A Unified Framework for Microscopy Defocus Deblur with Multi-Pyramid Transformer and Contrastive Learning

Yuelin Zhang[1] Pengyu Zheng[1] Wanquan Yan[1] Chengyu Fang[2] Shing Shin Cheng[1†]

[1] Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong

[2] Shenzhen International Graduate School, Tsinghua University

Abstract

Defocus blur is a persistent problem in microscope imaging that poses harm to pathology interpretation and medical intervention in cell microscopy and microscope surgery. To address this problem, a unified framework including the multi-pyramid transformer (MPT) and extended frequency contrastive regularization (EFCR) is proposed to tackle two outstanding challenges in microscopy deblur: longer attention span and data deficiency. The MPT employs an explicit pyramid structure at each network stage that integrates the cross-scale window attention (CSWA), the intra-scale channel attention (ISCA), and the feature-enhancing feed-forward network (FEFN) to capture long-range cross-scale spatial interaction and global channel context. The EFCR addresses the data deficiency problem by exploring latent deblur signals from different frequency bands. It also enables deblur knowledge transfer to learn cross-domain information from extra data, improving deblur performance for labeled and unlabeled data. Extensive experiments and downstream task validation show the framework achieves state-of-the-art performance across multiple datasets. Project page: https://github.com/PieceZhang/MPT-CataBlur.

1. Introduction

Microscope offers observers enhanced resolution and magnification [9, 48, 49], which greatly promotes the advancement of cell microscopy [9] and surgical microscopy [48]. Cell microscopy employs various optical techniques to reveal the structure and function of cells [9]. Surgical microscopy assists surgeons in performing delicate operations [48] including neurosurgery [81], ophthalmology [4], dentistry [15], etc. In microscopy, out-of-focus, or defocus, is one of the most common visual impairments caused by inferior optical quality, lens aperture, or object magnification [9, 48], resulting in blurred or distorted imaging. It poses harm to the downstream tasks [88], including segmentation [29, 30], detection [61], and classification [6]. While various microscopes with auto-focusing [35, 55, 79], assisted-focusing [67], or multi-focus [39, 82] capabilities have been developed to mitigate the defocus effect on-site, image degradation remains if the objects are distributed non-uniformly and not co-planar [50], or the cavities are too deep to be aligned with the focal plane [48]. Microscopy defocus deblur methods have thus been introduced as an offsite restoration approach.

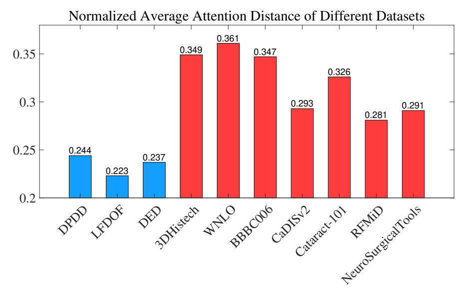

Figure 1. Normalized average attention distance of different datasets. The distance of real-world datasets (shown in blue) is significantly smaller than that of microscopy datasets (shown in red), showing the inter-domain feature difference.

Recent advances in deep learning have led to the development of various deep defocus deblur methods [34, 56, 57, 60], including those designed for microscopy [17, 18, 28, 32, 41, 50, 70, 72, 85, 91, 93]. Microscopy deblurring poses different challenges from real-world deblur tasks, due to the significant discrepancy between the features in the microscope images and natural scene images [88]. This difference can be quantified by calculating the normalized average attention distance for different datasets (attention intensity weighted by pixel distance then normalized by image size). The evaluation involves real-world datasets (DPDD [1], LFDOF [59], DED [47]), cell microscopy datasets (3DHistech [17], WNLO [17], BBBC006 [44]), and surgical microscopy datasets for cataract surgery

[†]Corresponding author.

(CaDISv2 [19], Cataract-101 [62]), retinal microsurgery (RFMiD [54]), neurosurgery (NeuroSurgicalTools [5]). As shown in Fig. 1, all cell microscopy datasets have normalized average attention distance around 0.35, and surgical microscope datasets around 0.3, indicating a much longer attention span than the real-world datasets at under 0.25. This result reveals the substantial discrepancy between these two domains, suggesting that **modeling attention in wider areas with larger receptive fields would benefit microscopy tasks**. Motivated by this analysis, we introduce a multi-pyramid transformer (MPT) with cross-scale window attention (CSWA), intra-scale channel attention (ISCA), and feature-enhancing feed-forward network (FEFN), to construct multiple pyramids explicitly on each stage of the network, fully exploiting latent cross-scale features in every projection space. CSWA captures the interaction between local $query$ and cross-scale key-$value$ pairs for long-range attention modeling with a quadratically enlarged receptive field while keeping computational efficiency. ISCA builds channel-wise attention on a local scale to provide global channel context, which is then integrated with the spatially correlated feature from CSWA by the proposed FEFN through an asymmetric activation mechanism.

Another problem in microscopy deblur is the insufficient data for training a robust model. Different from natural scene defocus deblur methods that use datasets captured with varying aperture sizes [1] or light field camera [47, 60], the high-quality training data for microscopy deblur can be much harder to obtain[17, 48]. For cell microscopy, insufficient training feature leads to a generalizability problem caused by different staining and imaging methods [17, 88]. The situation is worse for surgical microscopy because the imaging principle of microscope makes it impossible to simultaneously acquire blur-sharp pairs for model training [18, 48]. To alleviate the data deficiency problem, some training diagrams learn rich information by extending extra training data and then fine-tuning it with testing data [60]. This paradigm, however, may not be applicable to microscopy deblurring, as there is an **inter-domain gap** between natural scenes and microscopy images, and also **intra-domain gaps** among different microscopy datasets that are highly task-specific. The extended frequency contrastive regularization (EFCR) is proposed to address the data deficiency problem by encouraging the model to learn representations from decoupled frequency bands in the wavelet domain [23, 86], and further exploiting latent information leveraging the fact that model trained with synthetic reblurring images can deblur its naturally blurred counterpart [18]. It also enables cross-domain deblur knowledge transfer, facilitating multiple scenarios including extra data training and unlabeled data deblur.

This paper presents a unified deblur framework with MPT and EFCR to address the aforementioned two challenges in microscopy deblur. The surgical microscopy deblur is illustrated on cataract surgery, which is the most common surgery worldwide [13, 21, 66]. Extensive experiments are conducted on various open-source cell and surgical microscopy datasets, along with downstream tasks validation on cell detection and surgery scene semantic segmentation. For surgical microscopy deblur, we present a realistic blur synthesizing method, and collect a new dataset of defocus cataract microscopic surgery, which is the first dataset for surgical microscopy deblur. The method achieves state-of-the-art performance on not only microscope datasets but also real-world datasets, showing the universality of the proposed framework. The deblur results on unlabeled datasets also prove the effectiveness of the proposed EFCR on knowledge transferring. The main contributions are as follows, 1) The multi-pyramid transformer (**MPT**) is for the first time proposed for microscopy defocus deblur. It models the long-range spatial attention between local-scale and down-scale maps in each explicit pyramid using the proposed cross-scale window attention (**CSWA**) with a quadratically enlarged receptive field to adapt to the longer attention span of microscopy datasets. 2) The intra-scale channel attention (**ISCA**) is presented to incorporate global channel context in the CSWA spatial information via the proposed feature-enhancing feed-forward network (**FEFN**), providing additional intra-scale channel features to the pyramid. 3) A training strategy with extended frequency contrastive regularization (**EFCR**) is presented to alleviate data deficiency by exploiting latent deblur signal beyond the pixel constraint through synthetic reblurring, which is the first implementation of contrastive learning in microscopy deblur. It also enables cross-domain deblur knowledge transfer, facilitating extra data training and enhancing unlabeled image deblur.

2. Related Work

Single image defocus deblurring For learning-based deblur model, end-to-end method is widely applied [34, 56, 57, 60, 65] for its better performance and robustness [56] than methods based on defocus map estimation [33, 34, 47, 59]. Among them, many deblurring works have been done on cell microscopy to alleviate defocus brought by mechanical axial shift [70] and non-coplanar cells [50], and enhance the imaging quality of human cell [50, 72], pathology [17, 28], parasite [85], etc. Compared with cell microscopy, surgical microscopy deblur has not been well-explored due to the difficult acquisition of blur datasets with ground truth [18], although defocus blur is commonly encountered in microscopic surgery [18, 93].

Multi-scale methods The image pyramid in existing methods is usually built in two ways, i.e., explicitly stacking multi-scale feature maps in a single pyramid [16, 51, 80], or

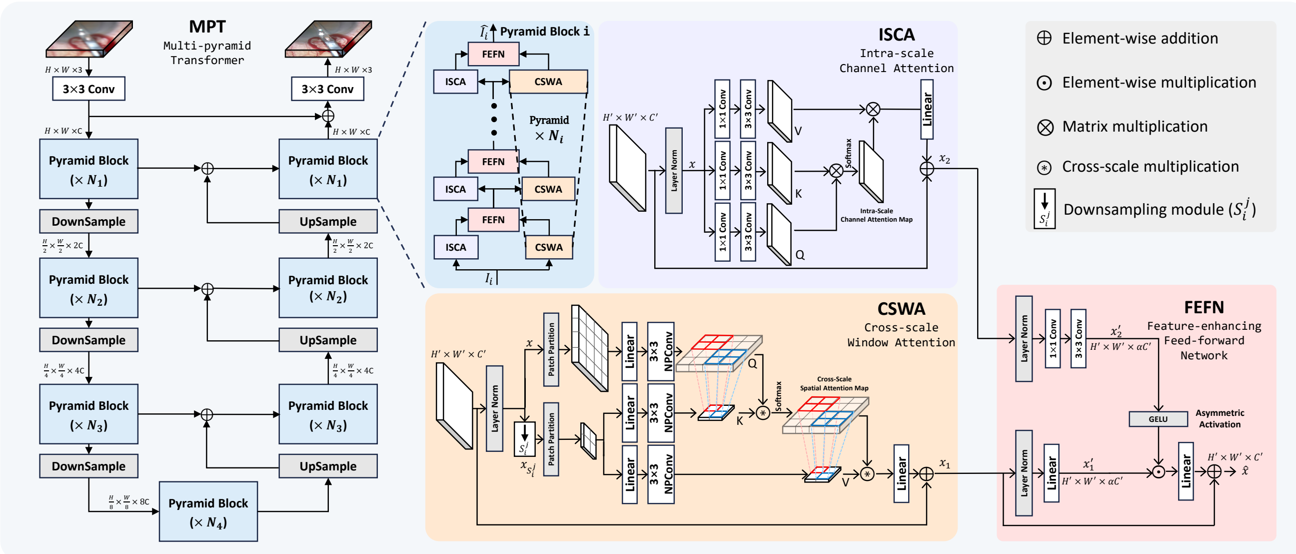

Figure 2. Overview of MPT. MPT constructs an explicit pyramid block at each stage. Inside the pyramid block, CSWAs constitute a coarse-to-fine pyramid, exploring cross-scale spatial interaction for each scale. The ISCA is built beside each CSWA to provide global channel context. Information from CSWA and ISCA is aggregated by FEFN using the asymmetric activation mechanism.

implicitly applying multi-stage structure [7, 8, 31, 56, 60, 71]. Explicit pyramid methods face single-level feature deficiency since the explicit pyramid is built on downscaling features in a single latent space [16, 51, 80]. Most of the existing microscope deblur methods [17, 18, 40, 70] adopt implicit multi-stage design to perform aggregation on separated latent space but suffer from inter-level feature discrepancy. The proposed MPT in this work addresses these drawbacks by building **multiple explicit pyramids** with CSWA, ISCA, and FEFN on **each feature level**, achieving cross-scale feature aggregation with an enlarged receptive field.

Contrastive learning Contrastive learning has been widely applied in low-level tasks [3, 14, 25, 27, 75, 92, 94]. They construct contrastive pairs on the feature space that take clean and corrupted images as positive and negative pairs, respectively [37, 76]. Some works have leveraged contrastive frequency information [94] and extracted frequency representations by wavelet transformation [3, 75, 86]. Following the fact that sharp and blurry images have similar low-frequency components but differ significantly in the high-frequency part [75], the idea of comparing different frequency bands separately is adopted in [90] by applying $\mathcal{L}_1$ loss directly to the frequency bands in contrastive regularization. In this paper, the proposed EFCR adopts basic CR and extended CR to encourage learning latent deblur signals and transferring cross-domain deblur information, thus addressing the data deficiency problem.

3. Method

The proposed framework consists of MPT and EFCR to address the two outstanding problems in microscopy defocus deblur, namely longer attention span and data deficiency.

3.1. Multi-pyramid Transformer (MPT)

MPT builds multiple explicit pyramids on each feature level, thus avoiding single-level feature deficiency and inter-level feature discrepancy [18, 51, 70, 80]. As shown in Fig. 2, the proposed MPT follows a U-shaped structure [22, 58, 74]. The blur image I_{in} with the size of $H \times W$ first gets input feature projection $I_1 \in \mathbb{R}^{H \times W \times C}$ through a convolution. Then, the feature goes through the network with seven pyramid blocks followed by re-sampling using pixel shuffles [64], and finally projects back to the image with a convolution. The shortcut connection is built between each encoder and decoder stage by element-wise addition.

Pyramid block The pyramid block i receives $I_i \in \mathbb{R}^{\frac{H}{2^i} \times \frac{W}{2^i} \times 2^i C}$ ($i \in \{1, 2, 3, 4\}$), and outputs the map $\hat{I}_i$ in the same size. There are N_i sub-blocks in pyramid block i. Each of them handles a local scale S_i^j ($j \in \{1, 2, ..., N_i\}$, $S_i^j \in \{\frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1\}$). Multiple stacked sub-blocks build the pyramid in a coarse-to-fine manner, i.e., $S_i^k \leq S_i^{k+1}$, $1 \leq k < N_i$. This design achieves progressive multi-scale feature aggregation and ensures the full exploration of each scale. In practice, N_i is set to be an even number. The sub-block in the even level adopts the common cyclic window shifting strategy [43] to gain cross-window interaction.

Cross-scale window attention (CSWA) CSWA captures the long-range interaction by modeling the attention between windows from the local-scale map and the down-scale map. The layer normalized [2] feature $x \in \mathbb{R}^{H' \times W' \times C'}$ first passes through a downsampling module to get $x_{S_i^j}$ with the size of $H'S_i^j \times W'S_i^j \times C'$. A strided average pooling followed by a linear projection with a shortcut is adopted in the downsampling module. The 3×3 neighboring padding convolution (NPConv) is proposed to generate Q, K, V projection ($Q \in \mathbb{R}^{\frac{H'W'}{M^2} \times M^2 \times C'}$, $K, V \in \mathbb{R}^{\frac{H'W'(S_i^j)^2}{M^2} \times M^2 \times C'}$) with inductive bias[20, 77], where M is the patch width. It pads a patch with its neighborhood pixels, providing the isolated edge pixels with neighboring information [69] (all 3×3 convolutions adopt bias-free grouped convolution by default with group size equal to the feature dimension). To get the cross-scale spatial attention map, the cross-scale multiplication ($\circledast$) is introduced as a one-to-many strategy, where each patch in K is multiplied with $\frac{1}{S_i^j} \times \frac{1}{S_i^j}$ patches in the corresponding location of Q, as illustrated in Fig. 2. This operation provides an $M \times M$ local patch in Q with interaction with a $M \times M$ patch in K whose information comes from a $\frac{M}{S_i^j} \times \frac{M}{S_i^j}$ region in the original map. Although the downscaled map loses information, it can still be highly instructive, since attention distributions of different scales are highly consistent [36]. It also leverages the pool of sharper patches generated by downscaling which serve as *priors* for deblurring [53, 95]. It makes the receptive field quadratically enlarged by $(S_i^j)^2$ times while keeping the computational complexity $O(M^2H'W'C')$ unchanged as the vanilla local window attention [43]. A similar strategy is adopted in multiplying the attention map and V. The self-attention in CSWA for a local window in the size of $M^2 \times C$ can be defined as:

$$Attention_1(q, k, v) = Softmax(qk^T/\sqrt{d} + B)v, \quad (1)$$

where $q, k, v \in \mathbb{R}^{M^2 \times d}$, and B is the relative positional encoding [43]. In practice, we implement the multi-head self-attention [68] by concatenating the result of h parallelly calculated attention. The output x_1 is then obtained by a linear projection with a shortcut.

Intra-scale channel attention (ISCA) ISCA handles a single scale feature $x \in \mathbb{R}^{H' \times W' \times C'}$ and generates cross-channel interaction with encoded global context [84]. By applying 1×1 convolutions followed by 3×3 convolutions, projections $Q, K, V \in \mathbb{R}^{H'W' \times C'}$ are generated from layer normalized x, introducing convolutional inductive bias and extracting cross-channel information in both point-wise and spatial-wise manners. The intra-scale channel attention map is then calculated by multiplying Q with K. The self-attention in ISCA can be formulated as:

$$Attention_2(Q, K, V) = Softmax(QK^T)V \quad (2)$$

Similar to CSWA, ISCA implements the multi-head self-attention [68] to get x_2.

Feature-enhancing feed-forward network (FEFN) The FEFN aggregates the spatial-wise feature x_1 with the channel-wise context x_2. The input features are first projected to x_1' and x_2' in the size of $H' \times W' \times \alpha C'$, where α is the expansion ratio. Instead of combining x_1' and x_2' by simply adding them together like [36], FEFN adopts an asymmetrical activation mechanism with GELU [26], where x_1 is element-wisely multiplied by GELU activated x_2. The FEFN can be formulated as

$$\hat{x} = W_p(GELU(x_2') \odot x_1') + x_1, \quad (3)$$

where $\hat{x} \in \mathbb{R}^{H' \times W' \times C'}$, and W_p refers to linear projection. Compared with the regular FN [12], this asymmetrical operation allows the spatial information from x_1 to be guided by the non-linearly activated signal from the channel context in x_2, offering x_1 an extra global view in terms of feature channels.

3.2. Extended Frequency Contrastive Regularization (EFCR)

The proposed EFCR contains basic CR and extended CR to explore latent deblur guidance beyond pixel constraints.

Constructions of contrastive pairs Given a training pair i with ground truth I_i^{gt}, blur input I_i^{in}, and deblurred output I_i^{out}, the Haar wavelet transformation [86] decouples the samples into low-low (LL), low-high (LH), high-low (HL), and high-high (HH) bands. For simplicity, here we define $f^h(\cdot)$ as the operator decoupling and concatenating high-frequency bands (LH, HL, HH), and $f^l(\cdot)$ as the operator for low-frequency band (LL). For basic CR, the frequency bands are directly taken as contrastive pairs. The positive and negative basic CR $\mathcal{L}_i^+$ and $\mathcal{L}_i^-$ are given by:

$$\mathcal{L}_i^+ = \|f^h(I_i^{out}) - f^h(I_i^{gt})\|_1 + \|f^l(I_i^{out}) - f^l(I_i^{gt})\|_1, \quad (4)$$

$$\mathcal{L}_i^- = \|f^h(I_i^{out}) - f^h(I_i^{in})\|_1. \quad (5)$$

Both bands are included in $\mathcal{L}_i^+$, since both high and low frequencies of I_i^{out} are expected to be pulled closer to I_i^{gt}. Only high frequency is taken for $\mathcal{L}_i^-$ to push the $f^h(I_i^{out})$ away from $f^h(I_i^{in})$ as blur degradation mainly happens in the high-frequency parts [8, 42].

The extended CR enforces the model to learn latent information from degraded high-frequency components beyond the pixel-wise constraint. Based on the idea that a model trained with synthetic blurred images can deblur natural blurry images in the dataset [18], the blurred image B_i^{in} is generated by applying random Gaussian blur (kernel size in $\{3, 5, 7\}$) on I_i^{in}, followed by calculating its deblurred result B_i^{out}. The extended CR $\mathcal{L}_i^{ext}$ based on extended training pair (B_i^{in}, B_i^{out}) is then formulated as:

$$\mathcal{L}_i^{ext} = \frac{\|f^h(B_i^{out}) - f^h(B_i^{in})\|_1}{\|f^h(I_i^{in}) - f^h(B_i^{in})\|_1}. \quad (6)$$

The $\mathcal{L}_i^{ext}$ is derived as a relative loss term by normalizing with $\mathcal{L}_1$ distance between the high-frequency components of I_i^{in} and its blurred counterpart B_i^{in} to alleviate the disturbance caused by blur variance from 3D objects with different depths [11].

The overall optimization objective $\mathcal{L}$ with the proposed EFCR $\mathcal{L}_{CR}$ is given by:

$$\mathcal{L} = \mathcal{L}_1 + \beta\mathcal{L}_{CR} = \mathcal{L}_1 + \beta\frac{1}{n}\sum_{i=1}^{n}\frac{\mathcal{L}_i^+}{\mathcal{L}_i^- + \mathcal{L}_i^{ext}}, \quad (7)$$

where $\mathcal{L}_1$ is the supervised pixel loss, n is the number of samples, and β is the scaling factor.

Knowledge transfer from extra data Defocus blur mainly causes high-frequency degradation [8, 42], implying that the high-frequency part can provide informative cross-domain deblur guidance. EFCR with extra data (denoted by EFCR$_{ex}$) constructs contrastive pairs on high-frequency components. Given an extra training pair $\{I_i^{gt'}, I_i^{in'}, I_i^{out'}\}$ from external dataset and corresponding extended samples $\{B_i^{in'}, B_i^{out'}\}$, EFCR$_{ex}$ with $\{\mathcal{L}_i^{+'}, \mathcal{L}_i^{-'}, \mathcal{L}_i^{ext'}\}$ can be formulated as:

$$\mathcal{L}_i^{+'} = \|f^h(I_i^{out'}) - f^h(I_i^{gt'})\|_1, \quad (8)$$

$$\mathcal{L}_i^{-'} = \|f^h(I_i^{out'}) - f^h(I_i^{in'})\|_1, \quad (9)$$

$$\mathcal{L}_i^{ext'} = \frac{\|f^h(B_i^{out'}) - f^h(B_i^{in'})\|_1}{\|f^h(I_i^{in'}) - f^h(B_i^{in'})\|_1}. \quad (10)$$

The overall optimization objective follows a similar pattern with Eq. (7), where the supervised training on the testing dataset ($\mathcal{L}_1$) proceeds simultaneously with EFCR$_{ex}$ ($\mathcal{L}_{CR}$).

EFCR$_{ex}$ facilitates two important applications. One is to transfer rich deblur signals from a real-world blur dataset to microscopy deblur tasks, in which EFCR$_{ex}$ is composed by $\{\mathcal{L}_i^{+'}, \mathcal{L}_i^{-'}, \mathcal{L}_i^{ext'}\}$. Another is to learn latent deblur direction from an unlabeled microscopy dataset thus enhancing the deblur performance, where the model is trained on a labeled dataset with an unlabeled microscopy dataset as the extra data. EFCR$_{ex}$ here is reformulated as $\{\mathcal{L}_i^{+}, \mathcal{L}_i^{-}, \mathcal{L}_i^{ext'}\}$.

4. Experiments

4.1. Datasets and Implementation

Extensive experiments are carried out on various real-world and microscopy datasets, including five labeled datasets: DPDD [1], LFDOF [59], BBBC006 [44], 3DHistech [17], CaDISBlur, and three unlabeled datasets: CUHK [63], WNLO [17], CataBlur. The LFDOF [59] is adopted as an extra training dataset for knowledge transfer using EFCR, since LFDOF has substantial samples with rich information and good cross correlation between defocused and ground truth pairs [60]. For surgical microscopy, two new surgical microscopy deblur datasets are presented, which are CaDISBlur and CataBlur. CaDISBlur is synthesized using images from a high-quality dataset CaDIS [19] by a novel realistic blur simulation method, in which the instruments and anatomies in the surgery scene are blurred respectively leveraging the object segmentation mask in CaDIS to simulate different focal planes. CataBlur is a new surgical microscope defocus blur dataset, including 1208 defocus images collected from 5 cataract surgeries, for evaluation on real surgical defocus blur. More details about the datasets, training settings, and proposed blur synthesizing method are provided in the supplementary material.

The proposed framework employs the same structure in all tests as follows. The MPT adopts a 4-stage design as shown in Fig. 2, with $[6, 6, 6, 6]$ sub-blocks, $[40, 80, 160, 320]$ feature dimensions, and $[1, 2, 4, 8]$ attention heads. The scale set of each pyramid block is set as $S_1 = [\frac{1}{8}, \frac{1}{8}, \frac{1}{4}, \frac{1}{4}, 1, 1]$, $S_2 = [\frac{1}{4}, \frac{1}{4}, \frac{1}{2}, \frac{1}{2}, 1, 1]$, $S_3 = S_4 = [\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 1, 1]$. The expansion ratio α in FEFN is set to 2.6, and the scaling factor β in EFCR is set to 1e^{-5}. The method is implemented using PyTorch and trained with AdamW optimizer [46] ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay is 1e^{-4}) for 3×10^5 iterations on NVIDIA A800 GPUs. The initial learning rate is set to 1e^{-4} and gradually decreases to 1e^{-6} by cosine annealing [45]. The batch size is set to 8 with training patches in the size of 256×256 augmented with random scaling, and horizontal and vertical flips. Three implementations are included, which are MPT, MPT with EFCR, and MPT with EFCR using LFDOF as extra data (noted as EFCR$_{ex}$). Then the result is reported in three metrics, including Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM) [73], and Learned Perceptual Image Patch Similarity (LPIPS) [87].

4.2. Comparison and Analysis

Evaluation on supervised deblur The evaluation of cell microscopy deblur and surgical microscopy deblur is conducted on three microscopy datasets covering a wide range of state-of-the-art defocus deblur methods and image restoration methods. Real-world deblur evaluation is also conducted on DPDD [1] to further prove the generalizability and universality. The result is shown in Tab. 1 and Tab. 2. The proposed framework demonstrates satisfactory performance on all microscopy datasets and real-world datasets, showing the advantages of the proposed MPT structure and EFCR training strategy. Compared with Restormer [84], which achieves the second-best performance on BBBC006 [44], MPT (76 FLOPs, 19.80 M) outperforms Restormer (141 FLOPs, 26.12 M) by 0.10 dB and 0.07 dB regarding PSNR while saving 46% FLOPs (for a 256×256 input) and 24% parameters, since MPT extracts richer representation than Restormer which only applies channel attention. GRL [36] achieves the second-best performance on 3DHistech [17] and CaDISBlur, but it directly models global spatial attention without leveraging the properties of downscaled

Table 1. Quantitative evaluation on microscopy deblur. The experiments on the sub-set $w1$ (stained by Hoechst to show nuclei structure) and $w2$ (stained by phalloidin to show cell structure) of BBBC006 [44] are conducted separately. Except for the methods using EFCR, the methods with the best and second best performance are noted in red and blue colors.

Method	BBBC006$_{w1}$ [44]			BBBC006$_{w2}$ [44]			3DHistech [17]			CaDISBlur		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM ↑	LPIPS↓
DRBNet [60]	32.83	0.737	0.381	26.66	0.589	0.458	32.83	0.853	0.131	42.54	0.776	0.243
GKMNet [56]	34.41	0.887	0.218	29.32	0.721	0.296	33.42	0.852	0.130	44.27	0.860	0.178
MIMO-UNet [7]	32.73	0.725	0.412	26.90	0.601	0.457	32.40	0.837	0.169	43.36	0.823	0.197
MSSNet [31]	34.01	0.790	0.289	28.68	0.736	0.361	33.09	0.870	0.126	44.09	0.871	0.160
SwinIR [38]	33.90	0.801	0.274	27.61	0.696	0.403	32.57	0.841	0.136	41.83	0.710	0.349
PANet [51]	34.45	0.890	0.230	29.07	0.743	0.290	33.24	0.869	0.129	44.49	0.917	0.134
GRL [36]	34.76	0.907	0.129	29.39	0.786	0.249	33.49	0.878	0.120	44.86	0.960	0.087
Restormer [84]	34.79	0.904	0.135	29.78	0.801	0.241	33.46	0.880	0.125	44.85	0.941	0.101
MPT	34.89	0.912	0.127	29.85	0.813	0.237	33.55	0.881	0.121	44.98	0.962	0.087
MPT+EFCR	34.96	0.917	0.119	29.89	0.820	0.230	33.58	0.887	0.119	45.09	0.969	0.082
MPT+EFCR$_{ex}$	35.16	0.935	0.083	30.11	0.829	0.205	33.63	0.892	0.116	45.25	0.971	0.077

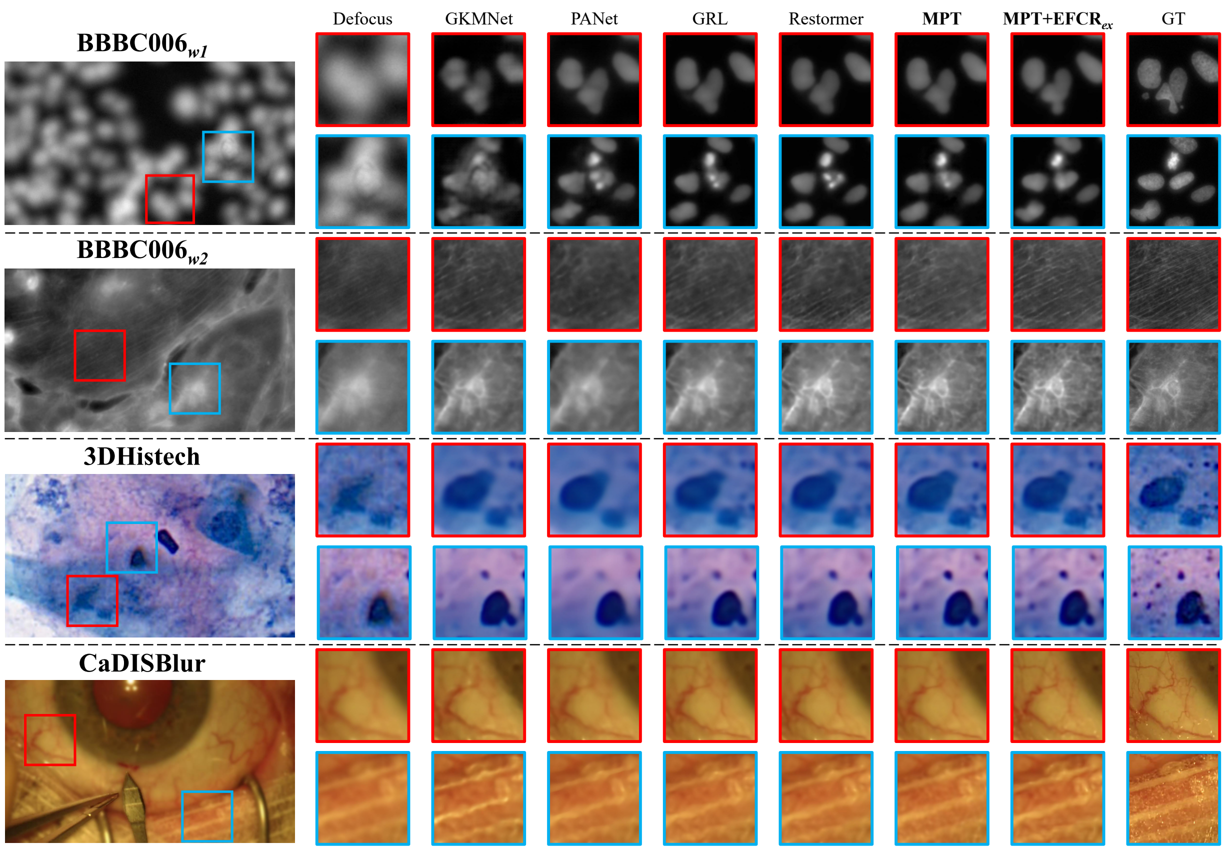

Figure 3. Qualitative evaluation on microscopy deblur. Our method achieves the best restoration of different types of defocus blur.

maps like CSWA, resulting in 1230 FLOPs for a 256×256 input that is 17× larger than ours. Compared to MSSNet [31], MIMO-UNet [7], and PANet [51] that adopt multi-scale or pyramid design, our method with multi-pyramid structure outperforms them in all tests. SwinIR [38] adopts the original local window attention [43], yet is hindered by the limited receptive field and fails to build long-range interaction. The visualizations shown in Fig. 3 prove that our method achieves the best restoration of fine details against strong defocus blur, especially for the miniature cell shape

Table 2. Quantitative real-world deblur evaluation on DPDD [1].

Method	PSNR↑	SSIM↑	LPIPS↓
KPAC [65]	25.24	0.774	0.226
IFAN [34]	25.37	0.789	0.217
DRBNet [60]	25.47	0.787	0.246
GKMNet [56]	25.47	0.789	0.219
Restormer [84]	25.98	0.811	0.178
NRKNet [57]	26.11	0.810	0.210
GRL [36]	26.18	0.822	0.168
MPT	26.21	0.826	0.175
MPT+EFCR	26.23	0.829	0.172
MPT+EFCR$_{ex}$	26.27	0.831	0.161

and complex cell structure, as well as precise features of surgical anatomies. For real-world deblur on DPDD [1], our method achieves the best performance in terms of SSIM and PSNR. It shows that our model is universally applicable to different types of images. Visualization of deblurring on DPDD is provided in Fig. 10 in supplementary materials.

For MPT trained with EFCR, the performance is improved by learning latent deblur information. By further applying EFCR$_{ex}$ to learn cross-domain deblur guidance, the deblur performance is significantly enhanced in all four microscopy datasets. It proves that deblurring benefits from cross-domain knowledge, despite the significant feature discrepancy between real-world extra data and microscope images. Improvements are also observed in SSIM and LPIPS, showing that EFCR$_{ex}$ enhances deblurring from the perspective of the human visual system, which is of great significance for clinical application. Visualizations in Fig. 3 draw a similar conclusion that the model with EFCR$_{ex}$ can restore the fine details more precisely. Real-world deblur can also benefit from EFCR as all three metrics are improved by integrating EFCR or EFCR$_{ex}$. Further discussion in Sec. 4.4 shows the superiority of this proposed training diagram against simply pretraining and fine-tuning.

Evaluation on unsupervised deblur The deblur experiments on unlabeled datasets are conducted to qualitatively evaluate the generalizability of the model, and also to prove the effectiveness of knowledge transfer based on the proposed EFCR$_{ex}$. Two unlabeled microscopy blur datasets are involved, including WNLO [17] and CataBlur. All methods are trained on LFDOF since the rich information in the real-world dataset benefits microscopy deblur (see proof in supplementary materials). Unlabeled datasets are adopted as the extra data to learn cross-domain latent information using EFCR$_{ex}$. The qualitative comparison is shown in Fig. 4. Even without EFCR$_{ex}$, our method still shows the best generalizability with fewer artifacts and successfully restores the detail from strong defocus degradation. With the help

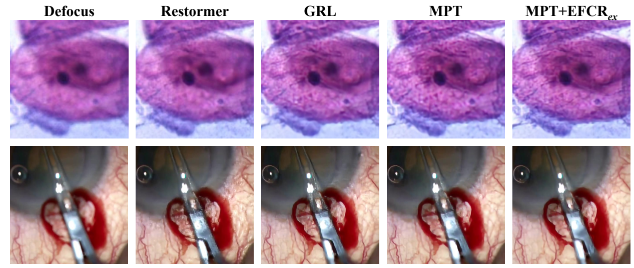

Figure 4. Qualitative evaluation on unsupervised deblur with WNLO (top) and CataBlur (bottom).

of EFCR$_{ex}$, the artifacts are further reduced, resulting in clearer deblurred images. Results on the real-world dataset CUHK [63] are shown in Fig. 11b in supplementary materials, which demonstrates the universality of our method.

4.3. Validation on downstream tasks

To demonstrate the clinical-related improvement, validations on medical downstream tasks are conducted.

Cell detection on BBBC006 Defocus blur can cause failure in cell detection and segmentation [6, 17] that is essential for many biological tasks [52]. The cell segmentation is performed using StarDist [61] on images in BBBC006 before and after deblur. The result is reported in Tab. 6 regarding the average precision (AP) over different intersection-of-union (IoU) thresholds, where higher AP means more cells are successfully detected. Our deblur framework significantly improves the cell detection performance by 19.57% (with extra data) and 18.54% (without extra data) compared with blurry input, surpassing the improvement brought by Restormer (16.05%) and GRL (13.50%). The visualization shown in Fig. 5 also proves that our method achieves better restoration of the cell shape and structure.

Semantic segmentation on CaDISBlur Semantic segmentation plays an important role in surgical scene understanding [19]. The semantic segmentation of the cataract surgical scene is conducted using OCRNet [83] on CaDISBlur based on CaDIS [19]. Results of images with different focal plane positions (blurry instruments or blurry anatomies, noted as *ins* or *ana*) are reported separately in Tab. 7 regarding mean IoU (mIoU) and pixel accuracy (PA). The deblurred images from our method lead to the best performance in most metrics. Visualizations shown in Fig. 5 also demonstrate the superiority of our method.

4.4. Ablation Studies

For ablation studies, the model variants are evaluated regarding PSNR on DPDD [1], BBBC006$_{w1}$ [44] and CaDISBlur datasets, which are denoted by PSNR$_D$, PSNR$_B$ and

Table 3. Ablation studies on attention blocks with four variants, where WA refers to the original version of window attention [38, 43]. Performance degradation occurs in all variants.

Configuration	PSNR_D	PSNR_B	PSNR_C
V_1 (CSWA×2)	26.10	34.76	44.83
V_2 (WA×2)	25.92	33.98	44.02
V_3 (ISCA×2)	26.01	34.60	44.71
V_4 (WA+ISCA)	26.13	34.10	44.78
MPT (CSWA+ISCA)	26.21	34.89	44.98

Table 4. Ablation studies on FEFN with three variants, which are symmetric structures: concatenation (V_1) and adding (V_2) followed by GELU activation, and reversed structure that uses the feature from CSWA for activation (V_3)

Configuration	PSNR_D	PSNR_B	PSNR_C
V_1 (Cat+GELU)	26.18	34.86	44.84
V_2 (Add+GELU)	26.03	34.75	44.87
V_3 (reversed)	25.98	34.72	44.50
MPT (FEFN)	26.21	34.89	44.98

Table 5. Ablation studies on EFCR. ΔPSNR refers to the changes in PSNR compared to the baseline.

Configuration	ΔPSNR_D	ΔPSNR_B	ΔPSNR_C
MPT+V_1	+0.01	+0.04	+0.05
MPT+EFCR	+0.02	+0.07	+0.11
Restormer+EFCR	+0.04	+0.07	+0.09
MPT+pretrain	+0.05	-0.02	+0.01
MPT+V_{ex1}	+0.05	+0.21	+0.18
MPT+EFCR_{ex}	+0.06	+0.27	+0.27
Restormer+EFCR_{ex}	+0.07	+0.19	+0.21

Table 6. Cell detection result on deblurred BBBC006.

IoU	0.5	0.7	0.9	Mean AP
blur	0.7010	0.5623	0.2194	0.4942
Restormer [84]	0.7789	0.6703	0.2714	0.5735
GRL [36]	0.7702	0.6433	0.2691	0.5609
GRL+EFCR	0.7710	0.6440	0.2695	0.5615
GRL+EFCR_{ex}	0.7769	0.6491	0.2733	0.5664
MPT (w/o EFCR)	0.7808	0.6778	0.2956	0.5847
MPT+EFCR	0.7814	0.6791	0.2970	0.5858
MPT+EFCR_{ex}	0.7865	0.6843	0.3019	0.5909
sharp	0.8021	0.7192	0.3518	0.6244

Table 7. Semantic segmentation result on deblurred CaDISBlur.

Method	Blurry instrument		Blurry anatomies	
	mIoU_{ins}	PA_{ins}	mIoU_{ana}	PA_{ana}
blur	0.7577	0.8677	0.7092	0.8194
Restormer [84]	0.7558	0.8803	0.8149	0.8674
GRL [36]	0.7606	0.8849	0.8135	0.8689
GRL+EFCR	0.7611	0.8851	0.8140	0.8691
GRL+EFCR_{ex}	0.7625	0.8857	0.8162	0.8710
MPT (w/o EFCR)	0.7607	0.8836	0.8293	0.8824
MPT+EFCR	0.7610	0.8842	0.8295	0.8830
MPT+EFCR_{ex}	0.7667	0.8859	0.8361	0.8896
sharp	0.7733	0.8886	0.8582	0.9305

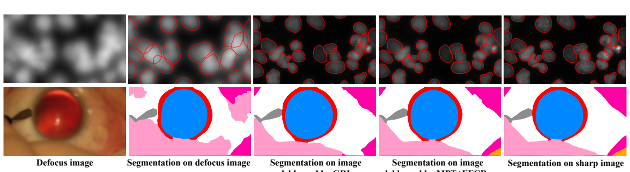

Figure 5. Downstream tasks result on BBBC006 (top) and CaDISBlur (bottom). Our method leads to less false segmentation.

PSNR_C, respectively. More ablation experiments and analyses are provided in supplementary materials.

Configurations of pyramid block Ablation studies on CSWA and ISCA are first carried out. As shown in Tab. 3, a performance drop is observed when changing CSWA to WA (V_4) since WA only models attention within a local window. Although hierarchical network structure in MPT may provide WA with a larger receptive field at low-resolution stages, it still causes inferior performance than CSWA since cross-scale interactions are not built. The situation gets worse if the two attention blocks are all changed to WA (V_2) since the model lost long-range modeling ability in both channel and spatial means. Experiments are then carried out on variants of FEFN, as shown in Tab. 4, which shows the superiority of the proposed asymmetrical activation.

Improvements in EFCR The result is shown in Tab. 5. V_1 and V_{ex1} refer to $\{\mathcal{L}_i^+, \mathcal{L}_i^-\}$ and $\{\mathcal{L}_i^{+'}, \mathcal{L}_i^{-'}\}$. The proposed EFCR and EFCR_{ex} yield significant improvements over baseline, not only on our method but also on Restormer [84], proving the effectiveness of the proposed training diagram. Following a similar approach in [60] to pretrain and fine-tune (denoted as *MPT+pretrain*), the trained model leads to a trivial improvement or even degradation. It further demonstrates the superiority of our EFCR_{ex}.

5. Conclusion

This paper presents a unified framework to address the outstanding problems in microscopy defocus deblur. The MPT outperforms existing multi-scale networks by incorporating spatial-channel context from CSWA and ISCA using FEFN. The proposed EFCR enforces the model to explore latent deblur guidance and further learn cross-domain knowledge from the extra data, yielding significant performance gain in both supervised and unsupervised image deblur. In the future, larger-scale datasets, e.g. ImageNet [10], will be adopted for knowledge transfer using EFCR, along with experiments on weakly supervised or unsupervised learning [24] and domain adaptation [89]. Experiments on MPT variants incorporating varied window mechanisms [78] will be carried out.

Acknowledgement We would like to thank Dr. Danny Siu-Chun Ng from The Department of Ophthalmology and Visual Sciences at The Chinese University of Hong Kong for providing the cataract surgery dataset for research.

References

[1] Abdullah Abuolaim and Michael S. Brown. Defocus Deblurring Using Dual-Pixel Data. In *Computer Vision – ECCV 2020*, pages 111–126, Cham, 2020. Springer International Publishing. 1, 2, 5, 7, 3, 4

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 3

[3] Yunpeng Bai and Chun Yuan. Contrastive Learning in Wavelet Domain for Image Dehazing. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, 2022. 3

[4] Josef F Bille. High resolution imaging in microscopy and ophthalmology: new frontiers in biomedical optics. 2019. 1

[5] David Bouget, Rodrigo Benenson, Mohamed Omran, Laurent Riffaud, Bernt Schiele, and Pierre Jannin. Detecting Surgical Tools by Modelling Local Appearance and Global Shape. *IEEE Transactions on Medical Imaging*, 34(12): 2603–2617, 2015. 2

[6] Xingyu Chen and Fujiao Ju. Automatic Classification of Pollen Grain Microscope Images Using a Multi-Scale Classifier with SRGAN Deblurring. *Applied Sciences*, 12(14): 7126, 2022. 1, 7

[7] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4641–4650, 2021. 3, 6

[8] Yuning Cui, Yi Tao, Wenqi Ren, and Alois Knoll. Dual-Domain Attention for Image Deblurring. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1):479–487, 2023. 3, 4, 5

[9] Andreas P. Cuny, Fabian P. Schlottmann, Jennifer C. Ewald, Serge Pelet, and Kurt M. Schmoller. Live cell microscopy: From image to insight. *Biophysics Reviews*, 3(2):021302, 2022. 1

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 8

[11] Han Dong, Aodong Shen, Youyong Kong, Yu Shen, and Huazhong Shu. No-Reference Defocused Image Quality Assessment Based on Human Visual System. In *2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP)*, pages 1–6, 2019. 5

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[13] Rui Fang, Yang-Fan Yu, En-Jie Li, Ning-Xin Lv, Zhao-Chuan Liu, Hong-Gang Zhou, and Xu-Dong Song. Global, regional, national burden and gender disparity of cataract: findings from the global burden of disease study 2019. *BMC Public Health*, 22(1):2068, 2022. 2

[14] Xin Feng, Yifeng Xu, Guangming Lu, and Wenjie Pei. Hierarchical contrastive learning for pattern-generalizable image corruption detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12076–12085, 2023. 3

[15] Manuel García Calderón, Daniel Torres Lagares, Carmen Calles Vázquez, Jesús Usón Gargallo, and José Luis Gutiérrez Pérez. The application of microscopic surgery in dentistry. *Medicina Oral, Patología Oral y Cirugía Bucal (Internet)*, 12(4):311–316, 2007. 1

[16] Garas Gendy, Nabil Sabor, Jingchao Hou, and Guanghui He. Balanced Spatial Feature Distillation and Pyramid Attention Network for Lightweight Image Super-resolution. *Neurocomputing*, 509:157–166, 2022. 2, 3

[17] Xiebo Geng, Xiuli Liu, Shenghua Cheng, and Shaoqun Zeng. Cervical cytopathology image refocusing via multi-scale attention features and domain normalization. *Medical Image Analysis*, 81:102566, 2022. 1, 2, 3, 5, 6, 7

[18] Negin Ghamsarian, Mario Taschwer, and Klaus Schoeffmann. Deblurring Cataract Surgery Videos Using a Multi-Scale Deconvolutional Neural Network. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 872–876, 2020. 1, 2, 3, 4

[19] Maria Grammatikopoulou, Evangello Flouty, Abdolrahim Kadkhodamohammadi, Gwenolé Quellec, Andre Chow, Jean Nehme, Imanol Luengo, and Danail Stoyanov. Cadis: Cataract dataset for surgical rgb-image segmentation. *Medical Image Analysis*, 71:102053, 2021. 2, 5, 7, 1

[20] Jianyuan Guo, Kai Han, Han Wu, Yehui Tang, Xinghao Chen, Yunhe Wang, and Chang Xu. Cmt: Convolutional neural networks meet vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12175–12185, 2022. 4

[21] Xiaotong Han, Jiaqing Zhang, Zhenzhen Liu, Xuhua Tan, Guangming Jin, Mingguang He, Lixia Luo, and Yizhi Liu. Real-world visual outcomes of cataract surgery based on population-based studies: a systematic review. *British Journal of Ophthalmology*, 107(8):1056–1065, 2023. 2

[22] Chunming He, Chengyu Fang, Yulun Zhang, Kai Li, Longxiang Tang, Chenyu You, Fengyang Xiao, Zhenhua Guo, and Xiu Li. Reti-diff: Illumination degradation image restoration with retinex-based latent diffusion model. 2023. 3

[23] Chunming He, Kai Li, Yachao Zhang, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Camouflaged object detection with feature decomposition and edge reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22046–22055, 2023. 2

[24] Chunming He, Kai Li, Yachao Zhang, Guoxia Xu, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping. *Advances in Neural Information Processing Systems*, 36, 2024. 8

[25] Chunming He, Kai Li, Yachao Zhang, Yulun Zhang, Zhenhua Guo, Xiu Li, Martin Danelljan, and Fisher Yu. Strategic preys make acute predators: Enhancing camouflaged object detectors by generating camouflaged objects. 2024. 3

[26] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 4

[27] Feng-Kai Jan and Chih-Wei Tang. Contrastive Learning Aided Single Image Deblurring. In *2022 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, pages 1–3, 2022. 3

[28] Cheng Jiang, Jun Liao, Pei Dong, Zhaoxuan Ma, De Cai, Guoan Zheng, Yueping Liu, Hong Bu, and Jianhua Yao. Blind deblurring for microscopic pathology images using deep learning networks. *arXiv preprint arXiv:2011.11879*, 2020. 1, 2

[29] Niveditha Kalavakonda, Blake Hannaford, Zeeshan Qazi, and Laligam Sekhar. Autonomous neurosurgical instrument segmentation using end-to-end learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1

[30] Matthew R Keaton, Ram J Zaveri, and Gianfranco Doretto. Celltranspose: Few-shot domain adaptation for cellular instance segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 455–466, 2023. 1

[31] Kiyeon Kim, Seungyong Lee, and Sunghyun Cho. MSSNet: Multi-Scale-Stage Network for Single Image Deblurring. In *Computer Vision – ECCV 2022 Workshops*, pages 524–539, Cham, 2023. Springer Nature Switzerland. 3, 6

[32] Anatasiia Kornilova, Mikhail Salnikov, Olga Novitskaya, Maria Begicheva, Egor Sevriugov, Kirill Shcherbakov, Valeriya Pronina, and Dmitry V. Dylov. Deep Learning Framework For Mobile Microscopy. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 324–328, 2021. 1

[33] Junyong Lee, Sungkil Lee, Sunghyun Cho, and Seungyong Lee. Deep defocus map estimation using domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12222–12230, 2019. 2

[34] Junyong Lee, Hyeongseok Son, Jaesung Rim, Sunghyun Cho, and Seungyong Lee. Iterative Filter Adaptive Network for Single Image Defocus Deblurring. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2034–2042, Nashville, TN, USA, 2021. IEEE. 1, 2, 7

[35] Chen Li, Adele Moatti, Xuying Zhang, H Troy Ghashghaei, and Alon Greenbaum. Deep learning-based autofocus method enhances image quality in light-sheet fluorescence microscopy. *Biomedical Optics Express*, 12(8):5214–5226, 2021. 1

[36] 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. 4, 5, 6, 7, 8

[37] Dong Liang, Ling Li, Mingqiang Wei, Shuo Yang, Liyan Zhang, Wenhan Yang, Yun Du, and Huiyu Zhou. Semantically contrastive learning for low-light image enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1555–1563, 2022. 3

[38] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 6, 8, 3

[39] Wei Lin, Dongping Wang, Yunlong Meng, and Shih-Chi Chen. Multi-focus microscope with hilo algorithm for fast 3-d fluorescent imaging. *PloS one*, 14(9):e0222729, 2019. 1

[40] Gaosheng Liu, Huanjing Yue, and Jingyu Yang. A coarse-to-fine convolutional neural network for light field angular super-resolution. In *CAAI International Conference on Artificial Intelligence*, pages 268–279. Springer, 2022. 3

[41] Jiahao Liu, Xiaoshuai Huang, Liangyi Chen, and Shan Tan. Deep learning–enhanced fluorescence microscopy via degeneration decoupling. *Optics Express*, 28(10):14859–14873, 2020. 1

[42] Keng-Hao Liu, Chia-Hung Yeh, Juh-Wei Chung, and Chuan-Yu Chang. A Motion Deblur Method Based on Multi-Scale High Frequency Residual Image Learning. *IEEE Access*, 8: 66025–66036, 2020. 4, 5

[43] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 3, 4, 6, 8

[44] Vebjorn Ljosa, Katherine L. Sokolnicki, and Anne E. Carpenter. Annotated high-throughput microscopy image sets for validation. *Nature Methods*, 9(7):637–637, 2012. 1, 5, 6, 7, 3

[45] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[46] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[47] 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. 1, 2

[48] Ling Ma and Baowei Fei. Comprehensive review of surgical microscopes: technology development and medical applications. *Journal of Biomedical Optics*, 26(1):010901, 2021. 1, 2

[49] Barry R Masters. History of the optical microscope in cell biology and medicine. *eLS*, 2008. 1

[50] Ioana Mazilu, Shunxin Wang, Sven Dummer, Raymond Veldhuis, Christoph Brune, and Nicola Strisciuglio. Defocus Blur Synthesis and Deblurring via Interpolation and Extrapolation in Latent Space. In *Computer Analysis of Images and Patterns*, pages 201–211, Cham, 2023. Springer Nature Switzerland. 1, 2

[51] Yiqun Mei, Yuchen Fan, Yulun Zhang, Jiahui Yu, Yuqian Zhou, Ding Liu, Yun Fu, Thomas S. Huang, and Humphrey Shi. Pyramid Attention Network for Image Restoration. *International Journal of Computer Vision*, 2023. 2, 3, 6

[52] Erik Meijering. Cell segmentation: 50 years down the road [life sciences]. *IEEE signal processing magazine*, 29(5): 140–145, 2012. 7

[53] Tomer Michaeli and Michal Irani. Blind Deblurring Using Internal Patch Recurrence. In *Computer Vision – ECCV 2014*, pages 783–798, Cham, 2014. Springer International Publishing. 4

[54] Samiksha Pachade, Prasanna Porwal, Dhanshree Thulkar, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, Luca Giancardo, Gwenolé Quellec, and Fabrice Mériaudeau. Retinal Fundus Multi-Disease Image Dataset (RFMiD): A Dataset for Multi-Disease Detection Research. *Data*, 6(2): 14, 2021. 2

[55] Henry Pinkard, Zachary Phillips, Arman Babakhani, Daniel A Fletcher, and Laura Waller. Deep learning for single-shot autofocus microscopy. *Optica*, 6(6):794–797, 2019. 1

[56] Yuhui Quan, Zicong Wu, and Hui Ji. Gaussian Kernel Mixture Network for Single Image Defocus Deblurring. In *Advances in Neural Information Processing Systems*, pages 20812–20824. Curran Associates, Inc., 2021. 1, 2, 3, 6, 7

[57] 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. 1, 2, 7

[58] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. 3

[59] Lingyan Ruan, Bin Chen, Jizhou Li, and Miu-Ling Lam. AIFNet: All-in-Focus Image Restoration Network Using a Light Field-Based Dataset. *IEEE Transactions on Computational Imaging*, 7:675–688, 2021. 1, 2, 5

[60] Lingyan Ruan, Bin Chen, Jizhou Li, and Miuling Lam. Learning to deblur using light field generated and real defocus images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16304–16313, 2022. 1, 2, 3, 5, 6, 7, 8

[61] Uwe Schmidt, Martin Weigert, Coleman Broaddus, and Gene Myers. Cell Detection with Star-Convex Polygons. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 265–273, Cham, 2018. Springer International Publishing. 1, 7

[62] Klaus Schoeffmann, Mario Taschwer, Stephanie Sarny, Bernd Münzer, Manfred Jürgen Primus, and Doris Putzgruber. Cataract-101: video dataset of 101 cataract surgeries. In *Proceedings of the 9th ACM Multimedia Systems Conference*, pages 421–425, New York, NY, USA, 2018. Association for Computing Machinery. 2

[63] Jianping Shi, Li Xu, and Jiaya Jia. Discriminative blur detection features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2965–2972, 2014. 5, 7, 1

[64] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 3

[65] 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. 2, 7

[66] Sameer Trikha, Andrew Michael John Turnbull, RJ Morris, David F Anderson, and Parwez Hossain. The journey to femtosecond laser-assisted cataract surgery: new beginnings or a false dawn? *Eye*, 27(4):461–473, 2013. 2

[67] Anna Trukhova, Marina Pavlova, Olga Sinitsyna, and Igor Yaminsky. Microlens-assisted microscopy for biology and medicine. *Journal of Biophotonics*, 15(9):e202200078, 2022. 1

[68] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 4

[69] Ashish Vaswani, Prajit Ramachandran, Aravind Srinivas, Niki Parmar, Blake Hechtman, and Jonathon Shlens. Scaling local self-attention for parameter efficient visual backbones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12894–12904, 2021. 4

[70] Jiahe Wang and Boran Han. Defocus deblur microscopy via head-to-tail cross-scale fusion. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2081–2086. IEEE, 2022. 1, 2, 3

[71] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021. 3

[72] Yanqi Wang, Zheng Xu, Yifan Yang, Xiaodong Wang, Jiaheng He, Tongqun Ren, and Junshan Liu. Deblurring microscopic image by integrated convolutional neural network. *Precision Engineering*, 82:44–51, 2023. 1, 2

[73] 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. 5

[74] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A General U-Shaped Transformer for Image Restoration. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17662–17672, New Orleans, LA, USA, 2022. IEEE. 3

[75] Gang Wu, Junjun Jiang, and Xianming Liu. A Practical Contrastive Learning Framework for Single-Image Super-Resolution. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–12, 2023. 3

[76] Haiyan Wu, Yanyun Qu, Shaohui Lin, Jian Zhou, Ruizhi Qiao, Zhizhong Zhang, Yuan Xie, and Lizhuang Ma. Contrastive learning for compact single image dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10551–10560, 2021. 3

[77] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22–31, 2021. 4

[78] 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. 8

[79] Jing Xu, Xiaolin Tian, Xin Meng, Yan Kong, Shumei Gao, Haoyang Cui, Fei Liu, Liang Xue, Cheng Liu, and Shouyu Wang. Wavefront-sensing-based autofocusing in microscopy. *Journal of Biomedical Optics*, 22(8):086012–086012, 2017. 1

[80] Ruikang Xu, Zeyu Xiao, Jie Huang, Yueyi Zhang, and Zhiwei Xiong. Edpn: Enhanced deep pyramid network for blurry image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 414–423, 2021. 2, 3

[81] Mahmut Gazi Yaşargil. *Microsurgery: applied to neurosurgery*. Elsevier, 2013. 1

[82] Seunghwan Yoo, Pablo Ruiz, Xiang Huang, Kuan He, Nicola J Ferrier, Mark Hereld, Alan Selewa, Matthew Daddysman, Norbert Scherer, Oliver Cossairt, et al. 3d image reconstruction from multi-focus microscope: axial super-resolution and multiple-frame processing. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1453–1457. IEEE, 2018. 1

[83] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-Contextual Representations for Semantic Segmentation. In *Computer Vision – ECCV 2020*, pages 173–190, Cham, 2020. Springer International Publishing. 7

[84] 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. 4, 5, 6, 7, 8

[85] Chi Zhang, Hao Jiang, Weihuang Liu, Junyi Li, Shiming Tang, Mario Juhas, and Yang Zhang. Correction of out-of-focus microscopic images by deep learning. *Computational and Structural Biotechnology Journal*, 20:1957–1966, 2022. 1, 2

[86] Dengsheng Zhang and Dengsheng Zhang. Wavelet transform. *Fundamentals of image data mining: Analysis, Features, Classification and Retrieval*, pages 35–44, 2019. 2, 3, 4

[87] 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. 5

[88] Yunlong Zhang, Yuxuan Sun, Honglin Li, Sunyi Zheng, Chenglu Zhu, and Lin Yang. Benchmarking the Robustness of Deep Neural Networks to Common Corruptions in Digital Pathology. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 242–252, Cham, 2022. Springer Nature Switzerland. 1, 2

[89] Yuelin Zhang, Sihao Xiang, Zehuan Wang, Xiaoyan Peng, Yutong Tian, Shukai Duan, and Jia Yan. Tdacnn: Target-domain-free domain adaptation convolutional neural network for drift compensation in gas sensors. *Sensors and Actuators B: Chemical*, 361:131739, 2022. 8

[90] Yanni Zhang, Qiang Li, Miao Qi, Di Liu, Jun Kong, and Jianzhong Wang. Multi-scale frequency separation network for image deblurring. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 3

[91] Yulun Zhang, Donglai Wei, Richard Schalek, Yuelong Wu, Stephen Turney, Jeff Lichtman, Hanspeter Pfister, and Yun Fu. High-Throughput Microscopy Image Deblurring with Graph Reasoning Attention Network. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2023. 1

[92] Bingxin Zhao, Weihong Li, and Weiguo Gong. Real-aware motion deblurring using multi-attention CycleGAN with contrastive guidance. *Digital Signal Processing*, 135: 103953, 2023. 3

[93] Huangxuan Zhao, Ziwen Ke, Ningbo Chen, Songjian Wang, Ke Li, Lidai Wang, Xiaojing Gong, Wei Zheng, Liang Song, Zhicheng Liu, Dong Liang, and Chengbo Liu. A new deep learning method for image deblurring in optical microscopic systems. *Journal of Biophotonics*, 13(3):e201960147, 2020. 1, 2

[94] Suiyi Zhao, Zhao Zhang, Richang Hong, Mingliang Xu, Yi Yang, and Meng Wang. FCL-GAN: A Lightweight and Real-Time Baseline for Unsupervised Blind Image Deblurring. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6220–6229, New York, NY, USA, 2022. Association for Computing Machinery. 3

[95] Maria Zontak, Inbar Mosseri, and Michal Irani. Separating signal from noise using patch recurrence across scales. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1195–1202, 2013. 4