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

Supplementary Material

6. Dataset Description

This section provides supplementary information about the datasets involved in the evaluation. The details are shown in Tab. 8.

For real-world scenarios, three datasets are included. For DPDD [1], the sharp-blur training pairs are collected sequentially by a DSLR camera with different aperture sizes. A labeled defocus deblur dataset LFDOF [59] collected by a light field camera is adopted in this paper as extra data for knowledge transfer since the training pairs captured by the light field camera have strict pixel-wise consistency [60]. LFDOF contains many more images than DPDD, which also benefits the microscopy deblur by transferring rich cross-domain information. For unsupervised deblur, an unlabeled blur dataset CUHK [63] is adopted, which is collected from the internet.

Three datasets are involved in cell microscopy deblur. BBBC006 comes from the Broad Bioimage Benchmark Collection [44], which contains images in two sub-sets stained by Hoechst and phalloidin captured by fluorescence microscope. It contains images with different focal planes (denoted as different z-stacks). Following dataset description [44], images collected on the optimal focal length (z-stack = 16) are set as the ground truth, and images above the optimal focal plane are used for training. To avoid redundancy, images with z-stack = [2, 6, 10] are set as blurry input for training. Since the images in BBBC006 only contain a single grayscale channel, for EFCR$_{ex}$ in BBBC006, the images in LFDOF are converted into grayscale with one channel. 3DHistech and WNLO [17] are two cell imaging datasets for cytopathology scanned by digital scanners. The labeled dataset 3DHistech is scanned using different focal planes, where the focal plane with the most cells in focus is set as the ground truth. WNLO is an unlabeled dataset with defocus images only.

Regarding the surgical microscopy deblur, two new datasets are presented, which are the labeled synthesized dataset CaDISBlur and the unlabeled cataract surgery defocus blur dataset CataBlur. CaDISBlur is synthesized based on CaDIS [19], which is a dataset for surgical scene semantic segmentation. Leveraging segmentation masks, the instruments and anatomies are blurred separately to simulate different focal planes. The original images in CaDIS are of high quality thus they can be treated as the sharp ground truth. CataBlur is an unlabeled real defocus blur dataset containing 1208 images acquired during 5 different cataract surgeries, from which the severity of defocus blur in microscope surgery can be observed. The privacy information is removed. The experiment's conduction and the dataset's collection are granted with ethical approval. The images in CataBlur are sampled from surgery videos with lower frames per second (fps) to remove redundancy.

Table 8. Dataset description.

Scenario	Dataset	#Image	Resolution	Label
Real world	DPDD [1]	500	1680×1120	Labeled
	LFDOF [59]	12,826	1008×688	Labeled
	CUHK [63]	704	$\sim$ 640×480	Unlabeled
Cell microscopy	BBBC006 [44]	6144	696×520	Labeled
	3DHistech [17]	94,973	256×256	Labeled
	WNLO [17]	108,065	256×256	Unlabeled
Surgical microscopy	CaDISBlur	9340	960×540	Labeled
	CataBlur	1208	1280×720	Unlabeled

7. Supplementary Experiments

Data deficiency As addressed in Sec. 1, the data deficiency in microscopy datasets can pose harm to generalizability. To further demonstrate the drawbacks brought by data deficiency, experiments are conducted on unlabeled microscopy datasets WNLO [17] and CataBlur regarding two settings. For the first setting S_1, the MPT is **trained on intra-domain microscopy images and tested on unlabeled microscopy datasets**, i.e., trained on 3DHistech [17] then tested on WNLO [17], and trained on CaDISBlur then tested on CataBlur. For S_2, the MPT is **trained on cross-domain real-world images (LFDOF [59]) and tested on unlabeled microscopy datasets**.

The deblur results are shown in Fig. 6, from which it can be observed that the model trained with cross-domain real-world dataset (S_2) leads to fewer artifacts and more fine details in its deblurred results than the model trained with intra-domain microscopy dataset (S_1). This phenomenon proves the existence of data deficiency in microscopy dataset, i.e., the model trained with microscopy dataset suffers from poor generalizability caused by the insufficient features contained in microscopy dataset. From visualization of S_2, the necessity of learning cross-domain rich deblur guidance is also proved.

Ablation studies on CSWA and FEFN Additional ablation studies regarding the structure of the proposed CSWA

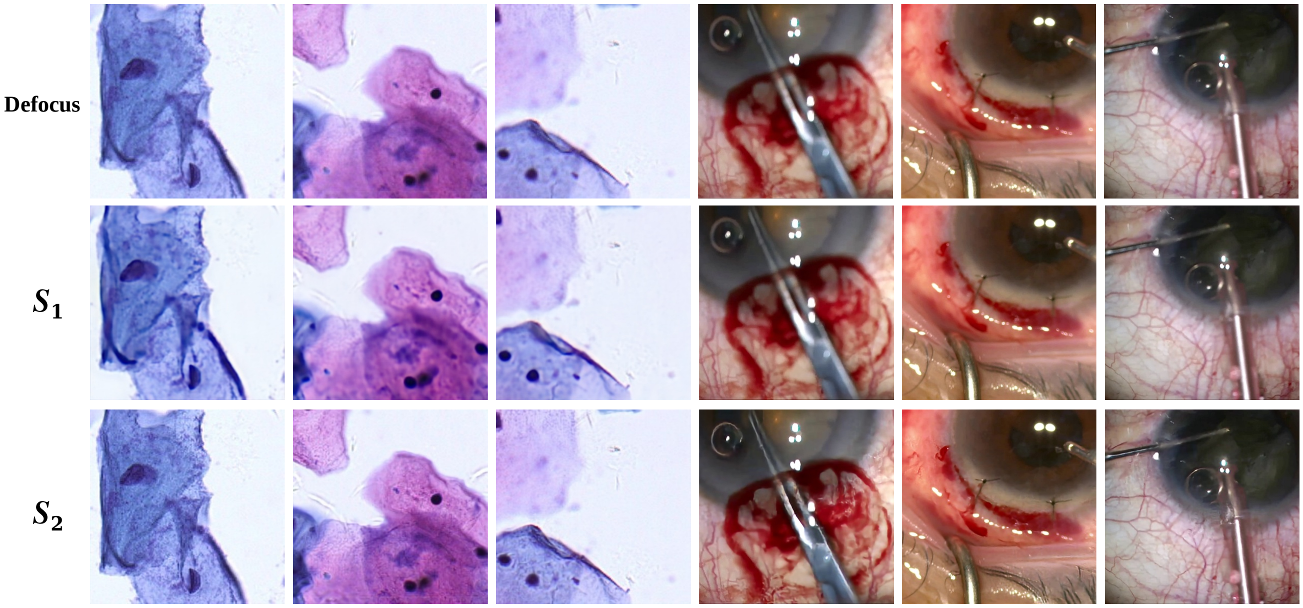


Figure 6. Illustration of models trained with data deficient microscopy dataset (S_1) and data sufficient real-world dataset (S_2), including unlabeled dataset WNLO [17] (left) and CataBlur (right). Even if the model is trained with intra-domain microscopy data (S_1), the deblur restoration turns out to be trivial for CataBlur, or even brings strong artifacts for WNLO. As for S_2, the deblur result tends to have fewer artifacts, and more fine details are restored. It proves the existence of data deficiency problem and the necessity of learning cross-domain knowledge.

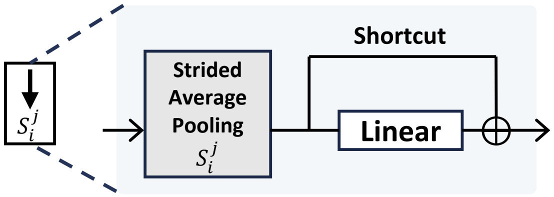


Figure 7. Downsampling module adopted in CSWA.

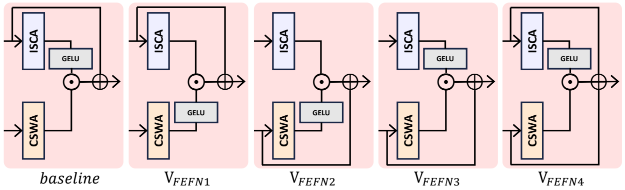


Figure 8. Baseline and variant design of ablation studies for FEFN. V$_{FEFN1}$ is the reversed structure (V$_3$) shown in Tab. 4

Table 9. Ablation studies on CSWA and FEFN. ΔPSNR refers to the change in PSNR compared with the MPT baseline.

Configuration	ΔPSNR$_D$	ΔPSNR$_B$	ΔPSNR$_C$
V_{ds1} (w/o shortcut)	-0.14	-0.02	-0.15
V_{ds2} (w/o linear)	-0.05	-0.04	-0.07
V_{ds3} (max pooling)	-0.37	-0.49	-0.63
V_{ds4} (convolution)	-0.03	-0.09	-0.06
V_{ds5} (interpolate)	-0.01	+0.02	-0.04
$V_{w/o\ NPC}$	-0.09	-0.07	-0.12
$V_{w/o\ sw}$	-0.06	-0.13	-0.09
V_{FEFN1}	-0.23	-0.17	-0.48
V_{FEFN2}	-0.09	-0.12	-0.15
V_{FEFN3}	-0.60	-0.74	-0.91
V_{FEFN4}	-0.24	-0.31	-0.20

and FEFN are conducted. The results are shown in Tab. 9. For the downsampling module adopted in CSWA (Fig. 7), a series of variants are evaluated, including variant without shortcut connection (V_{ds1}), the variant without linear projection and shortcut connection (V_{ds2}), and variants changing the downsampling operation from strided average pooling to strided maximum pooling (V_{ds3}), strided convolution (V_{ds4}), and bicubic interpolation (V_{ds5}). The results in Tab. 9 show that the current baseline outperforms the variants in most of the situations. Although V_{ds5}, which adopts bicubic interpolation, achieves almost the same performance as the baseline or even trivial improvement, it leads to complex computation that hinders parallel training and inference. Ablation studies on NPConv and shifting window mechanism in CSWA are also carried out, which are denoted by $V_{w/o\ NPC}$ and $V_{w/o\ sw}$, respectively. The results demonstrate the superiority of our design.

Going further from the experiments in Sec. 4.4, more ablation studies regarding the asymmetric activation mechanism and shortcut connection in FEFN are conducted, including four variants as shown in Fig. 8. Based on the result reported in Tab. 9, it can be concluded that the baseline structure adopted in this paper achieves the best performance among all the variants.

Ablation studies on pyramid scales Following the description in Sec. 4.1 to keep the sub-block number, feature dimensions, and attention heads unchanged, ablation studies on pyramid scales (S_i) are carried out regarding variants with similar FLOPs and parameters:

1) V$_1$: $S_1 = S_2 = S_3 = S_4 = [1, 1, 1, 1, 1, 1]$. In this variant, CSWA actually downgrades to the original WA (the variant with WA+ISCA shown in Tab. 3), and the image pyramid is not constructed.

2) V$_2$: $S_1 = [\frac{1}{16}, \frac{1}{16}, \frac{1}{8}, \frac{1}{8}, 1, 1]$, $S_2 = [\frac{1}{8}, \frac{1}{8}, \frac{1}{4}, \frac{1}{4}, 1, 1]$, $S_3 = S_4 = [\frac{1}{4}, \frac{1}{4}, \frac{1}{2}, \frac{1}{2}, 1, 1]$. This variant explores the pyramid structure with smaller scales than the baseline.

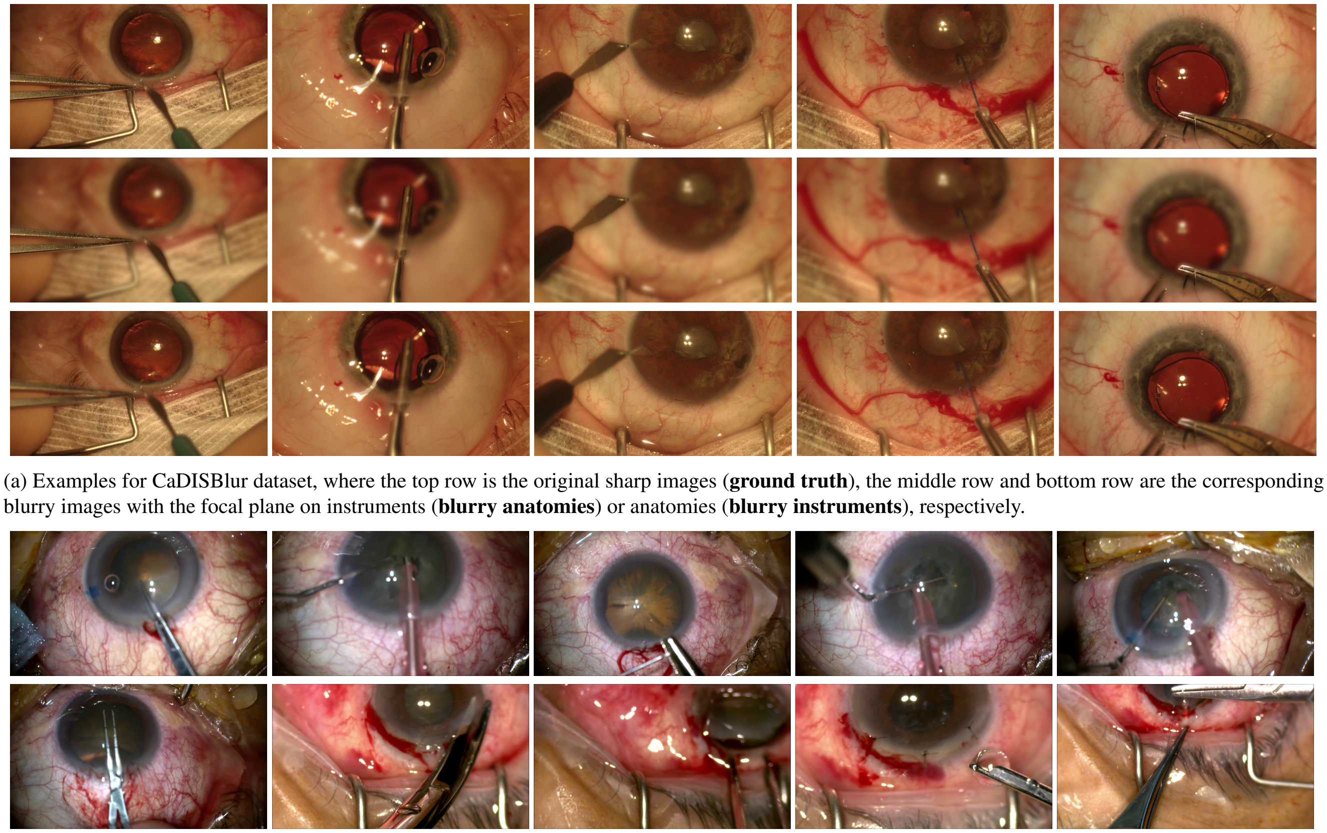
3) V$_3$: $S_1 = [\frac{1}{4}, \frac{1}{4}, \frac{1}{2}, \frac{1}{2}, 1, 1]$, $S_2 = [\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 1, 1]$, $S_3 = S_4 = [\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 1, 1]$. This variant adopts larger-scale pyramids than the baseline.

As shown by the results in Tab. 10, all the variants lead to performance drops. For V$_1$ with no pyramid structure, significant performance degradation is observed on BBBC006, which is the dataset with one of the longest attention spans. A similar phenomenon is observed in SwinIR [38] that does not feature a multi-scale pyramid. It achieves inferior performance on BBBC006 as shown in Tab. 1. These together prove the effectiveness of our multi-pyramid design, especially for the microscopy deblur tasks. The performance degradation in V$_2$ and V$_3$ shows the superiority of the pyramid scales in the baseline model.

Table 10. Ablation studies on pyramid scales.

Pyramid	ΔPSNR$_D$	ΔPSNR$_B$	ΔPSNR$_C$
V$_1$	-0.08	-0.79	-0.20
V$_2$	-0.05	-0.02	-0.07
V$_3$	-0.01	-0.04	-0.03

Additional visualization Demonstrations of supervised real-world deblur are shown in Fig. 10 for labeled datasets DPDD [1]. Illustrations of unsupervised deblur on microscopy dataset and real-world dataset are provided in Fig. 11a and Fig. 11b, respectively. The visualization proves that our method achieves the best performance on microscopy datasets and real-world datasets regarding various patterns in both supervised and unsupervised scenarios. More visualizations of the results of cell detection on BBBC006 [44] and surgical scene semantic segmentation on CaDISBlur are provided in Fig. 12, from which it can be concluded that the downstream tasks results on deblurred images from our method achieves more satisfactory outcomes.



(a) Examples for CaDISBlur dataset, where the top row is the original sharp images (**ground truth**), the middle row and bottom row are the corresponding blurry images with the focal plane on instruments (**blurry anatomies**) or anatomies (**blurry instruments**), respectively.

(b) Examples for CataBlur dataset, which is an unlabeled dataset containing only defocus images.

Figure 9. Examples of samples in CaDISBlur and CataBlur.

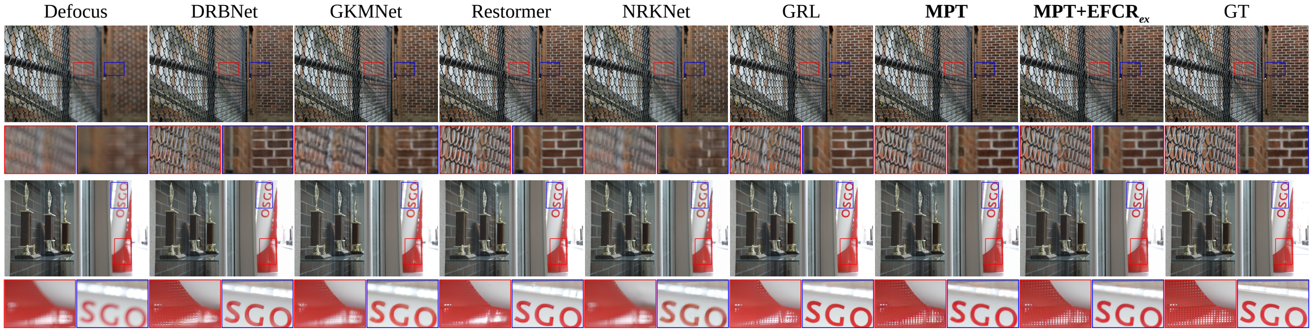
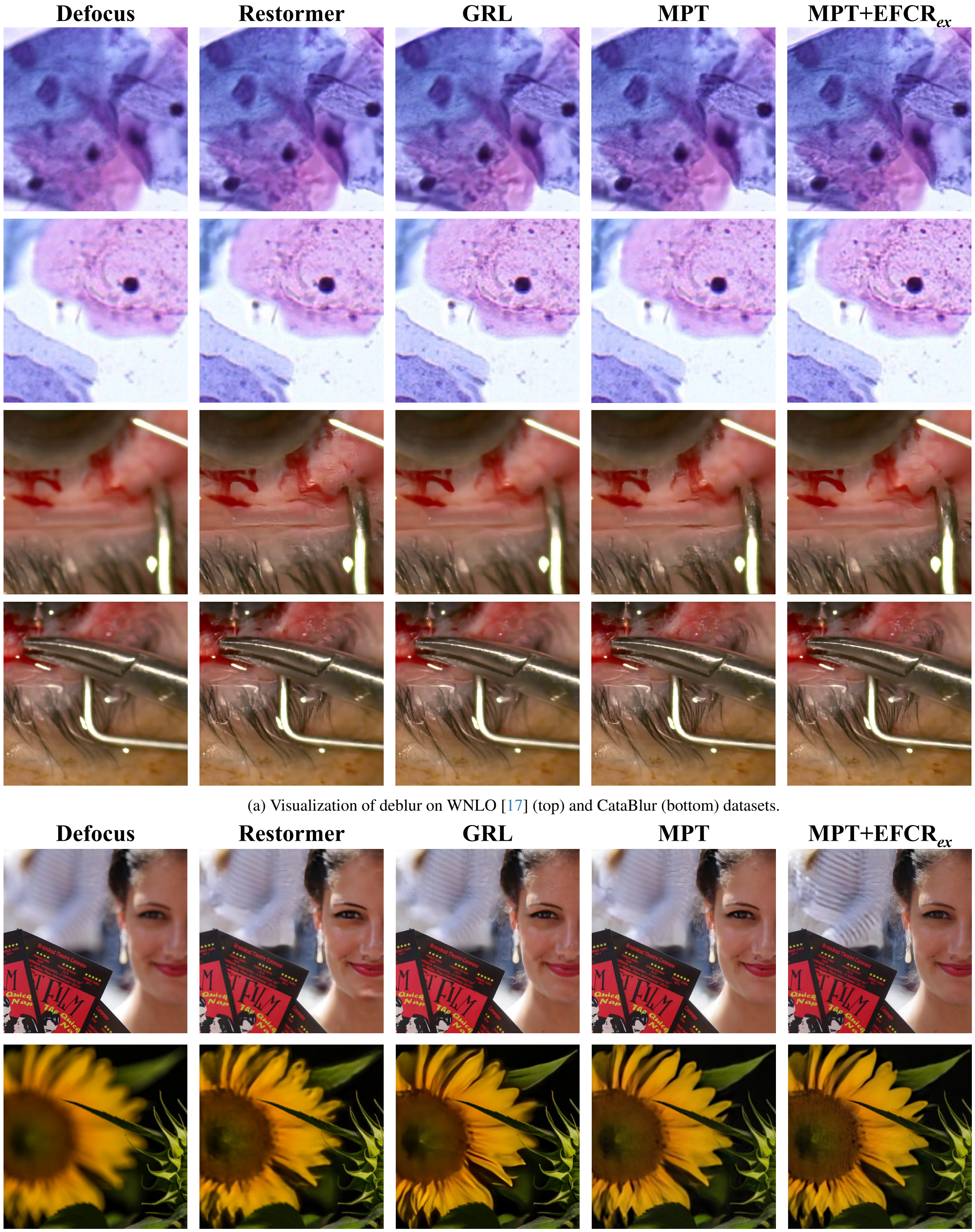


Figure 10. Visualization of deblur on DPDD [1] dataset. As the comparison shows, our MPT achieves the best deblur performance in real-world scenes, restoring most of the small-scale fine details and large-scale patterns. With the help of EFCR$_{ex}$, the performance is further enhanced. It shows the superiority and generalizability of our framework.



(a) Visualization of deblur on WNLO [17] (top) and CataBlur (bottom) datasets.

(b) Visualization of deblur on CUHK [63] dataset.

Figure 11. Visualization of unsupervised deblur on CataBlur, WNLO [17], and CUHK [63]. The proposed framework achieves the best unsupervised deblur performance on both microscopy datasets and real-world datasets, showing high generalizability. By further applying EFCR$_{ex}$, the deblur performance is significantly enhanced, proving the effectiveness of the proposed EFCR for knowledge transfer.

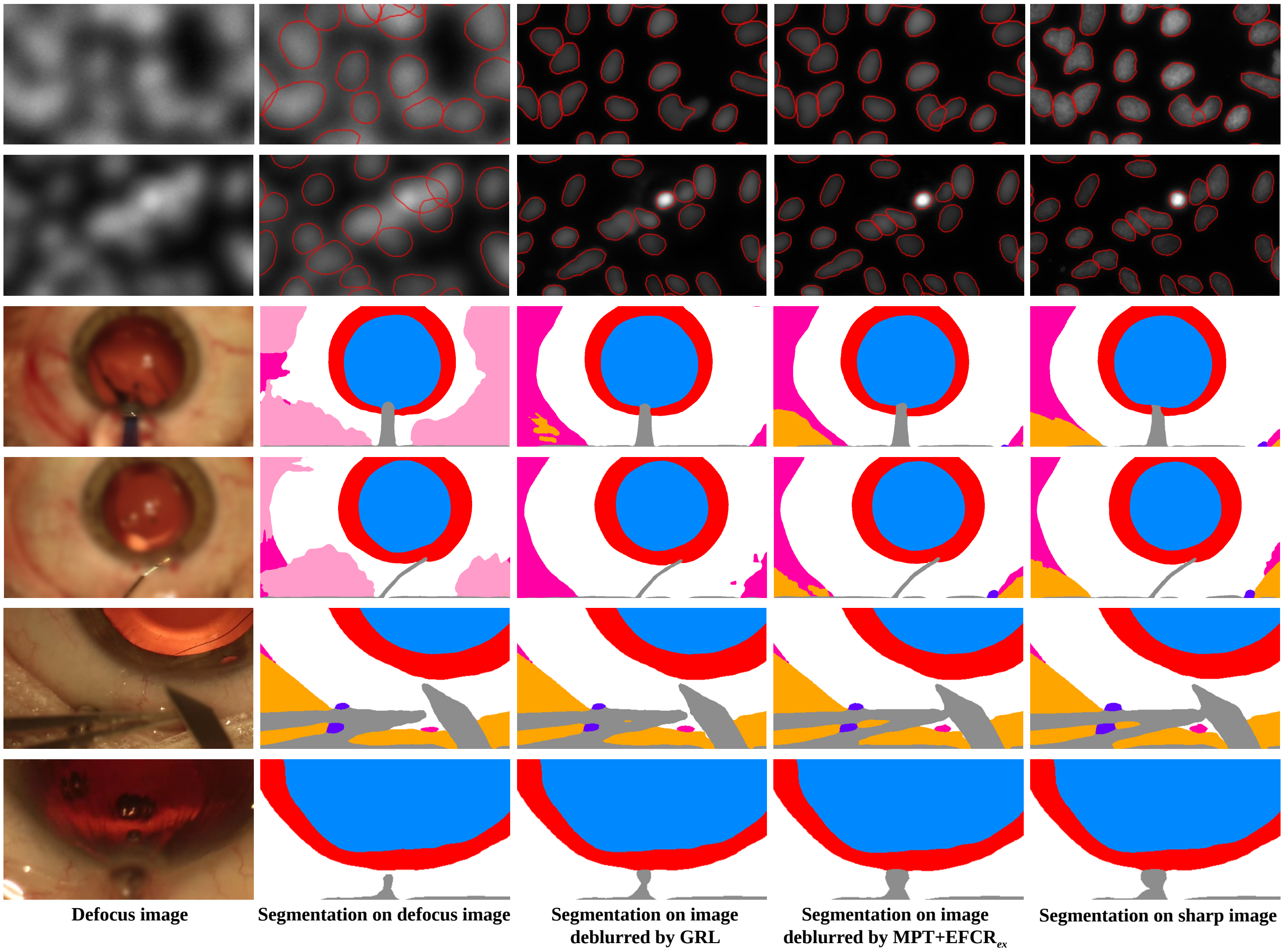

Defocus image
Segmentation on defocus image
Segmentation on image deblurred by GRL
Segmentation on image deblurred by MPT+EFCR$_{ex}$
Segmentation on sharp image

Figure 12. Results of downstream tasks on BBBC006 [44] (top) and CaDISBlur (bottom). For BBBC006, our method can restore precise cell shape with sharper outline, achieving more accurate segmentation and detection. For CaDISBlur, the deblurred images by our framework contain more differentiable features, leading to more accurate semantic segmentation in both anatomies and instruments. (Colormap: 🟦 Pupil, 🟥 Iris, ■ surgeon's hand, ⬜ Cornea, ■ Skin, 🟧 Surgical tape, 🟪 Eye retractors, ■ Instruments)