Deep Discriminative Spatial and Temporal Network for Efficient Video Deblurring Supplemental Material

Jinshan Pan[1*], Boming Xu[1*], Jiangxin Dong[1†], Jianjun Ge[2], and Jinhui Tang[1†]
[1]Nanjing University of Science and Technology [2]China Electronics Technology Group Corporation

Overview

In this document, we first present the network details in Section 1. Then, we provide additional analysis on the effect of the discriminative temporal feature fusion module and the wavelet-based feature propagation method in Section 2 and Section 3. To examine the effect of the number of frames on video deblurring, we further analyze it in Section 4. In Section 5, we analyze the limitations of the proposed approach. Finally, we show more visual comparisons on both synthetic and real-world images in Section 6.

1. Network Details

As stated in Section 3 of the main manuscript, our method contains a channel-wise gated dynamic network, a discriminative temporal feature fusion module, and a wavelet-based feature propagation for video deblurring. We also show the network details of the proposed channel-wise gated dynamic network and discriminative temporal feature fusion module in Figures 2 and 3 of the main manuscript. In this document, we show the overall network details of the proposed method in Figure 1.

2. Further Analysis on the Discriminative Temporal Feature Fusion Module

As the contents of each frame are different, we split the features and estimate the individual weight to respectively utilize useful information from each feature in the DTFF module. One may wonder whether the method without using the splitting operation or only using a concatenation followed by several ResBlocks works well or not. To answer this question, we compare the method without splitting operations in the DTFF module and the one using a concatenation followed by several ResBlocks using the same settings as stated in Section 5 of the manuscript. We use one additional 1×1 convolutional layer to ensure that the fused features by these two baselines have the same channel numbers as the proposed method. Table 1 shows that using the splitting operation generates better deblurred results.

Table 1. Quantitative evaluations of the splitting operation in the DTFF on the GoPro dataset. “w/ Concatenation by 25 ResBlocks” denotes that we replace the DTFF module with the concatenation operation followed by 25 ResBlocks in the proposed method.

Methods	w/o split	w/ Concatenation by 25 ResBlocks	Ours
PSNRs	33.27	33.08	33.33
SSIMs	0.9612	0.9593	0.9611
Network parameters	7.05M	7.54M	7.45M
FLOPs	42G	65G	48G

3. Further Analysis on the Wavelet-based Feature Propagation

The wavelet-based feature propagation (WaveletFP) is mainly used to avoid the error accumulation and reduce computational cost during the feature propagation. As significant blur increases the difficulty of structural detail estimation, if one

[*] Co-first authorship
[†] Corresponding author

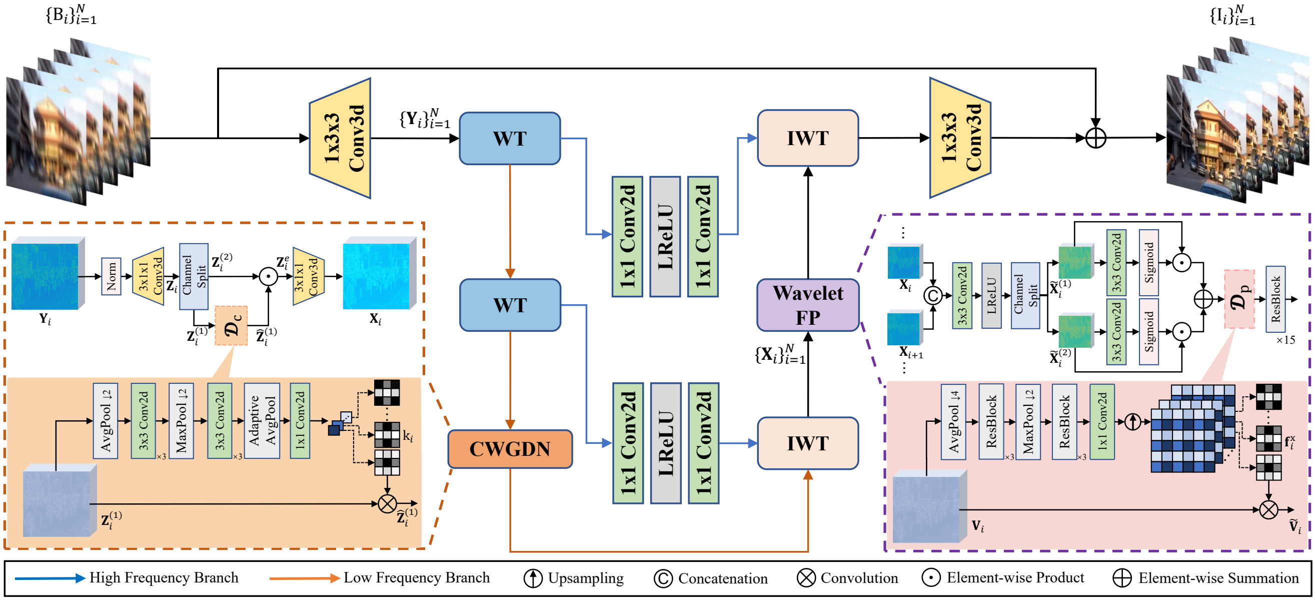


Figure 1. An overview of the proposed network for video deblurring. Given a blurred video $\mathbf{B} = \{\mathrm{B}_i\}_{i=1}^N$ with N frames, we first use a feature extraction module to extract features $\{\mathbf{Y}_i\}_{i=1}^N$ from $\mathbf{B}$, where $\mathbf{Y}_i \in \mathbb{R}^{H\times W\times C}$, $H \times W$ denotes the spatial dimension, and C is the number of channels. Then, we apply the channel-wise gated dynamic network (CWGDN) to the low-frequency part of $\{\mathbf{Y}_i\}_{i=1}^N$ by wavelet transformer and use the inverse wavelet transformer to obtain the reconstructed feature $\{\mathbf{X}_i\}_{i=1}^N$ as the input of the wavelet-based feature propagation (WaveletFP) that takes the discriminative temporal feature fusion (DTFF) module as the basic component. With the generated features. Finally, the latent frames are reconstructed based on the features by the wavelet-based feature propagation. "WT" and "IWT" denote the wavelet transform and inverse wavelet transform.

frame contains significant blur, the errors of the inaccurate estimated structural details will be accumulated during the feature propagation, which affects video deblurring. We have analyzed the effect of the WaveletFP method on video deblurring in Section 5 of the manuscript. In this document, we further analyze the effect of the wavelet transform used in the WaveletFP. To this end, we put the WaveletFP after the second IWT block in Figure 1. That is, the feature propagation is conducted on the original image resolutions without using the wavelet transform. We use the "w/o wavelet in FP" to denote this baseline and train it using the same settings as the proposed method for comparisons. Table 2 shows that our method generates better results with lower FLOPs values. In contrast, the baseline, "w/o wavelet in FP", requires higher computational cost than the proposed method.

The above evaluations demonstrate that the WaveletFP can alleviate the influence of the inaccurate structural details from non-local frames during the feature propagation process, thus facilitating video deblurring.

Table 2. Quantitative evaluations of the WaveletFP on the GoPro dataset. "w/ Bilinear in FP" denotes that we use the Bilinear downsampling and upsampling operations instead of the wavelet transform and the inverse wavelet transform in the proposed method.

Methods	w/o wavelet in FP	w/ Bilinear in FP	Ours
PSNRs	33.26	32.87	33.33
SSIMs	0.9608	0.9587	0.9611
Network parameters	7.45M	7.40M	7.45M
FLOPs	180G	47G	48G

4. Effect of the Number of Frames on Video Deblurring

We quantitatively evaluate the effect of the number of frames on video deblurring using the GoPro dataset [1]. Figure 2 shows that using more frames is able to improve the performance. However, the improvement is not significant when the number of frames is larger than 20.

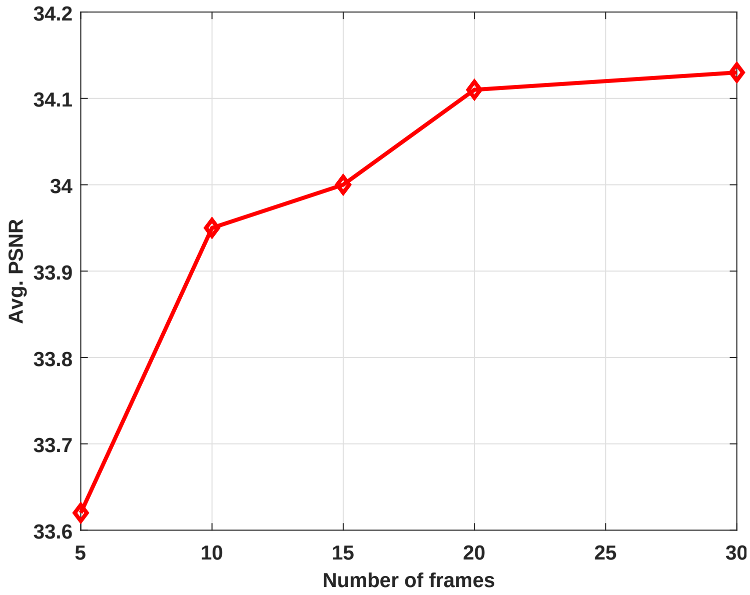
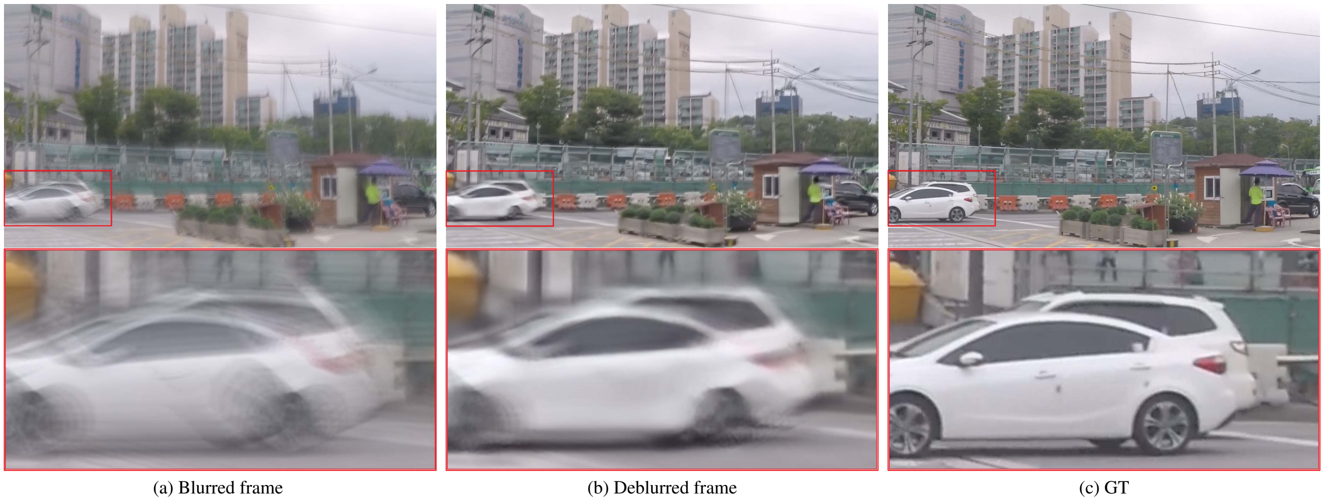


Figure 2. Effect of the number of frames on video deblurring. The results are obtained from the GoPro dataset [1].



(a) Blurred frame

(b) Deblurred frame

(c) GT

Figure 3. Limitations of the proposed method. Due to the fast motion of the car and the camera during the exposure time, the captured frame contains significant blur effects, e.g., the wheels of the car. The proposed method does not effectively remove the blur caused by abrupt motions. For example, the wheels are not recovered well as shown in (b).

5. Limitations

As demonstrated in Section 5 of the manuscript, although the proposed method achieves favorable performance on several video deblurring datasets, it cannot effectively handle the scenes with abrupt changes as it is difficult to find useful temporal information from both adjacent and long-range frames. Figure 3 shows an example, where the object and cameras have abrupt motions. Our method does not remove the blur effect well.

6. More Experimental Results

In this section, we provide more visual comparisons of the proposed method and state-of-the-art ones on both synthetic and real-world videos. Figures 4-12 show the comparisons, where our method generates better deblurred frames.

References

[1] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, pages 257–265, 2017. 2, 3, 5, 6, 7

[2] Jinshan Pan, Haoran Bai, and Jinhui Tang. Cascaded deep video deblurring using temporal sharpness prior. In *CVPR*, pages 3040–3048, 2020. 5, 6, 7, 8, 9, 10, 11, 12, 13

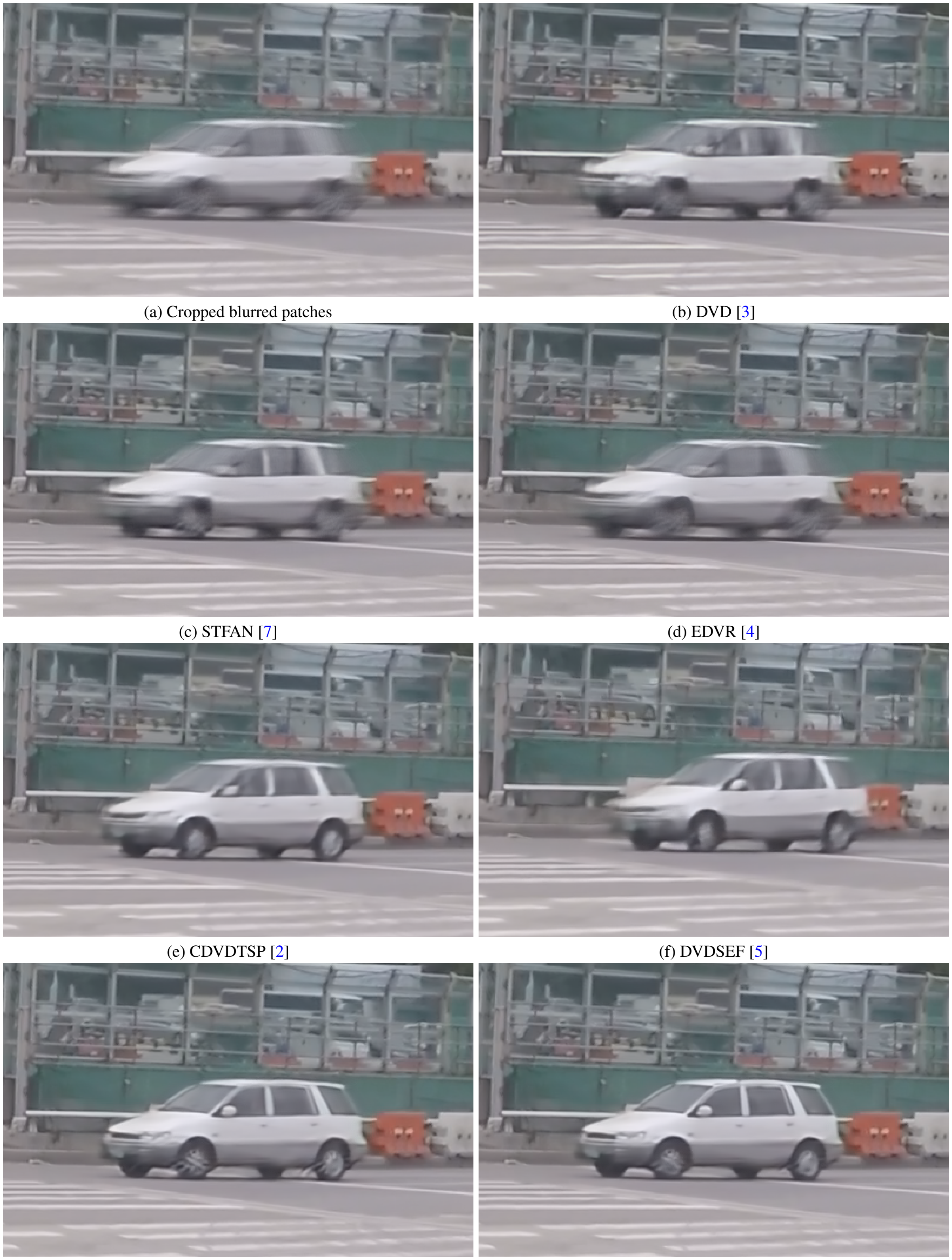
[3] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *CVPR*, pages 237–246, 2017. 5, 6, 7, 8, 9, 10, 11, 12

[4] Xintao Wang, Kelvin C.K. Chan, Ke Yu, Chao Dong, and Chen Change Loy. EDVR: Video restoration with enhanced deformable convolutional networks. In *CVPR Workshops*, pages 1954–1963, 2019. 5, 6, 7

[5] Xinguang Xiang, Hao Wei, and Jinshan Pan. Deep video deblurring using sharpness features from exemplars. *IEEE TIP*, 29:8976–8987, 2020. 5, 6, 7

[6] Zhihang Zhong, Ye Gao, Yinqiang Zheng, and Bo Zheng. Efficient spatio-temporal recurrent neural network for video deblurring. In *ECCV*, pages 191–207, 2020. 8, 9, 10, 11, 12, 13

[7] Shangchen Zhou, Jiawei Zhang, Jinshan Pan, Haozhe Xie, Wangmeng Zuo, and Jimmy Ren. Spatio-temporal filter adaptive network for video deblurring. In *ICCV*, pages 2482–2491, 2019. 5, 6, 7, 8, 9, 10, 11, 12



(a) Cropped blurred patches

(b) DVD [3]

(c) STFAN [7]

(d) EDVR [4]

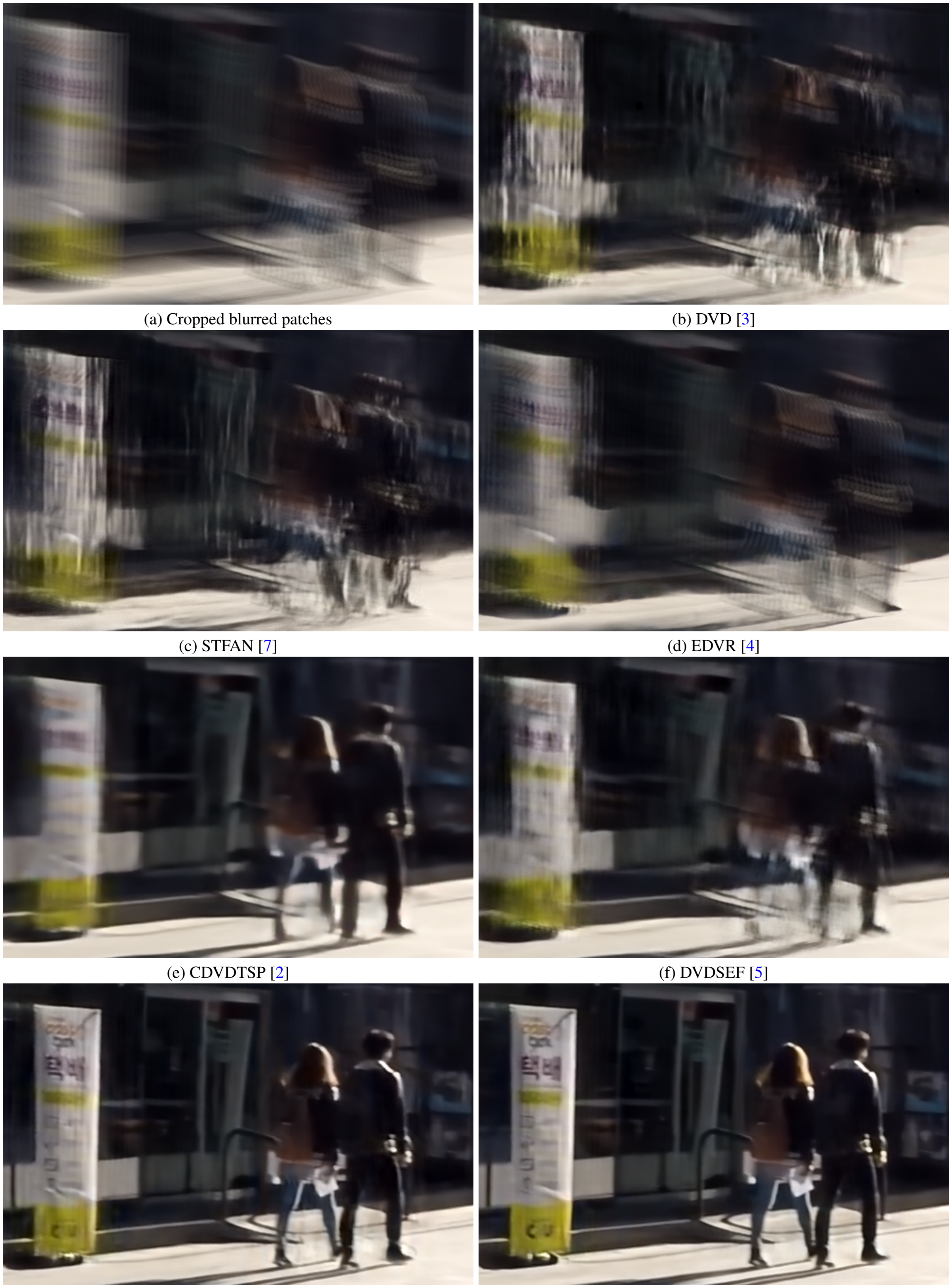
(e) CDVDTSP [2]

(f) DVDSEF [5]

(g) Ours

(h) Ours-L

Figure 4. Deblurred results on the GoPro test dataset [1]. The results shown in (b)-(d) still contain significant blur effects. The CDVDTSP method [2] and the DVDSEF method [5] do not recover the car well. In contrast, our method generates a better-deblurred frame, where the car is restored well.



(a) Cropped blurred patches

(b) DVD [3]

(c) STFAN [7]

(d) EDVR [4]

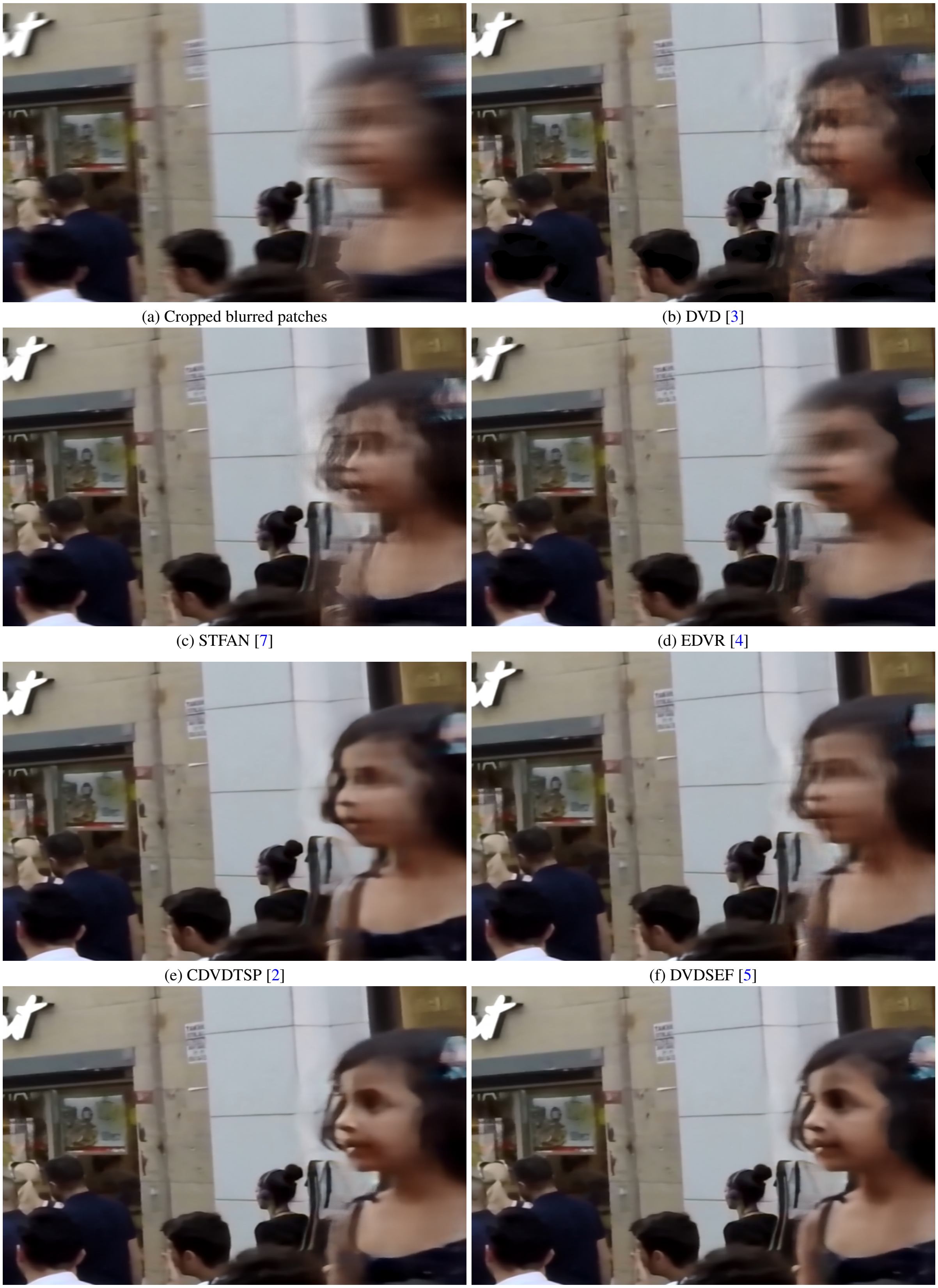
(e) CDVDTSP [2]

(f) DVDSEF [5]

(g) Ours

(h) Ours-L

Figure 5. Deblurred results on the GoPro test dataset [1]. The results shown in (b)-(f) still contain significant blur effects. In contrast, our method generates a better-deblurred frame.



(a) Cropped blurred patches

(b) DVD [3]

(c) STFAN [7]

(d) EDVR [4]

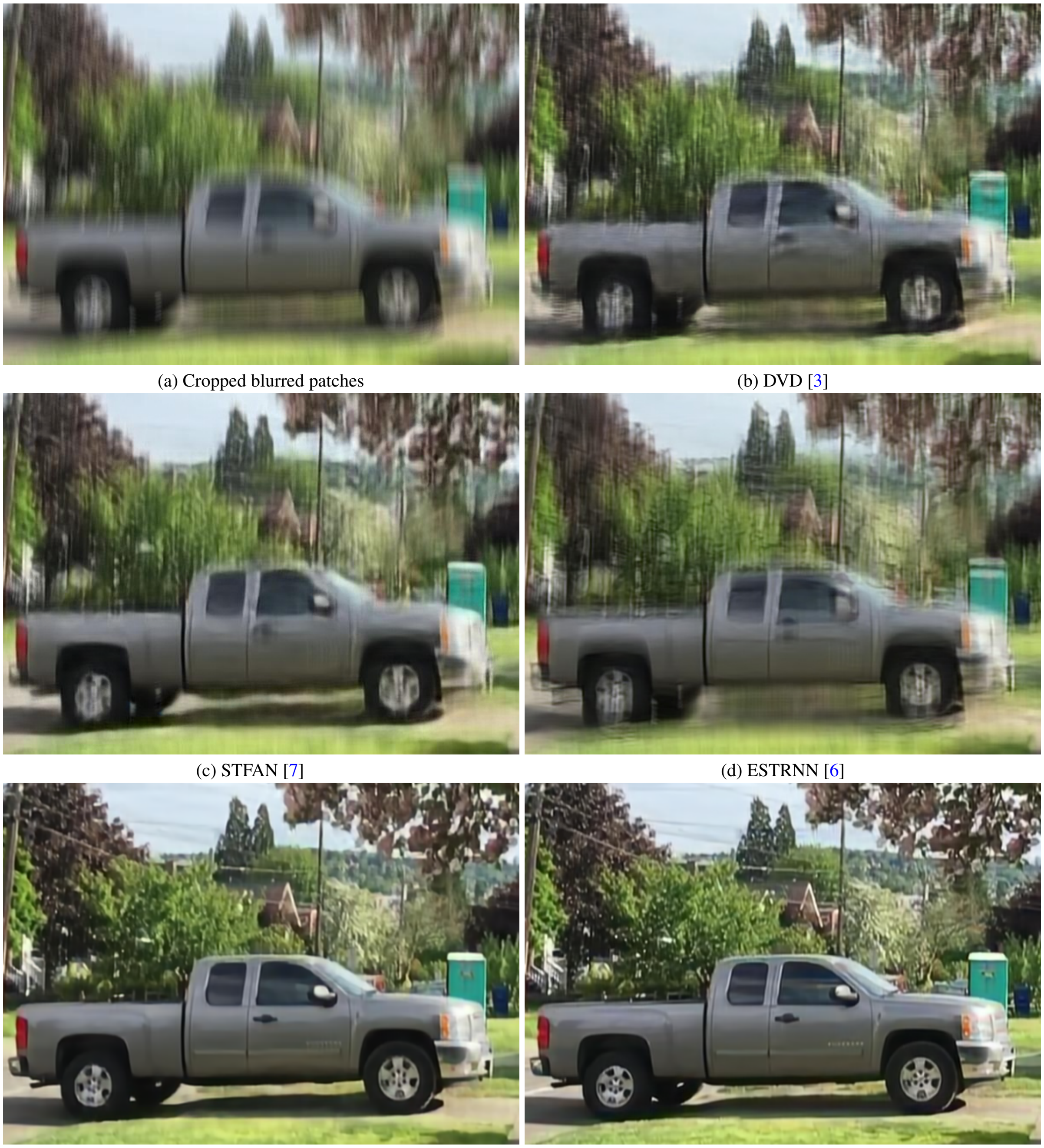
(e) CDVDTSP [2]

(f) DVDSEF [5]

(g) Ours

(h) Ours-L

Figure 6. Deblurred results on the GoPro test dataset [1]. The results shown in (b)-(f) still contain significant blur effects. In contrast, our method generates a better-deblurred frame, where the face of the girl is restored well.



(a) Cropped blurred patches

(b) DVD [3]

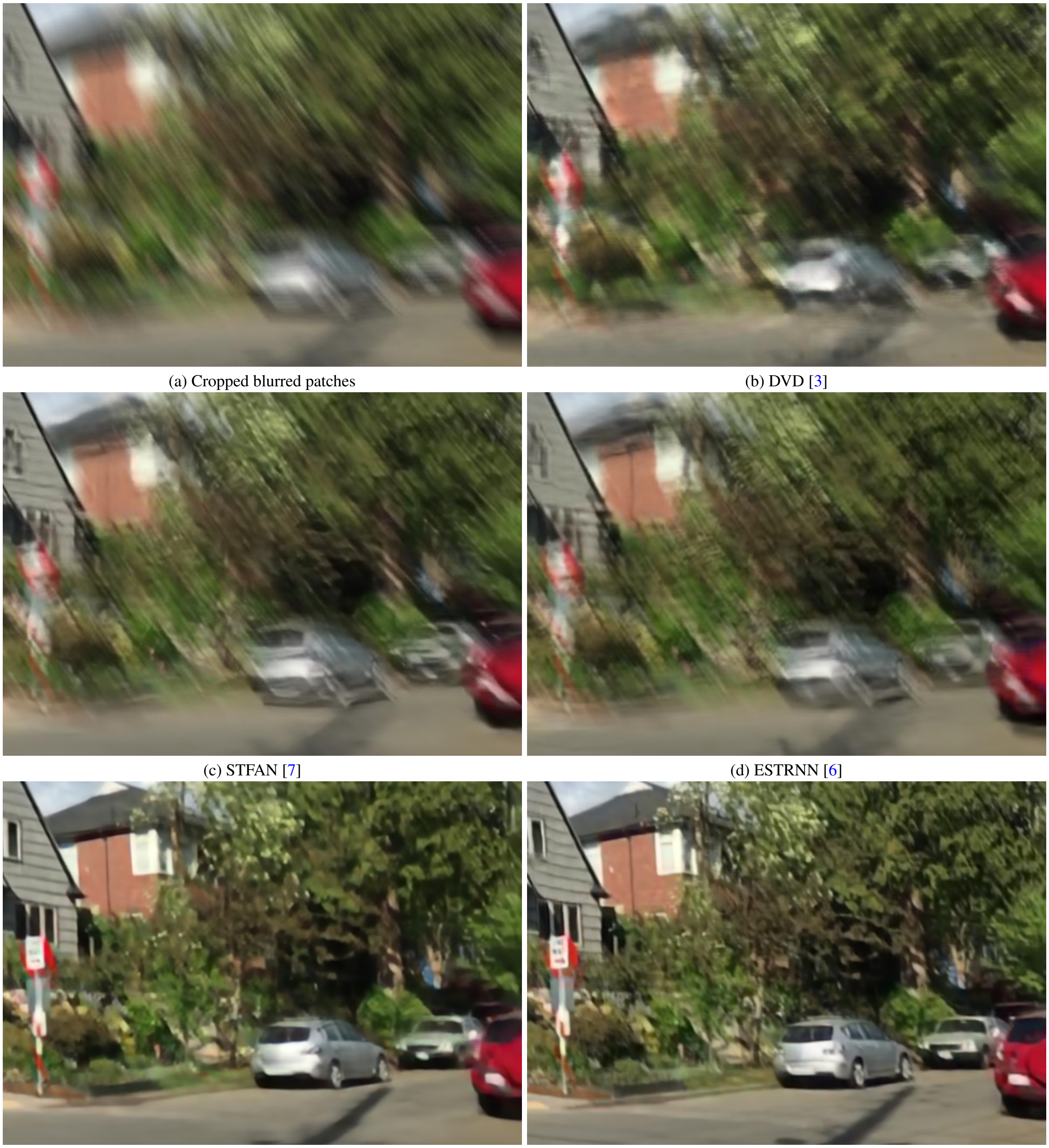
(c) STFAN [7]

(d) ESTRNN [6]

(e) CDVDTSP [2]

(f) Ours

Figure 7. Deblurred results on the DVD test dataset [3]. The results shown in (b)-(d) still contain significant blur effects. The deblurred frame by [2] is better as shown in (e). However, some structural details are not recovered well. In contrast, our method generates a better-deblurred frame with finer structural details.



(a) Cropped blurred patches

(b) DVD [3]

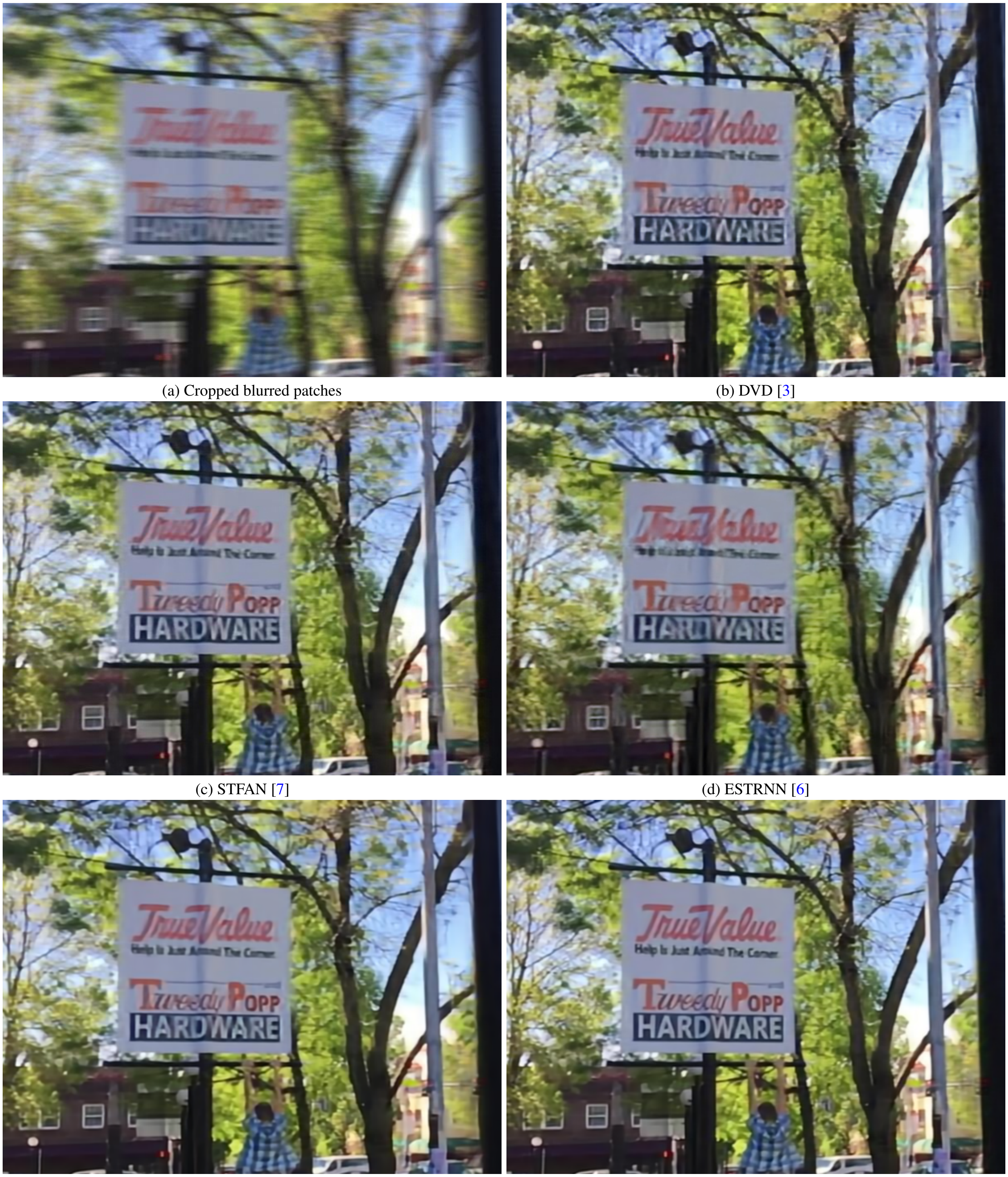
(c) STFAN [7]

(d) ESTRNN [6]

(e) CDVDTSP [2]

(f) Ours

Figure 8. Deblurred results on the DVD test dataset [3]. The results shown in (b)-(d) still contain significant blur effects. The deblurred frame by [2] is better as shown in (e). However, some structural details are not recovered well. In contrast, our method generates a better-deblurred frame with finer structural details.



(a) Cropped blurred patches

(b) DVD [3]

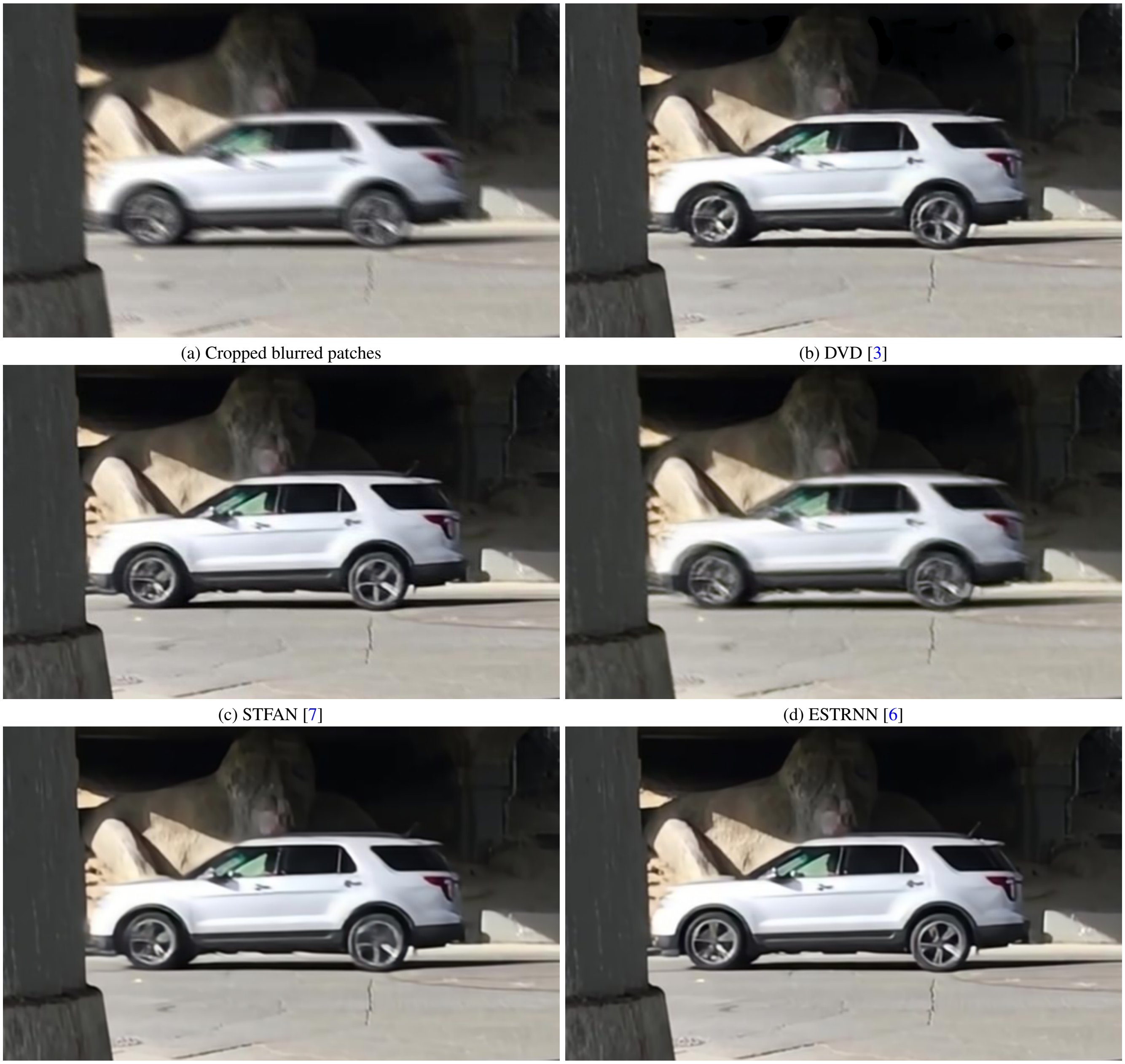
(c) STFAN [7]

(d) ESTRNN [6]

(e) CDVDTSP [2]

(f) Ours

Figure 9. Deblurred results on the DVD test dataset [3]. The evaluated methods do not generate clear frames, where the characters are not restored well. In contrast, our method generates a better-deblurred frame with clearer characters.



(a) Cropped blurred patches

(b) DVD [3]

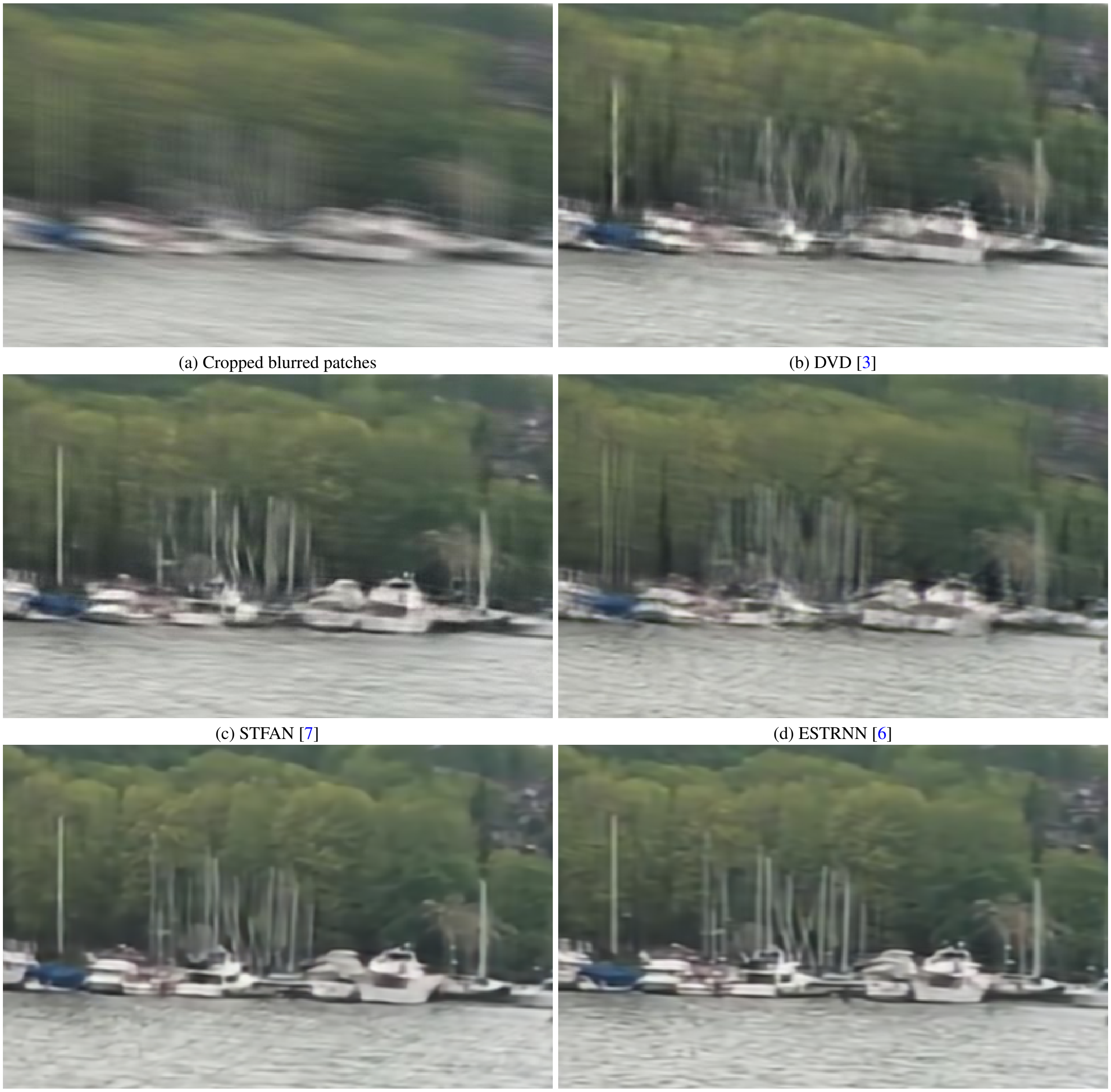
(c) STFAN [7]

(d) ESTRNN [6]

(e) CDVDTSP [2]

(f) Ours

Figure 10. Deblurred results on the DVD test dataset [3]. The evaluated methods do not generate clear frames. For example, the wheels of the car still contain significant blur effects. In contrast, our method generates a better-deblurred frame, where the wheels of the car are restored well.



(a) Cropped blurred patches

(b) DVD [3]

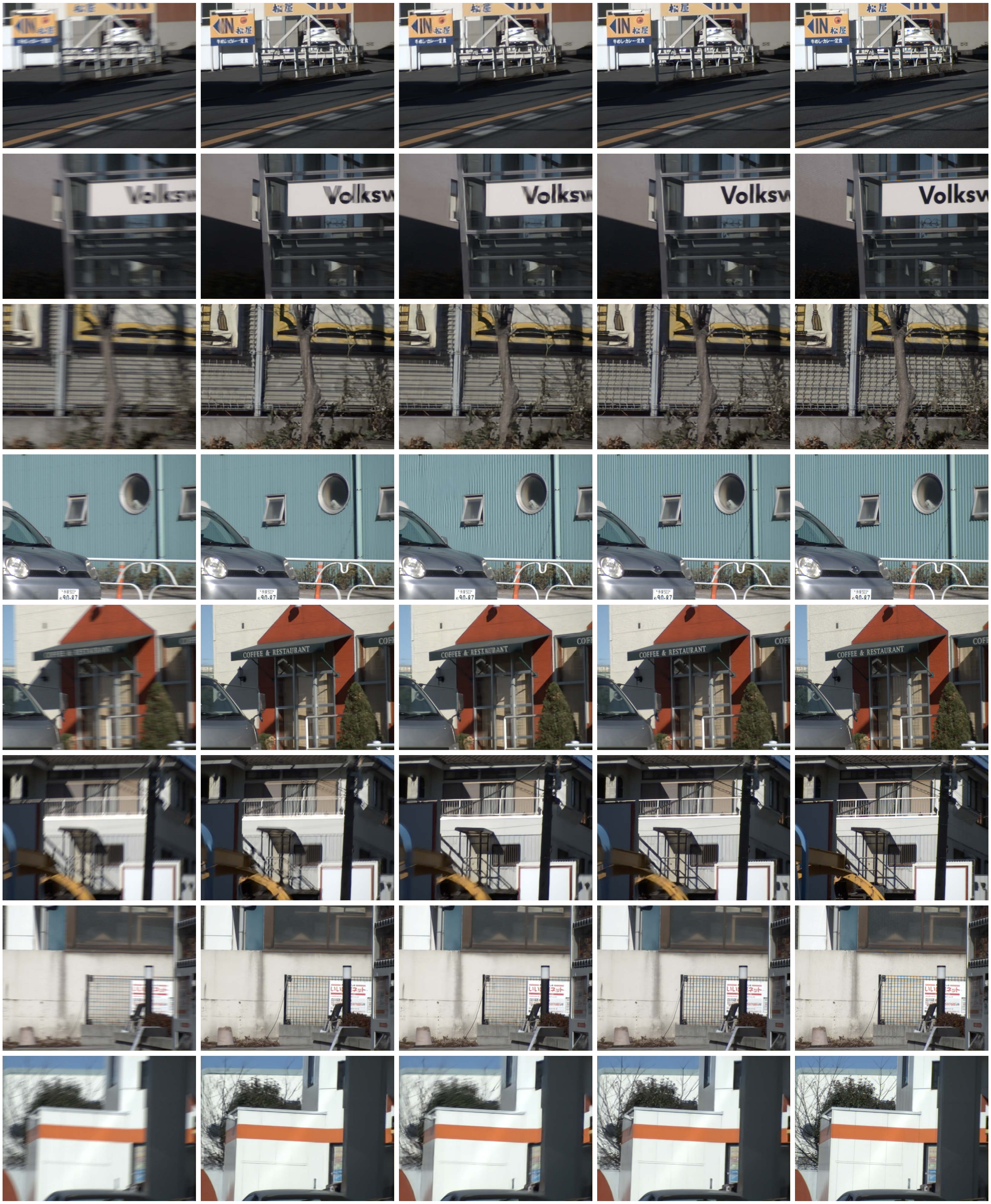
(c) STFAN [7]

(d) ESTRNN [6]

(e) CDVDTSP [2]

(f) Ours

Figure 11. Deblurred results on the DVD test dataset [3]. The evaluated methods do not generate clear frames In contrast, our method generates a better-deblurred frame.



(a) Blurred frames (b) ESTRNN [6] (c) CDVDTSP [2] (d) Ours (e) GT

Figure 12. Qualitative comparisons on the real-world dataset [6]. The results by the ESTRNN [6] and CDVDTSP [2] still contain blur effects and some structures are not restored well. In contrast, the proposed method generates much clearer frames, which are visually close to the ground truth ones.