2},\end{aligned} \tag{14}$$

where $*$ is element-wise product, and ε is empirically set as 1×10^{-3} in all the experiments.

To address the significant deviation of the estimated defocus blur map $\hat{\boldsymbol{B}}$ from the true defocus blur map, we introduce an additional loss term $\mathcal{L}_P$ that uses the pseudo supervision $\boldsymbol{B}_P$ derived in Section 3.2.3 as supervision

$$\mathcal{L}_P = \sqrt{\| \boldsymbol{B}_P - \hat{\boldsymbol{B}} \|^2 + \varepsilon^2}. \tag{15}$$

3.4 A New Dataset for Image Defocus Deblurring

To validate the effectiveness of our approach on a specific device, we employed a HUAWEI X2381-VG camera to establish a new dataset for image defocus deblurring, referred to SDD dataset. This camera has adjustable DOF through vertical lens movement and aperture modulation, allowing us to collect pairs of blurry images and corresponding ground-truth sharp images from the same scene. Despite our best efforts to ensure proper alignment between the blurry images and ground-truth sharp images, the SDD dataset exhibits noticeable misalignment in training pairs compared to DPDD [1], as depicted in Fig. 1. This discrepancy can be attributed to adjustments made by using an electronic motor to manipulate the camera lens.

Fig. 9: Sample images from our SDD dataset. The first and second rows show the outdoor and indoor scenes respectively.

This dataset comprises 150 high-resolution blurry and sharp image pairs with dimensions 4096×2160. Sample images from the SDD dataset are shown in Fig. 9. These pairs are divided into 115 training pairs and 35 testing pairs. Similar to [1], the training image pairs are resized and cropped into 4,830 patches sized 512×512. The SDD dataset encompasses diverse indoor and outdoor scenes including 50 indoor scenes and 65 outdoor scenes in the training set, along with 11 indoor scenes and 24 outdoor scenes in the test set. Misalignment between blurry and sharp images occurs in two forms: zoom misalignment obtained by vertical camera movement and shift misalignment achieved through horizontal movement, referring to Fig. 1.

4 Experiments

Datasets: We evaluate the performance of our proposed method on four datasets: DPDD [1], RealDOF [2], DED [19] and our SDD.

- DPDD dataset was built by using a dual-pixel camera, capturing the defocus and all-in-focus pairs in two successive shots. It contains 350/76/74 image triplets for training/testing/-validation, respectively. Each blurry image is paired with a corresponding sharp image. We use 7,000 processed blurry-sharp pairs for training and 76 blurry images for testing.

- RealDOF dataset was constructed using a dual-camera system with a beam splitter, as described in [2]. It provides only a test set, consisting of 50 scenes, for evaluation.
- DED dataset was the first large-scale realistic dataset for defocus map estimation and defocus image deblurring. It comprises a total of 1,112 image pairs, with some sourced from the multi-view dataset [19] and others captured using a light field camera.
- SDD dataset contains 4,830 training pairs, and 35 image testing images with high resolution 4096×2160.

Evaluation Metrics: We report five metrics to quantitatively assess the compared methods: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [44], Learned Perceptual Image Patch Similarity (LPIPS) [45], Fréchet Inception Distance (FID) [46] and Deep Image Structure and Texture Similarity (DISTS) [47]. In defocusing deblurring scenarios with misaligned training data, existing methods may restore the image, but they often lead to severe distortions and truncations. Quantitative metrics, *e.g.*, PSNR and SSIM, are sensitive to pixel-level differences and may not fully capture the local structural improvements, such as edge sharpness and texture clarity, which are more perceptually significant. This aligns with the findings in image restoration research [46, 47], where perceptual quality is increasingly valued over strict pixel-level fidelity. The study [47] also analyzes the unsatisfactory performance of existing metrics on certain image restoration techniques, partially due to their low tolerance to spatial misalignment. Thus, in addition to PSNR and SSIM [44], we also calculated LPIPS [45], FID [46], and DISTS [47].

4.1 Implementation Details

At the beginning of training, the predicted optical flow $\mathbf{\Phi}_{\boldsymbol{I}_S \to \hat{\boldsymbol{I}}}$ and $\mathbf{\Phi}_{\hat{\boldsymbol{I}} \to \boldsymbol{I}_S}$ are hardly informative due to the low quality of $\hat{\boldsymbol{I}}$. Therefore, during the initial training stage, we calculate $\mathbf{\Phi}_{\boldsymbol{I}_S \to \hat{\boldsymbol{I}}}$ using $\mathcal{F}_{flow}(\boldsymbol{I}_S, \boldsymbol{I}_B)$ over T training epochs, and subsequently by Eq. (11). To ensure the quality of $\hat{\boldsymbol{I}}$, we empirically set T as 15 in our experiments. During training, we set λ as 0.35 to generate the calibration masks. The maximal radius of blur kernels m is set as 8, and the sampling points number N_p in deformable attention is set as 4. The trade-off parameters α and β in Eq. (2) are both set as 0.5.

All the experiments are conducted using PyTorch on two A100 GPUs. The input images, along with the corresponding sharp ground truths and defocus maps, are randomly cropped to a size of 512×512. The batch size is set to 1. The parameters are initialized using the strategy proposed by He et al. [48], and are optimized using the Adam optimizer [49] by setting $\beta_1 = 0.9, \beta_2 = 0.999$ and $\epsilon = 10^{-8}$. The learning rate is initialized as 2×10^{-5} and is halved every 60 epochs. The entire training stage ends with 200 epochs.

4.2 Evaluation on SDD Dataset

In this section, we compare our proposed defocus deblurring method against state-of-the-art single-image defocus deblurring approaches. we retrain state-of-the-art deblurring methods on the SDD dataset for further comparison, including DMPHN [50], MPRNet [51], UformerT [16], Restormer [15] and Loformer [22]. To validate the effectiveness of our method, We have applied the proposed framework to both CNN-based and Restormer-based deblurring methods. As shown in Table 1, our proposed framework is evaluated from two perspectives: (i) reblurring-based learning framework, *i.e.*, deblurring model+Ours(Eq. (1)), and (ii) reblurring-based blurring framework and deblurring blocks $\mathcal{E}$ and $\mathcal{F}$, *i.e.*, deblurring model+Ours(Eq. (2)). One can see that for plain UNet, PSNR gains +1.2dB and +1.58dB are obtained by Ours(Eq. (1)) and Ours(Eq. (2)), respectively. For Loformer with top performance, PSNR gains +0.44dB and +0.68dB are obtained by Ours(Eq. (1)) and Ours(Eq. (2)), respectively. Considering that UNet, Restormer and Loformer are representative network architectures, covering from plain CNN to top Transformer, we believe that our proposed framework can be successfully applied to improve existing deblurring models.

Since our dataset is designed for tackling misalignment tasks, directly calculating pixel-level metrics may not be accurate. Therefore, when testing all methods, we used an existing pretrained optical flow network to warp the sharp images $\boldsymbol{I}_S$, aligning them with the deblurred images $\hat{\boldsymbol{I}}$, and then calculated the evaluation metrics between $\hat{\boldsymbol{I}}$ and $\boldsymbol{I}_S^w$.

Table 1 summarizes the results of all competing methods on SDD dataset. In our conference paper [21], our method directly adopt MPRNet as the deblurring model, and achieves notable performance gains over MPRNet. In this work, our reblurring-guided learning frameworks Eq. (1) and Eq. (2) both obtain better deblurring performance, while Ours (Eq. (2)) further improves the quantitative metrics benefiting from the defocus map as degradation prior in the baseline deblurring network. Nevertheless, all the models trained by reblurring-guided framework can well handle the misalignment issues in training data.

In Table 2, we provide comparison of parameters and FLOPs for several representative methods, including UFormerT [16], MPRNet [51], Loformer [22] and Restormer [15]. Since Ours(Eq. (1)) does not modify the architecture of deblurring model, parameters and FLOPs of the methods with * are exactly same with their original models. For the methods with †, the blocks $\mathcal{E}$ and $\mathcal{F}$ introduce additional computational cost, but the increases are very slight, where only ∼0.1M parameters and ∼18G FLOPs are negligible in comparison to these deblurring models.

The visualizations shown in Fig. 10 prove that our method overcomes the limitations of misaligned training pairs, achieving superior restoration of fine details in the presence of strong defocus blur. In the green box in the fifth row, pixel misalignment leads to optimization discrepancies during training, causing severe distortion in the image restoration process of basic models such as DPDNet [1], MPRNet [51], and Restormer [15]. The edges of the table and chair, which should be straight, are restored into pronounced curves, which contradicts our visual perception. In the first row, a similar phenomenon can also be observed. Although Loformer [22] performs well in image restoration, it experiences local "collapse" in training tasks with misaligned data. In the green box, Loformer [22] exhibits a "truncation" phenomenon during the restoration process, resulting in poor visual quality. It is clear that this issue can be effectively addressed by our training framework.

4.3 Evaluation on DPDD Dataset

For the DPDD [1] dataset, besides the methods mentioned above, two state-of-the-art single image defocus deblurring methods Son et al. [24] and IFAN [2] are also compared. It is worth noting that although the DPDD dataset is meant to be aligned, there actually exists slight misalignment as shown in Fig. 1. Therefore, we calculate the evaluation metrics in view of $\hat{\boldsymbol{I}}$ *v.s.* $\boldsymbol{I}_S^w$. The experimental results are reported in Table 3.

By tolerating the slight misalignment existing in DPDD dataset [1], JDRL contributes to better performance on testing set of DPDD. Notably, the DPDD dataset was collected under strictly controlled conditions using remote control and strictly controlled capture conditions to ensure precise alignment. However, in real-world scenarios, it is impossible to achieve perfectly

Table 1: Quantitative comparison of competing methods on SDD dataset. The methods marked with notation $*$ are trained using Ours(Eq. (1)), and the methods marked with notation † are trained using Ours(Eq. (2)).

Method	PSNR↑	SSIM↑	LPIPS↓	FID↓	DISTS↓
UNet	24.62	0.758	0.344	81.82	0.212
DMPHN [50]	25.00	0.769	0.326	71.47	0.208
DPDNet$_S$ [1]	24.81	0.760	0.343	75.66	0.210
MPRNet [51]	26.28	0.796	0.302	62.32	0.202
UFormerT [16]	25.68	0.774	0.321	70.24	0.208
Restormer [15]	26.39	0.806	0.301	58.90	0.189
Loformer [22]	26.51	0.808	0.296	54.80	0.179
UNet*	25.82	0.783	0.305	58.53	0.181
MPRNet*	26.88	0.810	0.265	55.30	0.177
UformerT*	26.53	0.808	0.297	56.18	0.180
Restormer*	26.89	0.817	0.257	51.22	0.168
Loformer*	26.95	0.820	0.255	48.77	0.165
UNet†	26.20	0.796	0.298	55.49	0.180
Restormer†	27.08	0.819	0.253	50.64	0.166
Loformer†	**27.20**	**0.826**	**0.251**	**47.91**	**0.164**

Table 2: Comparison of computational costs of competing methods. For the methods marked with $*$, deblurring models are exactly same with their original ones. For the methods marked with †, deblurring module introduces $\mathcal{E}$ and $\mathcal{F}$.

Method	FLOPs(G)	Params(M)
MPRNet [51]	6830	20.1
UFormerT [16]	42.7	5.20
Restormer [15]	564	26.1
Loformer [22]	331	27.9
MPRNet*	6830	20.1
UformerT*	42.7	5.20
Restormer*	564	26.1
Loformer*	331	27.9
Restormer†	582	26.2
Loformer†	348	28.0

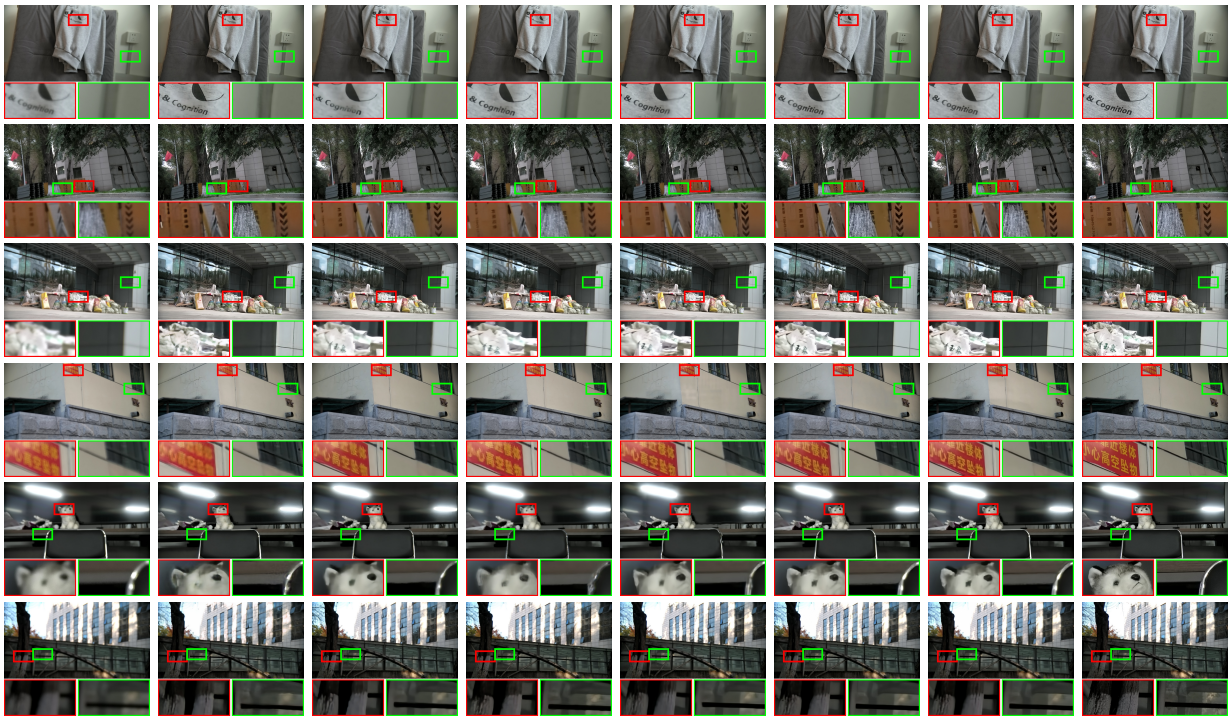


Fig. 10: Visual comparison of competing methods on SDD dataset.Loformer* is trained using Ours(Eq. (1)),while Loformer† is trained using ours(Eq. (2)).

Table 3: Quantitative comparison on DPDD dataset [1]. We only apply our framework on Restormer that achieves top performance on this dataset, resulting in Restormer†.

Method	PSNR↑	SSIM↑	LPIPS↓	FID↓	DISTS↓
DPDNet$_S$ [1]	24.48	0.778	0.262	76.46	0.158
Son et al. [24]	25.49	0.807	0.212	60.93	0.135
IFAN [2]	25.86	0.825	0.192	52.79	0.133
MPRNet [51]	26.03	0.820	0.214	60.68	0.140
Restormer [15]	26.65	0.850	0.158	45.57	0.104
Loformer [22]	26.10	0.840	0.197	53.24	0.126
Restormer†	**26.81**	**0.854**	**0.152**	**43.22**	**0.102**

aligned blurry-sharp pairs through manual control. Our misalignment training strategy can, to some extent, compensate for the discrepancies introduced during the data collection process.

4.4 Evaluation on DED Dataset

The DED dataset [19] provides training triplets, while the testing set only releases blurred images without their corresponding ground-truth, making it not possible to directly obtain the evaluation results in [19]. Additionally, the pre-trained model of DID-ANet [19] released by the authors is corrupted[1]. Therefore, for testing on the DED dataset [19], we randonly re-split the training set of the DED dataset into training and testing sets in a ratio 8:2, and DID-ANet [19] is trained by adopting their default training settings. Moreover, we trained Restormer and our model for comparison. The re-splitted dataset has also been released by us[2]. The results are presented in Table 4. Since DED dataset is captured by a Lytro camera, the training triplets are well aligned, it is reasonable that our method only achieves minor improvements than Restormer. The performance gain is mainly from the design of our baseline deblurring model. We also visualized our defocus blur maps in Fig. 12. We note that ground-truth defocus maps in DED dataset are also not truly captured, and instead it is estimated by from the light field images by Lytro camera. Since our defocus blur maps are with $M = 35$ dimension, we reduced

[1] https://github.com/xytmhy/DID-ANet-Defocus-Deblurring
[2] https://github.com/ssscrystal/Reblurring-guided-JDRL

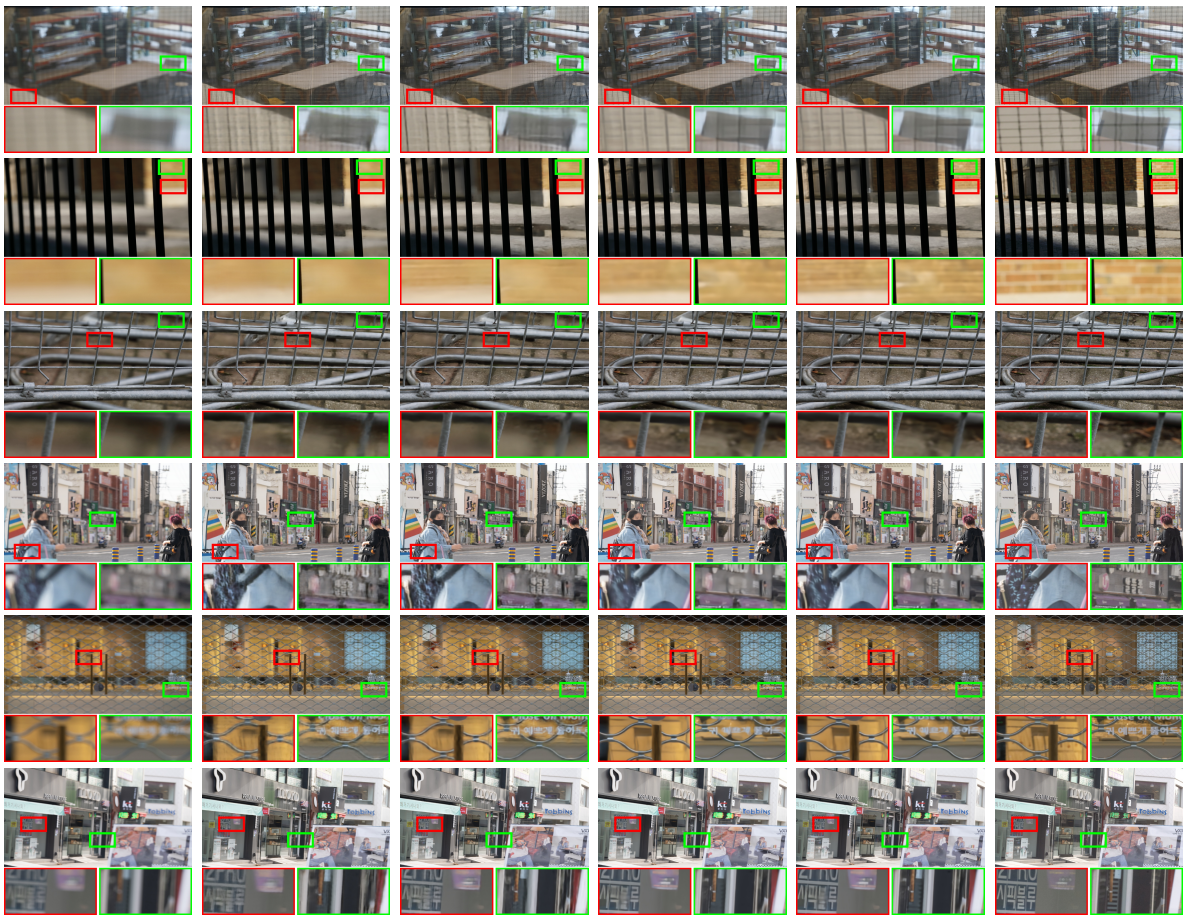


Fig. 11: Visual comparison of competing methods on the DPDD dataset and the RealDOF dataset. The first three rows are from the DPDD dataset, and the last three rows are from the RealDOF dataset. Our method demonstrates better visual results in terms of textures and structures.

Table 4: Quantitative comparison on DED dataset [19]. We note that the images from DED are generated from light field data captured by a Lytro camera, and are well aligned.

Method	PSNR↑	SSIM↑	LPIPS↓	FID↓	DISTS↓
DID-ANet [19]	29.59	0.860	0.186	72.97	0.193
Restormer [15]	31.29	0.890	0.121	58.74	0.141
Restormer†	**31.37**	**0.894**	**0.112**	**57.26**	**0.132**

their dimensionality from 35 to 1 by adopting principal component analysis. We can see that defocus map by our method can better reflect the spatially variant blur amount than that by DID-ANet.

4.5 Generalization Evaluation

Evaluation of Pre-trained models on Different Dataset. This section begins with an evaluation of the performance of models pre-trained on the DPDD [1] and DED datasets [19], *i.e.*, Restormer from Table 3 and Restormer from Table 4, when applied to images from our SDD dataset. Although defocus blur in the input blurry image is not severe as depicted in Figure 13, both Restormer models trained on DPDD and DED datasets are limited in removing defocus blur, due to the domain gaps between different sensors for capturing images. The limited generalization ability of trained models to new sensors emphasizes

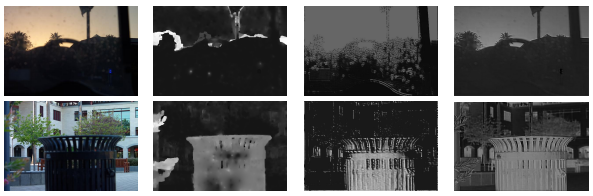


Image Defocus map DID-ANet [19] Ours

Fig. 12: Visual comparison of defocus blur maps. Since defocus blur map cannot be truly captured using a camera, it in this case is estimated from light field data captured by a Lytro camera. The defocus blur map estimated by Ours better reflects spatially variant blur amounts than that predicted by DID-ANet [19]. Note that our estimated defocus blur maps are with $H \times W \times M$ dimension with $M = 35$, and we reduced their dimensionality to $H \times W \times 1$ by adopting principal component analysis.

the necessity for a rapid data collection and training framework to enhance model adaptability and performance across diverse sensor types.

Evaluation of Generalization Ability. Furthermore, we follow existing works to adopt RealDOF dataset for evaluating generalization ability. We used the trained model from the DPDD dataset [1] for evaluation. As shown in Fig. 1, the RealDOF dataset also exhibits a certain degree of misalignment. Therefore, the strategy we designed for misaligned datasets also demonstrates effectiveness on the RealDOF dataset. Our evaluation results were calculated between the input images and the ground-truth images after warping, as presented in Table 5. Although trained on the DPDD dataset, Our learning framework shows robust generalization to other datasets. The experimental results also indicate that the domain gap between RealDOF and DPDD [1] is smaller compared to the gap between DPDD [1] and SDD. Through our framework, training with loosely aligned data pairs not only reduces data collection costs but also achieves better visual performance, providing an efficient and low-cost deblurring strategy for rapid adaptation to new sensors.

Cross-dataset Evaluation. Moreover, we perform cross-dataset evaluation. Within our reblurring-based learning framework, we utilize the UNet architecture to train models on three datasets, *i.e.,* SDD, DPDD [1], and DED [19]. As shown in Table 6, the model trained on DPDD [1] achieves the best performance, followed by our SDD-trained model, while the DED-trained model

Table 5: Quantitative comparison on RealDOF dataset [2].

Method	PSNR↑	SSIM↑	LPIPS↓	FID↓	DISTS↓
DPDNet$_S$ [1]	22.66	0.702	0.397	73.07	0.213
Son et al. [24]	24.35	0.756	0.308	64.57	0.187
IFAN [2]	25.39	0.794	0.264	43.39	0.162
MPRNet [51]	24.66	0.771	0.313	52.96	0.186
Restormer [15]	25.80	0.815	0.249	40.40	0.153
Loformer [22]	24.72	0.780	0.306	47.99	0.164
Restormer†	**25.98**	**0.819**	**0.220**	**39.69**	**0.151**

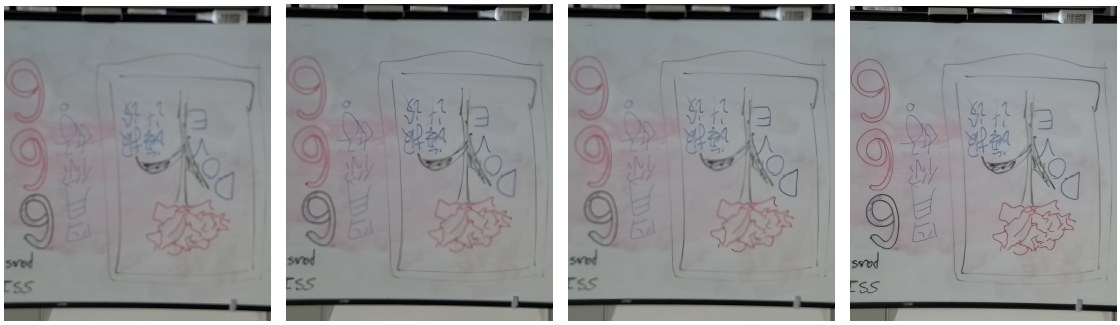
Table 6: Evaluation on RealDOF [2] by training UNet models on different datasets.

Training Data	PSNR↑	SSIM↑	LPIPS↓	FID↓	DISTS↓
DED [19]	21.81	0.672	0.414	79.20	0.221
DPDD [1]	22.28	0.687	0.401	74.27	0.216
SDD	22.03	0.680	0.406	75.41	0.219

exhibited the lowest performance. We attribute the observed experimental results primarily to the varying scales and data distribution of the training datasets. Specifically, DPDD [1] comprises 7,000 training pairs, DED [19] contains 1,112 pairs, and our SDD includes 4,830 pairs. We note that the primary contribution of our work lies in providing an approach for effective deployment on target devices by learning a device-specific image defocus deblurring model, while relaxing the strict requirement for perfect alignment in training pairs. Thus, our SDD serves as a testbed for addressing misalignment issues in training pairs and validating the effectiveness of the deblurring models and training strategies, without requiring generalization across different cameras. Consequently, it is acceptable that the model trained on our SDD (captured using a HUAWEI camera) does not achieve the best metrics on RealDOF (captured using a Google Pixel camera).

4.6 Ablation Study

To systematically validate the effectiveness of our proposed deblurring framework, we conduct comprehensive ablation studies from two perspectives: (1) the fusion model and (2) the reblurring-guided learning components. All experiments are performed on the SDD dataset using the UNet architecture as the backbone, with quantitative comparisons against variants that systematically



Blurry image | Restormer from Table 4 | Restormer from Table 3 | Ours

Fig. 13: An image with mild defocus blur from our SDD dataset is handled by two models trained on DPDD and DED datasets respectively, *i.e.*, Restormer from Table 3 and Restormer from Table 4. The limited generalization ability to new sensor emphasizes the necessity of an effective learning framework that supports misaligned training pairs.

Table 7: Comparison of different deformable modules for fusing defocus blur map and blurry image.

Method	PSNR↑	SSIM↑	LPIPS↓	FID↓	DISTS↓
DConv	26.01	0.787	0.303	57.68	0.183
LDConv	25.89	0.774	0.305	58.14	0.184
Ours	26.20	0.796	0.298	55.49	0.180

remove key modules. These studies aim to isolate the contributions of each component and verify the necessity of our design choices.

4.6.1 Fusion Model Effectiveness Analysis

To validate the superiority of our deformable cross-attention fusion model, we provide experimental comparison against these methods, *i.e.,* DConv [52], LDConv [53], and DAT [54].

- **DConv** learns per-position offsets within a fixed $k \times k$ grid, adapting locally to geometric variations in single-modality features. While we also employ learned offsets, we apply them in the Transformer attention domain, using query-driven reference points rather than a fixed convolution grid. This allows for more flexible, spatially-global alignment during fusion of blurred images and defocus maps.
- **LDConv** extends deformable convolution by enabling arbitrary sampling patterns and linear scaling of filter parameters, offering parameter-efficient but flexible sampling. Unlike LDConv's convolutional sampling and interpolation, our method leverages Transformer cross-attention to fuse features between modalities using deformable sampling guided by cross-modality queries—providing a more expressive and context-aware fusion mechanism.
- **DAT** learns shared offset groups to shift key/-value positions from a uniform reference grid, leading to data-dependent, sparse attention. We similarly employ query-conditioned deformable attention, but specifically tailor it to spatial alignment across two modalities (blurry image & defocus map). Our reference points are explicitly optimized to correct misalignment in defocus estimation, rather than general semantic aggregation.

As shown in Table 7, DConv and LDConv exhibit inferior performance compared to Ours, due to their limited capability in capturing long-range correspondence. Overall, deformable attention is likely to outperform deformable convolution, especially in scenarios with significant spatial misalignment, owing to its broader perception field. In Table 7, the results of DAT are omitted, because of its excessive GPU memory consumption, such as running out of memory on an 80G A100 GPU. The primary computational burden stems from the integration of local and deformable attention mechanisms, where local attention may not benefit resolving spatial misalignment when fusing defocus blur maps with blurry images.

4.6.2 Framework Components Ablation

To systematically evaluate the contributions of our reblurring-guided learning framework, we design ablation variants based on the UNet deblurring model, all trained on the SDD dataset: The effectiveness of the deblurring model involving pseudo defocus blur map for supervision has already been demonstrated in Table 1, where Ours with Eq. (2) outperforms Ours with Eq. (1). Our ablation studies are performed on the SDD dataset, and six variants are designed to analyze their contributions. Variant #1 represents a vanilla UNet trained using $\mathcal{L}_1$ loss. Variants #2 and #3 correspond to Ours with Eq. (1) by excluding the bi-directional optical flow deforming process and calibration mask, respectively. Variants #4, #5, and #6 aim to demonstrate the effectiveness of components within our reblurring module. Specifically, they variants of Ours with Eq. (1) by removing the reblurring module, isotropic kernels module $\mathcal{R}_{kpn}$ and weight prediction module $\mathcal{R}_{wpn}$ from, respectively.

Table 8: Ablation study on SDD dataset.

Variant	PSNR/SSIM/LPIPS
#1 w/ $\mathcal{L}_1$ loss	24.62/0.758/0.344
#2 w/ Warp Operation	25.51/0.776/0.322
#3 w/o Cycle Deformation	24.80/0.758/0.347
#4 w/o Calibration Mask	25.78/0.780/0.309
#5 w/o reblurring module	25.58/0.777/0.311
#6 w/o $\mathcal{R}_{kpn}$	25.68/0.778/0.311
#7 w/o $\mathcal{R}_{wpn}$	25.74/0.778/0.332
Ours (Eq. (1))	25.82/0.783/0.305
Ours (Eq. (2))	26.20/0.796/0.298

The quantitative results are presented in Table 8. Observations from experiments #1, #2 and #3 demonstrate that the absence of the deformation process significantly diminishes network performance. Moreover, the inclusion of the calibration mask appears to moderately enhance performance, as depicted in experiment #4. Experiment #5 clearly indicates that the lack of reblurring module substantially impacts overall network performance. The results of experiments #6 and #7 validate the advantages of isotropic blur kernels prediction module $\mathcal{R}_{kpn}$ and weight prediction module $\mathcal{R}_{wpn}$ within reblurring module for the reblurring process.

5 Conclusion

In this paper, we propose a reblurring-guided learning framework, designed specifically to tackle the significant challenge of misalignment in training pairs for single image defocus deblurring. The proposed method is distinctively composed of a deblurring module that integrates prior knowledge through a lightweight prediction model and a bi-directional optical flow-based deformation technique. This enables the framework to adeptly accommodate spatial misalignments between training pairs. Furthermore, the spatially variant reblurring module plays a pivotal role in reblurring the deblurred output to achieve spatial alignment with the original blurry image, utilizing predicted isotropic blur kernels and generating weighting maps for this purpose. Moreover, we also establish a new single image defocus deblurring dataset to evaluate our method and benefit future research. Extensive results on SDD, DPDD, DED and RealDOF datasets validate the effectiveness of our method in comparison to state-of-the-art methods.

References

[1] Abdullah Abuolaim and Michael S Brown. Defocus deblurring using dual-pixel data. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 111–126. Springer, 2020.

[2] Junyong Lee, Hyeongseok Son, Jaesung Rim, Sunghyun Cho, and Seungyong Lee. Iterative filter adaptive network for single image defocus deblurring. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2034–2042, 2021.

[3] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 30(11):3212–3232, 2019.

[4] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3):257–276, 2023.

[5] Constantine P Papageorgiou, Michael Oren, and Tomaso Poggio. A general framework for

object detection. In *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*, pages 555–562. IEEE, 1998.

[6] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. The secrets of salient object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 280–287, 2014.

[7] Joao Carreira and Cristian Sminchisescu. Constrained parametric min-cuts for automatic object segmentation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3241–3248. IEEE, 2010.

[8] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 447–456, 2015.

[9] Amit Goldstein and Raanan Fattal. Blur-kernel estimation from spectral irregularities. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 622–635. Springer, 2012.

[10] Jinsun Park, Yu-Wing Tai, Donghyeon Cho, and In So Kweon. 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 (CVPR)*, pages 1736–1745, 2017.

[11] Jianping Shi, Li Xu, and Jiaya Jia. Just noticeable defocus blur detection and estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 657–665, 2015.

[12] Dilip Krishnan and Rob Fergus. Fast image deconvolution using hyper-laplacian priors. *Advances in neural information processing systems*, 22, 2009.

[13] DA Fish, AM Brinicombe, ER Pike, and JG Walker. Blind deconvolution by means of the richardson–lucy algorithm. *JOSA A*, 12(1):58–65, 1995.

[14] Abdullah Abuolaim, Mauricio Delbracio, Damien Kelly, Michael S Brown, and Peyman Milanfar. Learning to reduce defocus blur by realistically modeling dual-pixel data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2289–2298, 2021.

[15] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022.

[16] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17683–17693, 2022.

[17] Yuelin Zhang, Pengyu Zheng, Wanquan Yan, Chengyu Fang, and Shing Shin Cheng. A unified framework for microscopy defocus deblur with multi-pyramid transformer and contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11125–11136, 2024.

[18] Yawei Li, Yuchen Fan, Xiaoyu Xiang, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Efficient and explicit modelling of image hierarchies for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18278–18289, 2023.

[19] Haoyu Ma, Shaojun Liu, Qingmin Liao, Juncheng Zhang, and Jing-Hao Xue. Defocus image deblurring network with defocus map estimation as auxiliary task. *IEEE Transactions on Image Processing*, 31:216–226, 2022.

[20] Ren Ng, Marc Levoy, Mathieu Brédif, Gene Duval, Mark Horowitz, and Pat Hanrahan. *Light field photography with a hand-held plenoptic camera*. PhD thesis, Stanford university, 2005.

[21] Yu Li, Dongwei Ren, Xinya Shu, and Wangmeng Zuo. Learning single image defocus deblurring with misaligned training pairs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1495–1503, 2023.

[22] Xintian Mao, Jiansheng Wang, Xingran Xie, Qingli Li, and Yan Wang. Loformer: Local

frequency transformer for image deblurring. In *ACM Multimedia 2024*, 2024.

[23] Li Xu and Jiaya Jia. Two-phase kernel estimation for robust motion deblurring. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part I 11*, pages 157–170. Springer, 2010.

[24] Hyeongseok Son, Junyong Lee, Sunghyun Cho, and Seungyong Lee. Single image defocus deblurring using kernel-sharing parallel atrous convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2642–2650, 2021.

[25] Wenda Zhao, Guang Hu, Fei Wei, Haipeng Wang, You He, and Huchuan Lu. Attacking defocus detection with blur-aware transformation for defocus deblurring. *IEEE Transactions on Multimedia*, 2023.

[26] Lufei Chen, Xiangpeng Tian, Shuhua Xiong, Yinjie Lei, and Chao Ren. Unsupervised blind image deblurring based on self-enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25691–25700, 2024.

[27] Xiaole Tang, Xile Zhao, Jun Liu, Jianli Wang, Yuchun Miao, and Tieyong Zeng. Uncertainty-aware unsupervised image deblurring with deep residual prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9883–9892, 2023.

[28] Dongwon Park, Byung Hyun Lee, and Se Young Chun. All-in-one image restoration for unknown degradations using adaptive discriminative filters for specific degradations. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5815–5824. IEEE, 2023.

[29] Vaishnav Potlapalli, Syed Waqas Zamir, Salman H Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one image restoration. *Advances in Neural Information Processing Systems*, 36, 2024.

[30] Yuang Ai, Huaibo Huang, Xiaoqiang Zhou, Jiexiang Wang, and Ran He. Multimodal prompt perceiver: Empower adaptiveness generalizability and fidelity for all-in-one image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25432–25444, 2024.

[31] Jingbo Lin, Zhilu Zhang, Yuxiang Wei, Dongwei Ren, Dongsheng Jiang, Qi Tian, and Wangmeng Zuo. Improving image restoration through removing degradations in textual representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2866–2878, 2024.

[32] Hao Yang, Liyuan Pan, Yan Yang, Richard Hartley, and Miaomiao Liu. Ldp: Language-driven dual-pixel image defocus deblurring network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24078–24087, 2024.

[33] Li Pang, Xiangyu Rui, Long Cui, Hongzhong Wang, Deyu Meng, and Xiangyong Cao. Hir-diff: Unsupervised hyperspectral image restoration via improved diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3005–3014, 2024.

[34] Yuhui Quan, Xin Yao, and Hui Ji. Single image defocus deblurring via implicit neural inverse kernels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12600–12610, 2023.

[35] Yuhui Quan, Zicong Wu, and Hui Ji. Neumann network with recursive kernels for single image defocus deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5754–5763, 2023.

[36] Libin Sun, Sunghyun Cho, Jue Wang, and James Hays. Edge-based blur kernel estimation using patch priors. In *IEEE international conference on computational photography (ICCP)*, pages 1–8. IEEE, 2013.

[37] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

[38] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Bjorn Stenger, Wei Liu, and Hongdong Li. Deblurring by realistic blurring.

In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2737–2746, 2020.

[39] Huaijin Chen, Jinwei Gu, Orazio Gallo, Ming-Yu Liu, Ashok Veeraraghavan, and Jan Kautz. Reblur2deblur: Deblurring videos via self-supervised learning. In *2018 IEEE International Conference on Computational Photography (ICCP)*, pages 1–9. IEEE, 2018.

[40] Pierre Charbonnier, Laure Blanc-Feraud, Gilles Aubert, and Michel Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In *Proceedings of 1st International Conference on Image Processing*, volume 2, pages 168–172. IEEE, 1994.

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 234–241. Springer, 2015.

[42] Qian Ye, Masanori Suganuma, and Takayuki Okatani. Accurate single-image defocus deblurring based on improved integration with defocus map estimation. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 750–754. IEEE, 2023.

[43] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018.

[44] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

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

[46] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6228–6237, 2018.

[47] Jinjin Gu, Haoming Cai, Haoyu Chen, Xiaoxing Ye, Ren Jimmy S, and Chao Dong. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 633–651. Springer, 2020.

[48] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.

[49] Kingma DP and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. of the 3rd International Conference for Learning Representations (ICLR)*, 2015.

[50] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14821–14831, 2021.

[51] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Deep stacked hierarchical multi-patch network for image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5978–5986, 2019.

[52] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.

[53] Xin Zhang, Yingze Song, Tingting Song, Degang Yang, Yichen Ye, Jie Zhou, and Liming Zhang. Ldconv: Linear deformable convolution for improving convolutional neural networks. *Image and Vision Computing*, 149:105190, 2024.

[54] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4794–4803, 2022.