Supplementary Material for "LDP: Language-driven Dual-Pixel Image Defocus Deblurring Network"

In this supplementary material, we provide additional LDP implementation details (Sec. 1) and more ablation studies (Sec. 2), and additional results (Sec. 3).

1. Implementation Details

We build our network structure, Ours (Small) and Ours (Large), by adopting the backbone network structure from [2] with *small* and *large* configurations, respectively, and insert one BPA block at the 1st layer of the network backbone. The initial channel width for the small and large of our LDP is 32, and is doubled/halved after downsampling/upsampling. There are 4 layers for the encoder and decoder, each encoder/decoder layer is followed by a downsampling/upsampling operation, and the depth of each layer is presented below,

- Ours (Small): encoder depth=[2, 2, 2, 2], decoder depth=[2, 2 ,2, 2].
- Ours (Large): encoder depth=[4, 4, 8, 8], decoder depth=[8, 8, 4, 4].

The BPA and model block architecture are visualized in Fig. 1 and Fig. 2. Ours (Small)/Ours (Large) has 7.6M/19.6M parameters and 270G/686G flops, where flops are measured based on the DPD-blur dataset DP pair resolution, *i.e.*, 1120×1680. The above measurement does not take into account CLIP for blur map estimation, which consumes 88M parameters and 220G flops.

2. Ablation Study

CLIP Design Choices. Our blur map estimation strategy uses CLIP. We study the impact of CLIP variants on blur map estimation. We compare three versions of CLIP which adopt ViT-B/16, ViT-B/32, and ResNet50 as encoder, respectively, and report the results in Tab. 2. We obtain the best results with ViT-B/32 as the backbone, and also have the following observations: i) ResNet50 leads to the worst performance because of the absence of long-range interaction that is required to model the symmetry between the left and right view. ii) ViT-B/16 achieves the second-best performance. Compared with the ViT-B/32, the patch size is reduced from 32×32 to 16×16, and the number of patches is increased quadratically. The ViT-B/16 thus has a high complexity when modeling pair-wise patch symmetry, decreasing the blur map estimation performance.

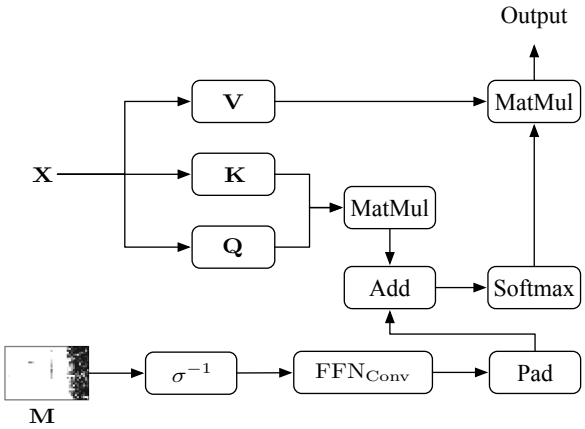

Figure 1. *Our BPA block architecture.*

DP-aware Format. We experimentally verify the effectiveness of our designed DP-aware format that is used for blur map estimation. To further demonstrate its advantages, we additionally study three baseline prompts by directly concatenating the left and right views in Tab. 1. Our method (prompting the horizontal symmetry between the left and flipped right views) achieves 0.15/0.03 higher in PSNR(dB)/SSIM compared to the second-best additional baseline setting. Moreover, we show the blur maps generated from different baseline settings in Fig. 3.

Loss Generalization Ability. We train state-of-the-art defocus deblurring methods, Restormer and DeepRFT, by additionally using our $\mathcal{L}_{\text{bwl}}$ and $\mathcal{L}_{\text{bal}}$ in Tab. 3. We denote the variations as Restormer$^+$ and DeepRFT$^+$. Compared to the original Restormer and DeepRFT, Restormer$^+$ and DeepRFT$^+$ achieve 0.25 dB and 0.41 dB improvements in PSNR, respectively. The improvements show the effectiveness of the loss functions.

Analysis of Attention map Our BPA block assumes that the attention map is an adaptive deblurring kernel. Here we provide analyses of the attention map in this section. For simplicity, we assume a single attention block-based deblurring network. We then solve the following optimization

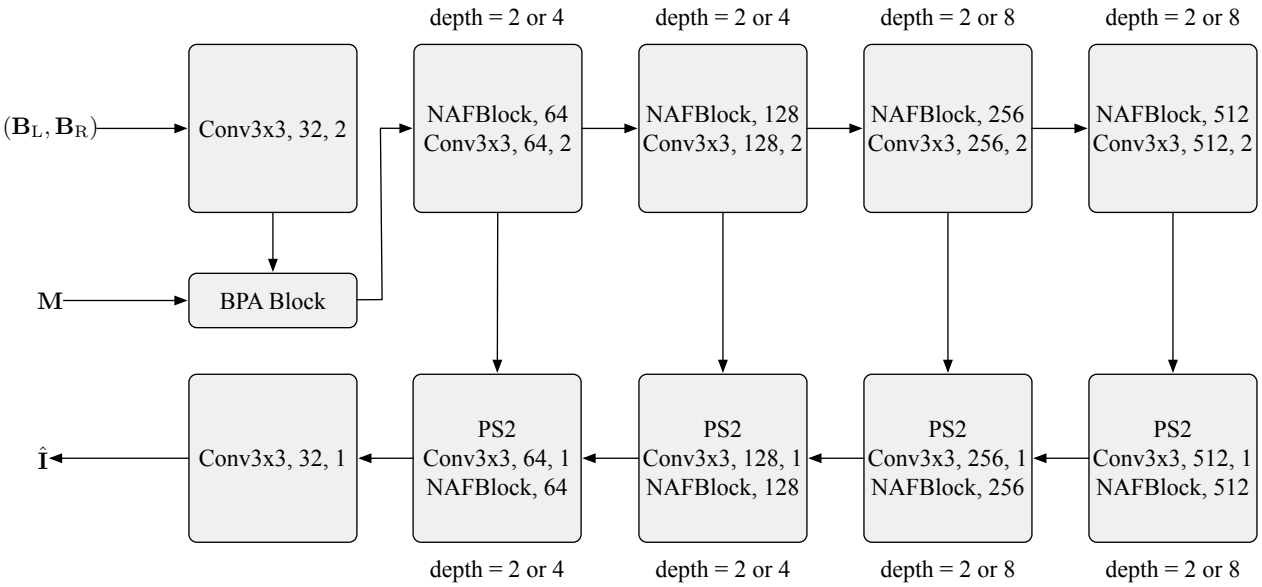

Figure 2. *Our LDP architecture. We use the deblurring backbone from [2] composed of NAFBlock, convolution layers, and pixel shuffle layers. For NAFBlock, we attach the output channel after it. We use 'Conv' as a convolution layer followed by kernel size, output channel, and stride. In the encoder branch (top row), each NAFBlock layer is followed by a convolution layer with stride 2 for downsampling. In the decoder branch (bottom row), a pixel shuffle layer with a factor of 2 (*i.e.*, 'PS2') and a convolution layer is used before each NAFBlock layer for upsampling. For Ours (Small) and Ours (Large), we have depth=2 and depth=4/8, respectively.*

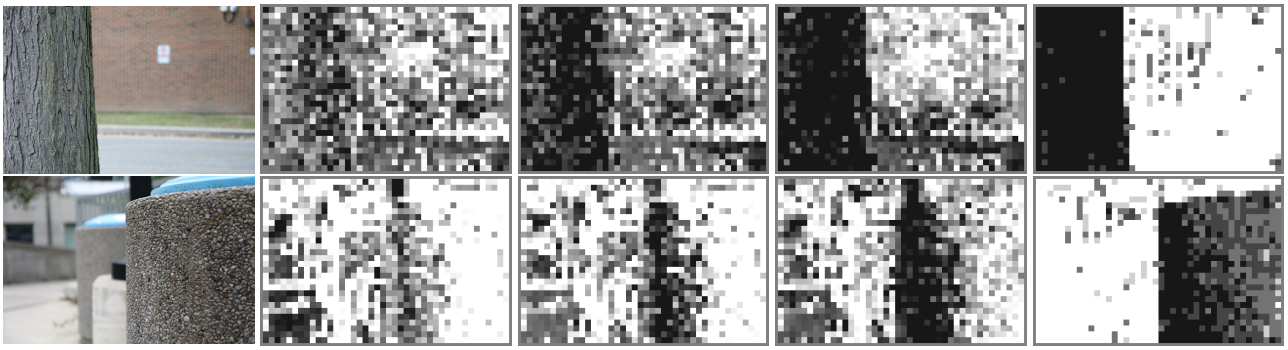

Figure 3. *Blur maps generated from different baseline settings. (a) is the blurred image. From (b) to (e), we present the estimated blur map by using prompts in Tab. 1.*

Table 1. *Comparison of using different DP-aware formats.*

Prompt	PNSR$_\uparrow$	SSIM$_\uparrow$	MAE$_\downarrow$	MSE_rel$_\downarrow$
`[The left and right of the image are different.]`	26.61	0.825	0.036	0.047
`[The image is inconsistent from left to right.]`	26.69	0.827	0.033	0.046
`[The left half and the right half are different.]`	26.76	0.828	0.034	0.046
Ours	**26.91**	**0.831**	**0.032**	**0.045**

problem,

$$\theta^\star = \arg\min_\theta \mathcal{L}_\text{total}\Big(\sum_{i,j,l,p} \frac{\exp(\mathbf{Q}(i,j)\mathbf{K}(l,p))}{\exp\sum_{l',p'}(\mathbf{Q}(i,j)\mathbf{K}(l',p'))}\mathbf{V}(l,p), \mathbf{I}\Big) ,$$
$$= \arg\min_\theta \mathcal{L}_\text{total}\Big(\sum_{i,j,l,p} \mathbf{A}(i,j,l,p)\mathbf{V}(l,p), \mathbf{I}\Big) ,$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are projected from the DP pair $(\mathbf{B}_\mathrm{L}, \mathbf{B}_\mathrm{R})$ by weight matrix $\mathbf{W}_\mathbf{Q}$, $\mathbf{W}_\mathbf{K}$, and $\mathbf{W}_\mathbf{V}$. (i,j), (l,p), and (l',j') are indices of the feature domain, and $\theta = \{\mathbf{W}_\mathbf{Q}, \mathbf{W}_\mathbf{K}, \mathbf{W}_\mathbf{V}\}$ is the set of all weight parameters. By the decomposition, we can

Table 2. *Comparison of using different CLIP versions.*

Setting	PNSR$_\uparrow$	SSIM$_\uparrow$	MAE$_\downarrow$	MSE_rel$_\downarrow$
ResNet50	26.81	0.828	0.032	0.046
ViT-B/16	26.88	0.830	0.032	0.045
ViT-B/32	**26.91**	**0.831**	**0.032**	**0.045**

Table 3. *Generalization ability of our proposed loss terms $\mathcal{L}_{bwl}$ and $\mathcal{L}_{bal}$. We use $^+$ to denote the model trained with additionally using our $\mathcal{L}_{bwl}$ and $\mathcal{L}_{bal}$.*

Setting	PSNR$_\uparrow$	SSIM$_\uparrow$	MAE$_\downarrow$	MSE_rel$_\downarrow$
Restormer	26.66	0.833	0.035	0.046
Restormer$^+$	26.91	0.834	0.033	0.045
DeepRFT	25.71	0.801	0.037	0.051
DeepRFT$^+$	26.12	0.806	0.036	0.049

(a) (b) (c) (d)

Figure 4. *Visualization of feature maps w/o our BPA Block. (a) is the blur image, and (b) is the feature map from self-attention (Eq. (8) in main paper). (c) is blur map generated by our method. (d) is the feature map from Our BPA block (Eq. (9) in main paper).*

assume that the attention map is an exponential kernel normalized for performing deblurring (see Fig. 4).

3. Additional Results

We show restorations of our method and state-of-the-art method on DPD-blur dataset [1], DPD-disp dataset [4], and our collected LDP-real data in Fig. 5, Fig. 6, and Fig. 7. The collected LDP-real data is captured by us using a Canon EOS 5D Mark IV under low lighting conditions, without ground truth sharp images.

References

[1] Abdullah Abuolaim and Michael S Brown. Defocus deblurring using dual-pixel data. In *European Conference on Computer Vision*, pages 111–126. Springer, 2020. 3, 4

[2] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 17–33. Springer, 2022. 1, 2

[3] 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. 4, 5, 6

[4] Abhijith Punnappurath, Abdullah Abuolaim, Mahmoud Afifi, and Michael S Brown. Modeling defocus-disparity in dual-pixel sensors. In *2020 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12. IEEE, 2020. 3, 5

[5] Yan Yang, Liyuan Pan, Liu Liu, and Miaomiao Liu. K3dn: Disparity-aware kernel estimation for dual-pixel defocus deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13263–13272, June 2023. 4, 5, 6

[6] 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

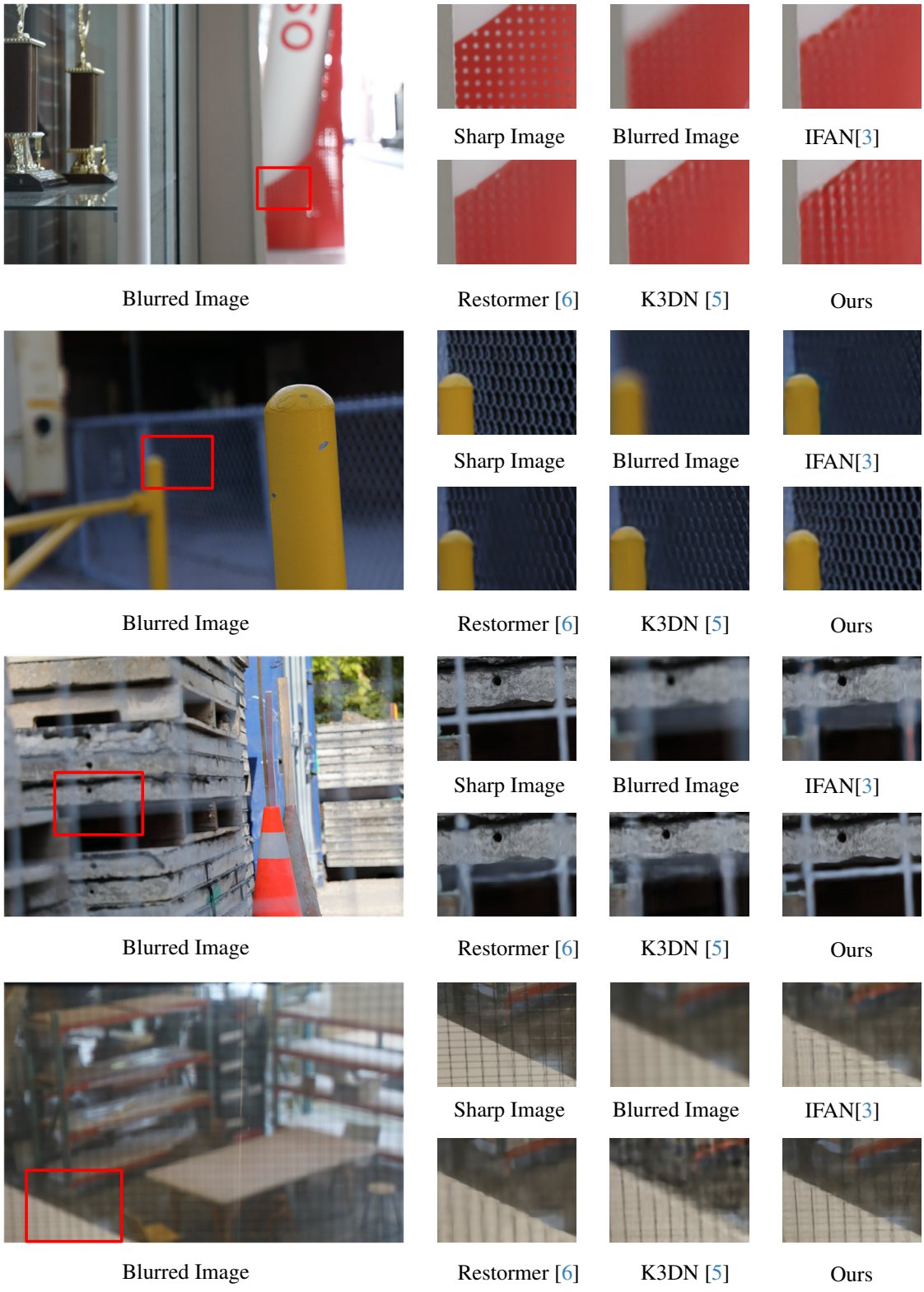

Figure 5. *Qualitative comparison on the DPD-blur dataset [1]. We present the ground truth large sharp image in the first column, and the regions residing in the red bounding box are cropped into a small ground truth sharp image in the second column. The corresponding regions of the blurred image ($\mathbf{B}_{\mathrm{L}}$) are in the third column.*

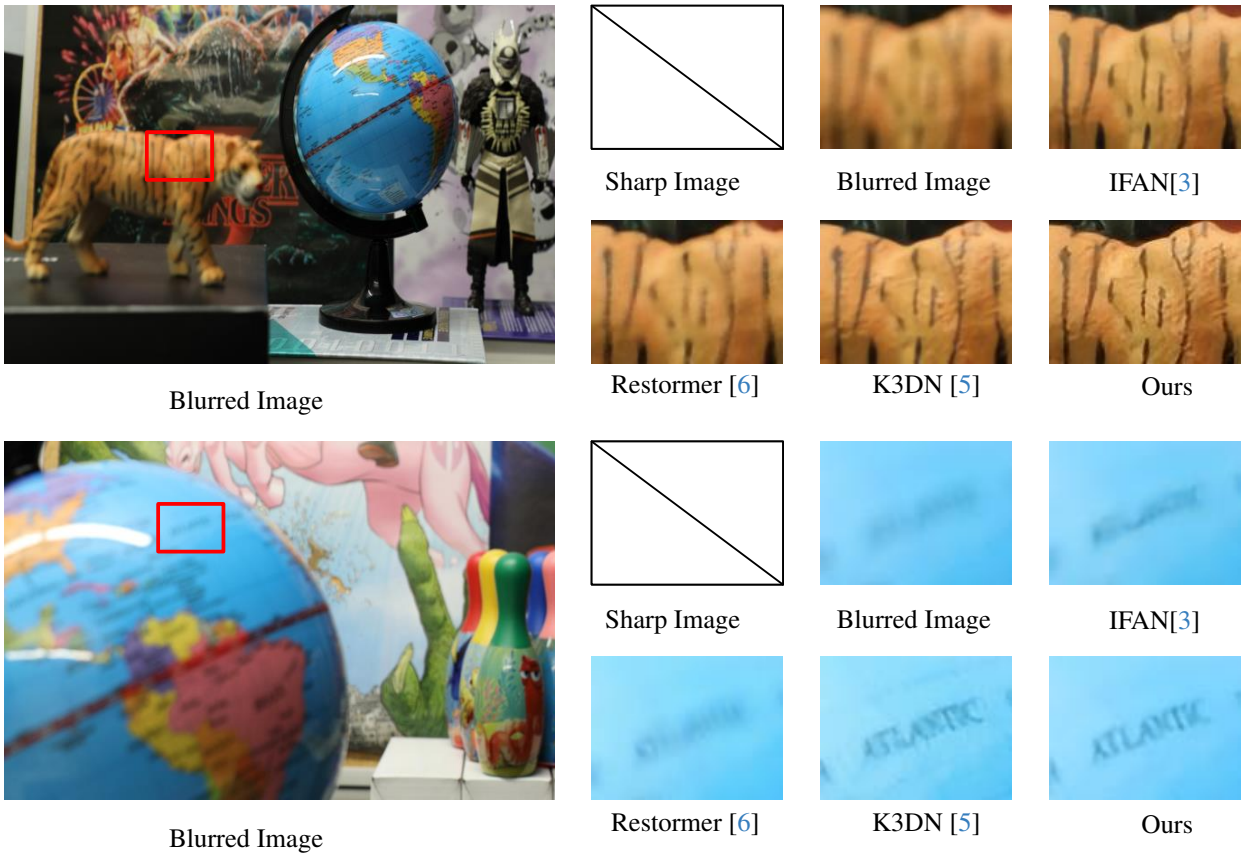

Figure 6. *Qualitative comparison on the collected DPD-disp dataset [4]. We present the left DP image ($\mathbf{B}_{\mathrm{L}}$) in the first column, and the regions residing in the red bounding box and its sharp image are cropped into small images in the third and second columns.*

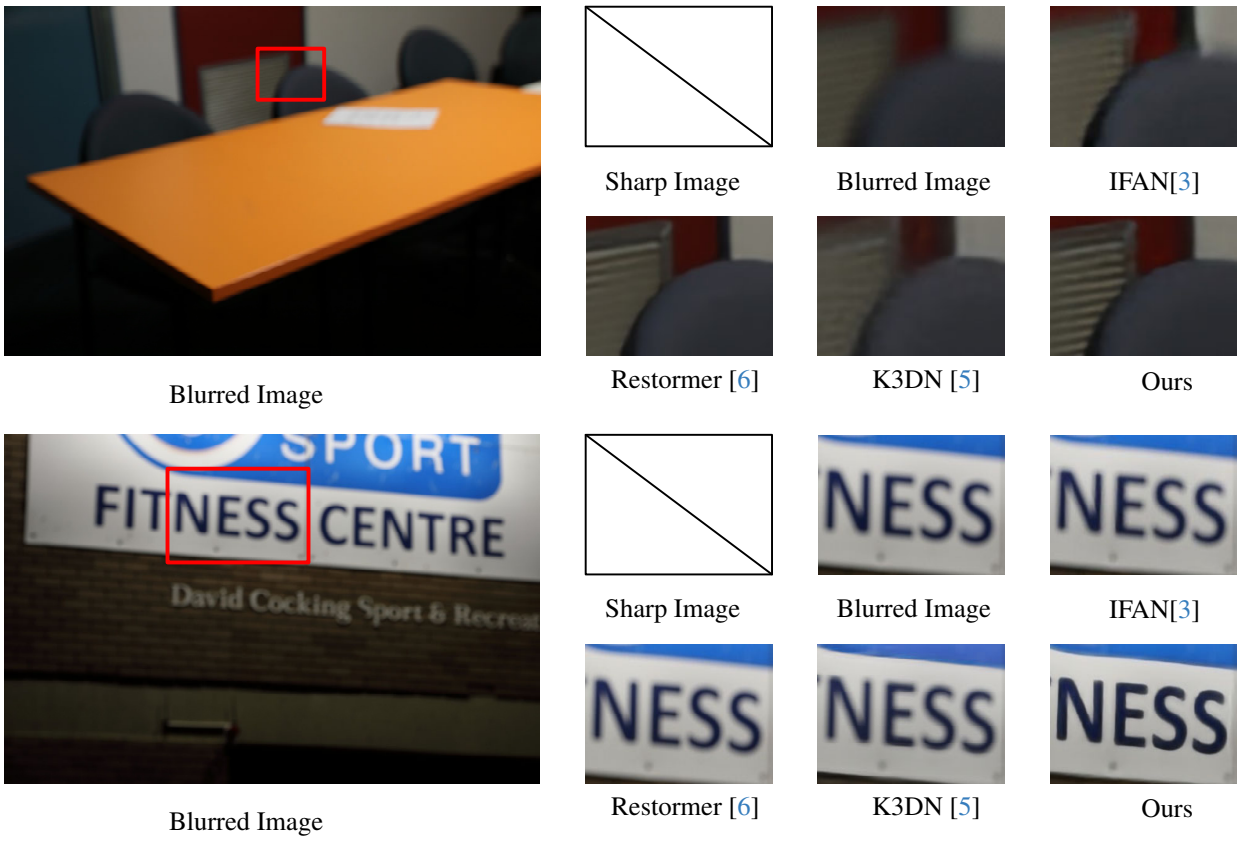

Figure 7. *Qualitative comparison on the collected LDP-real data. We present the left DP image ($\mathbf{B}_{\mathrm{L}}$) in the first column, and the regions residing in the red bounding box and its sharp image are cropped into small images in the third and second columns.*