Supplementary Material: Learning Event-Driven Video Deblurring and Interpolation

Songnan Lin[1⋆], Jiawei Zhang[2 ⋆⋆], Jinshan Pan[3], Zhe Jiang[2], Dongqing Zou[2], Yongtian Wang[1], Jing Chen[1 ⋆⋆], and Jimmy Ren[2]

[1] Beijing Institute of Technology, Beijing, China
[2] SenseTime Research, Shenzhen, China
[3] Nanjing University of Science and Technology, Nanjing, China

1 Overview

In this supplementary material, we first provide the detailed structures of the proposed *IntegralNet* and *GateNet* in Section 2. We present further analysis and discussion on each component of the proposed method in Section 3. We give a comprehensive quantitative evaluation in Section 4. We show additional results on video deblurring on synthetic and real data in Section 5 and Section 6, respectively. Finally, we provide more visual comparisons on simultaneous video deblurring and interpolation on real-world blurry videos in Section 7.

2 Network Architecture

The proposed event-driven video deblurring and interpolation algorithm consists of two sub-networks: 1) an *IntegralNet*, which contains event feature extraction, dynamic filter generation and multi-residual prediction, to estimate the residuals D between blurry and sharp frames as well as those I between sharp images, 2) a *GateNet* to predict the weights for fusing the initial reconstructed results in an adaptive selection manner. Table 1, Table 2, Table 3 and Table 4 list the configurations of the three modules of *IntegralNet* and *GateNet*, respectively.

⋆ This work was done when Songnan Lin was an intern at SenseTime.
⋆⋆ Corresponding authors: zhjw1988@gmail.com; chen74jing29@bit.edu.cn

Table 1. Configurations of the event feature extraction module in *IntegralNet*. 'Conv' denotes the convolution layer. 'Res' denotes the residual block including two convolution layers. '$C(\cdot)$' denotes the concatenation operation. E_{i-1}, E_i denote the previous and current events which are firstly divided into 2N equal-time-interval bins and further each bin is divided into M equal-size chunks, stacked as 2MN-channel features for 2N-time video reconstruction. U_i denotes the extracted features.

Layer	Input	Output	In Channel	Out Channel	Kernel Size	Stride
Conv1	$C(E_i, E_{i-1})$	conv1	4MN	64	3×3	1
Res1_1	conv1	res1_1	64	64	3×3	1
Res1_2	res1_1	res1_2	64	64	3×3	1
Conv2	res1_2	conv2	64	96	3×3	2
Res2_1	conv2	res2_1	96	96	3×3	1
Res2_2	res2_1	res2_2	96	96	3×3	1
Conv3	res2_2	conv3	96	128	3×3	2
Res3_1	conv3	res3_1	128	128	3×3	1
Res3_2	res3_1	U_i	128	128	3×3	1

Table 2. Configurations of the dynamic filter generation module in *IntegralNet*. The network inputs the previous and current blurry frames B_{i-1}, B_i, corresponding events E_{i-1}, E_i and the previously recovered sharp frames $S_{i-1,j}$. And it outputs the dynamic filters denoted as $\mathcal{F}_i$ with size $K \times K \times 1$. 'Conv' denotes the convolution layer. 'Res' denotes the residual block including two convolution layers. '$C(\cdot)$' denotes the concatenation operation.

Layer	Input	Output	In Channel	Out Channel	Kernel Size	Stride
Conv1	$C(B_i, B_{i-1}, E_i, E_{i-1}, S_{i-1,j})$	kconv1	2+4MN+2N	64	3×3	1
Res1_1	kconv1	kres1_1	64	64	3×3	1
Res1_2	kres1_1	kres1_2	64	64	3×3	1
Conv2	kres1_2	kconv2	64	96	3×3	2
Res2_1	kconv2	kres2_1	96	96	3×3	1
Res2_2	kres2_1	kres2_2	96	96	3×3	1
Conv3	kres2_2	kconv3	96	128	3×3	2
Res3_1	kconv3	kres3_1	128	128	3×3	1
Res3_2	kres3_1	kres3_2	128	128	3×3	1
Conv4	kres3_2	kconv4	128	$128 \times K \times K$	1×1	1
Reshape	kconv4	$\mathcal{F}_i$	-	-	-	-

Table 3. Configurations of the multi-residual prediction module in *IntegralNet*. 'Conv' denotes the convolution layer. 'Res' denotes the residual block including two convolution layers. 'Deconv' denotes the transposed convolution layer. '$C(\cdot)$' denotes the concatenation operation. V_i denotes the transformed event features via Eq. (5) (Please see the original paper). $D_{i\to i,0}$ denotes the residual between the blurry image and the keyframe. $I_{i-1,j\to i,0;j\in(-N,N]}$ denotes the residuals between the $2N$ previously recovered sharp frames and the current sharp keyframe. $I_{i,0\to i,j;j\neq 0}$ denotes the residuals between the latent keyframe and the $2N-1$ interpolated frames.

Layer	Input	Output	In Channel	Out Channel	Kernel Size	Stride
Deconv3	V_i	updeconv3	128	96	4×4	2
Conv3	C(updeconv3,res2_2)	upconv3	192	96	3×3	1
Res3_2	upconv3	upres3_2	96	96	3×3	1
Res3_1	upres3_2	upres3_1	96	96	3×3	1
Deconv2	upres3_1	updeconv2	96	64	4×4	2
Conv2	C(updeconv2,res1_2)	upconv2	128	64	3×3	1
Res2_2	upconv2	upres2_2	64	64	3×3	1
Res2_1	upres2_2	upres2_1	64	64	3×3	1
Conv1_D	upres2_1	upconv1_d	64	32	3×3	1
Res1_2_D	upconv1_d	upres1_2_d	32	32	3×3	1
Res1_1_D	upres1_2_d	upres1_1_d	32	32	3×3	1
Conv0_D	upres1_1_d	$D_{i\to i,0}$	32	1	3×3	1
Conv1_I1	upres2_1	upconv1_i1	64	32	3×3	1
Res1_2_I1	upconv1_i1	upres1_2_i1	32	32	3×3	1
Res1_1_I1	upres1_2_i1	upres1_1_i1	32	32	3×3	1
Conv0_I1	upres1_1_i1	$I_{i-1,j\to i,0}$	32	2N	3×3	1
Conv1_I2	upres2_1	upconv1_i2	64	32	3×3	1
Res1_2_I2	upconv1_i2	upres1_2_i2	32	32	3×3	1
Res1_1_I2	upres1_2_i2	upres1_1_i2	32	32	3×3	1
Conv0_I2	upres1_1_i2	$I_{i,0\to i,j}$	32	2N-1	3×3	1

Table 4. Configurations of *GateNet*. *GateNet* inputs the blurry input B_i, the corresponding events E_i and initial recovered frames $F_{i,j,k;j\in(-N,N],k\in[0,2N]}$. And it outputs a gate map for an adaptive selection, denoted as $M_{i,j,k;j\in(-N,N],k\in[0,2N]}$. '3DConv' denotes the 3D convolution layer. Please see the original paper for more details.

Layer	Input	Output	In Channel	Out Channel	Kernel Size	Stride
Reshape	$B_i, E_i, F_{i,j,k}$	In	-	-	-	-
3DConv1	In	3dconv1	M+2N+2	64	$3\times 3\times 3$	1
3DConv2	3dconv1	3dconv2	64	64	$3\times 3\times 3$	1
3DConv3	3dconv2	3dconv3	64	2N+1	$3\times 3\times 3$	1
Sigmoid	3dconv3	$M_{i,j,k}$	-	-	-	-

3 Analysis and Discussion

In this section, we give further discussions on the effectiveness of each component in our event-driven video deblurring and interpolation network.

3.1 Effectiveness of Physical-Based Framework

The proposed algorithm is designed based on the physical model of the event-based video reconstruction. We predict the residuals I and D and apply multiply operation to them according to Eq. 2 and Eq. 4 (Please see the original paper for details). To demonstrate the effectiveness of the physical-based framework, we conduct an experiment that adds the residuals and the intensity images up, as already used in pure image-based algorithms (denoted as 'Addition'). More visual comparisons are illustrated in Fig. 1. 'Addition' predicts blurry addition residuals (Fig. 1 (b)) and thus generates smooth results but with more artifacts finally (Fig. 1 (e)). However, as the proposed method is based on the physical model, which makes it easy to calculate the multiplication residuals (Fig. 1 (c)) from event data, it is robust to severely-blurred frames and restores images with more details and fewer artifacts (Fig. 1(f)).

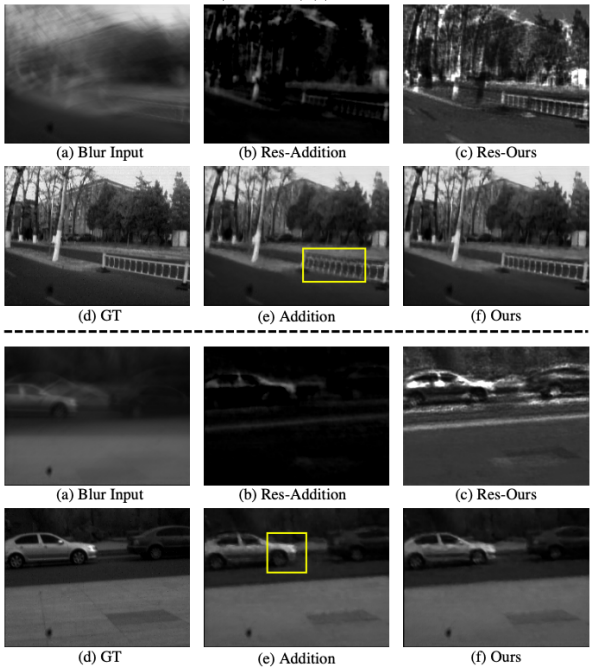


Fig. 1. Visual comparison with 'Addition'. (a) is the blurry input, while (d) is the ground truth. 'Addition' replaces multiplication with addition. 'Res-' in (b)(c) denote the learned residuals between the keyframes and the interpolated frames. The proposed method makes it easy to predict multiplication residuals and restores sharper results with finer details, which demonstrates the effectiveness of the physical-based framework.

3.2 Effectiveness of Dynamic Filtering

To handle the events triggered by the spatially variant threshold, we propose to integrate the dynamic filters when estimating residuals. To validate the above discussions, we remove the dynamic filter generation module and feed its inputs $(S_{i-1}, S_i, E_{i-1}, E_i, S_{i-1})$ into the event feature extraction directly for a fair comparison (denoted as 'w/o DF'). More visual comparisons between our method and 'w/o DF' are illustrated in Fig. 2. Due to the lack of compensation for the spatially variant triggering threshold, it provides overly-smooth residuals (see Fig. 2 (b)) compared to ours (see Fig. 2 (c)). And thus, it cannot restore the missing details in the final results (see Fig. 2 (e)). Moreover, it introduces strong accumulated noises, especially at edges.

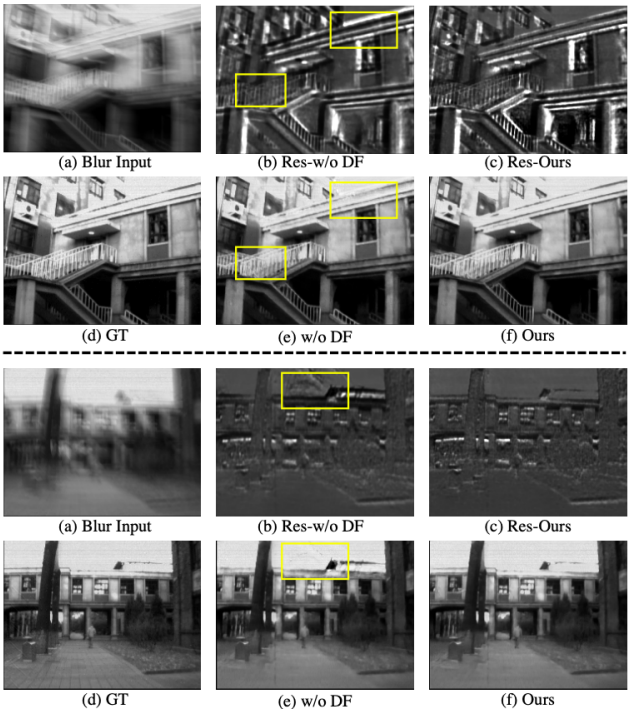


Fig. 2. Visual comparison with 'w/o DF'. (a) is the blurry input, while (d) is the ground-truth frame. 'Res-' in (b)(c) denote the learned residuals between the keyframes and the interpolated frames. 'w/o DF' represents removing the dynamic filtering. The proposed method restores sharper residuals and thus generates clearer images with more details and fewer artifacts, which demonstrates the effectiveness of dynamic filtering.

Furthermore, to intuitively understand the dynamic filtering, we provide generated filters in Fig. 3. Given sharp frames and the respective events, the average triggering thresholds can be inferred from Eq. 2 in the original paper as

$$\bar{c}(x,y) = \frac{log(S_{i',j'}(x,y)/S_{i,j}(x,y))}{\sum_{t_m \in \Omega_{i,j \to i',j'}} p_m \cdot \mathbf{1}(x_m, y_m, x, y)}. \tag{1}$$

As can be seen from Fig. 3 (e), the triggering thresholds are not uniform across the image plane. The predicted filters in Fig. 3 (f) are spatially variant and closely related to the average thresholds, which can facilitate to minimize the effects of the non-uniform threshold.

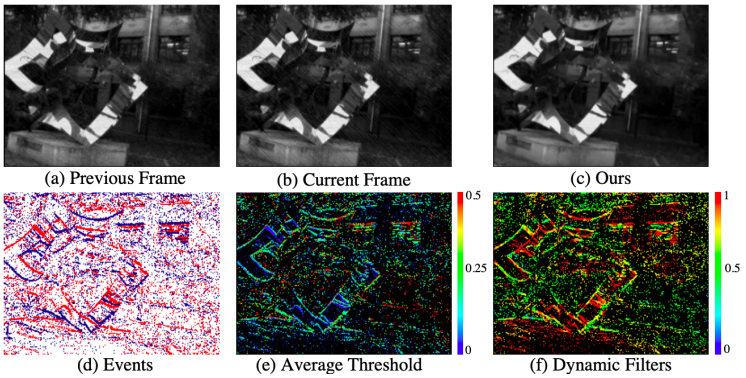


Fig. 3. Effectiveness of dynamic filtering. (a)(b) are two latent sharp frames, and (d) is the corresponding interval of events captured with an event camera. The average threshold of each pixel in this interval (e) is estimated from (a)(b)(d) using Eq. 1. (f) is the visualization of the generated filters that are spatially variant and closely related to (e).

3.3 Effectiveness of Previous Information

We note that the existing event-based video deblurring and interpolation algorithms [4,1] bring one blurry frame alive without considering additional information that exists across adjacent frames. To validate the effectiveness of utilizing previous information, we compare a method that only estimates the keyframes $C_{i,0}$ from current blurry inputs without the ones $P_{i,0,j}$ from the previously recovered frames (denoted as 'w/o Pre'. Please see the original paper). The visual comparisons shown in Fig. 4 indicate that involving previous information is more effective for video deblurring and reconstruction.

3.4 Effectiveness of Frame Fusion

To integrate the $2N + 1$ initial recovered results $F_{i,j,k}$ in an adaptive selection manner, the proposed frame fusion step utilizes the information from the blurry frame, event data and the initial results to generate a gate map and then obtains the final results by weighted summation. To demonstrate the effectiveness of this design, we compare the method that removes the estimation of the gate map but feeds the initial results into three 3D convolution layers directly to estimate the final results (denoted as 'w/o ASF'). We show more visual comparisons in Fig. 5. The proposed method keeps finer details, which validates the effectiveness of the proposed frame fusion.

Furthermore, initial results and gate maps are illustrated in Fig 5 as well. It can be noted that as the frames inferred from the blurry images (Fig. 5 (c)) seem

Fig. 4. Visual comparison with 'w/o Pre'. (a) is the blurry input, while (b) is the ground-truth frame. 'w/o Pre' represents removing the previous information in the keyframe estimation step. The proposed method restores sharper results with finer details, which demonstrates the effectiveness of previous information.

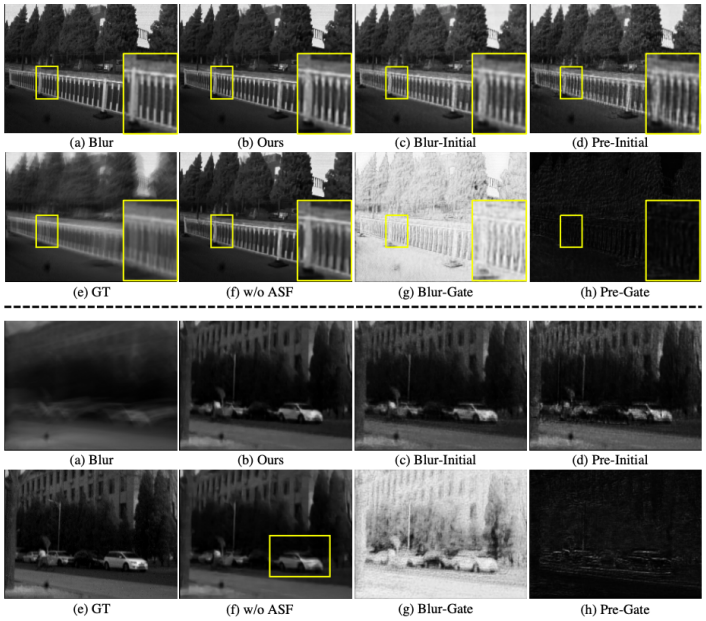


Fig. 5. Visual comparison with 'w/o ASF'. (a) blurry input. (b) our reconstruction results. (c) initial results estimated from blurry inputs. (d) an example of initial results estimated from the previous recovered frames. (e) ground truth. (f) reconstruction results of 'w/o ASF' which removes the adaptively-selected fusion. (g)(h) corresponding gate maps of (c)(d). The proposed method can integrate the initial results in an adaptive selection scheme and keep more details, which demonstrates the effectiveness of our frame fusion.

photorealistic but blurry at edges, the gate map tends to 0 at edges but 1 for other regions (Fig. 5 (g)). Instead, the ones from the last recovered frames are sharp with more details but contain significant artifacts (Fig. 5 (d)), thus the gate map tends to 1 at edges but 0 for other places (Fig. 5 (h)). The proposed frame fusion module can integrate the initial results in an adaptive selection scheme and keep more details, which is more effective for video reconstruction.

4 Comprehensive Quantitative Evaluation

In this section, we give comprehensive comparison against the state-of-the-art algorithms, including STFAN [7], TNTT [2], E2V [5], BHA [4] and LEMD [1]. To demonstrate the effectiveness of the proposed framework, we further compare the enhanced version of the single-sensor algorithms (denoted as *) which feed both event and intensity data into networks. BHA runs on an Intel Xeon E5 CPU and other algorithms run on a GeForce GTX 1080 GPU. Table 5 shows quantitative results in terms of average PSNR, SSIM, model size and running time. The proposed network performs favorably against the state-of-the-art algorithms. Moreover, it contains smaller model size and it is more effective than the most of the state-of-the-art algorithms except E2V and E2V*.

Table 5. Video deblurring and reconstruction performance on the synthetic subset of Blur-DVS [1], in terms of average PSNR, SSIM, parameter numbers ($\times 10^6$) of different networks and running time (ms per recovered frame) with image size of 180×240.

Average results of video deblurring							
Methods	E2V[5]	E2V*	STFAN[7]	STFAN*	BHA[4]	LEMD[1]	Ours
PSNR	16.89	24.81	19.03	30.18	22.43	26.48	**30.57**
SSIM	0.597	0.790	0.518	0.897	0.715	0.839	**0.904**
Params (M)	10.71	10.71	5.36	5.38	-	5.37	**4.80**
Time (ms)	**4.33**	4.36	11.49	11.53	205.14	26.27	14.17
Average results of video deblurring and interpolation							
Methods	E2V[5]	E2V*	TNTT[2]	TNTT*	BHA[4]	LEMD[1]	Ours
PSNR	16.60	24.10	19.05	29.02	22.06	25.33	**29.65**
SSIM	0.587	0.777	0.521	0.875	0.699	0.827	**0.890**
Params (M)	10.71	10.71	10.68	10.88	-	9.13	**5.00**
Time (ms)	**4.33**	4.36	7.36	7.90	57.13	13.13	4.68

* denotes the enhanced version of the corresponding single-sensor algorithm. See text for more details.

5 More Results on Synthetic Blurry Videos

In this section, we present additional comparisons with the state-of-the-art algorithms, including STFAN [7], E2V [5], BHA [4], CIE [6] and LEMD [1], on GoPro [3] and the synthetic subset of Blur-DVS [1]. To demonstrate the effectiveness of the proposed framework, we further compare the enhanced versions of the single-sensor algorithms. STFAN* feeds additional event data into the spatio-temporal filter adaptive module of STFAN to assist frame alignment and deblurring. E2V* feeds events together with intensity frames into E2V for each of its recurrent reconstruction step. The compared methods as well as the enhanced variants are trained with hybrid inputs following the corresponding released procedures. The proposed method generates much sharper results with fewer noises and artifacts.

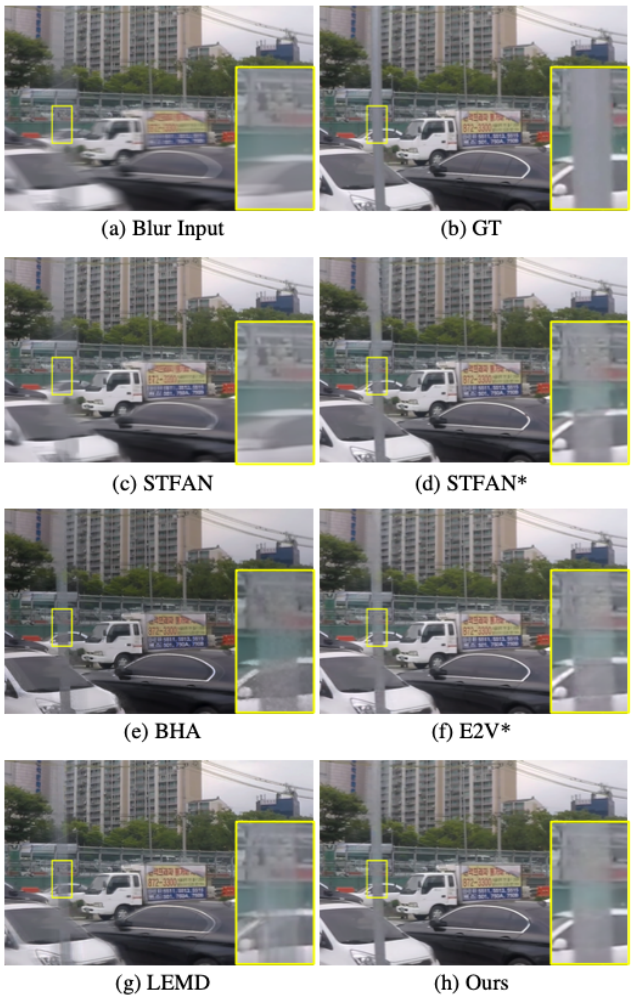


Fig. 6. Visual comparisons on video deblurring on the GoPro [3] dataset.

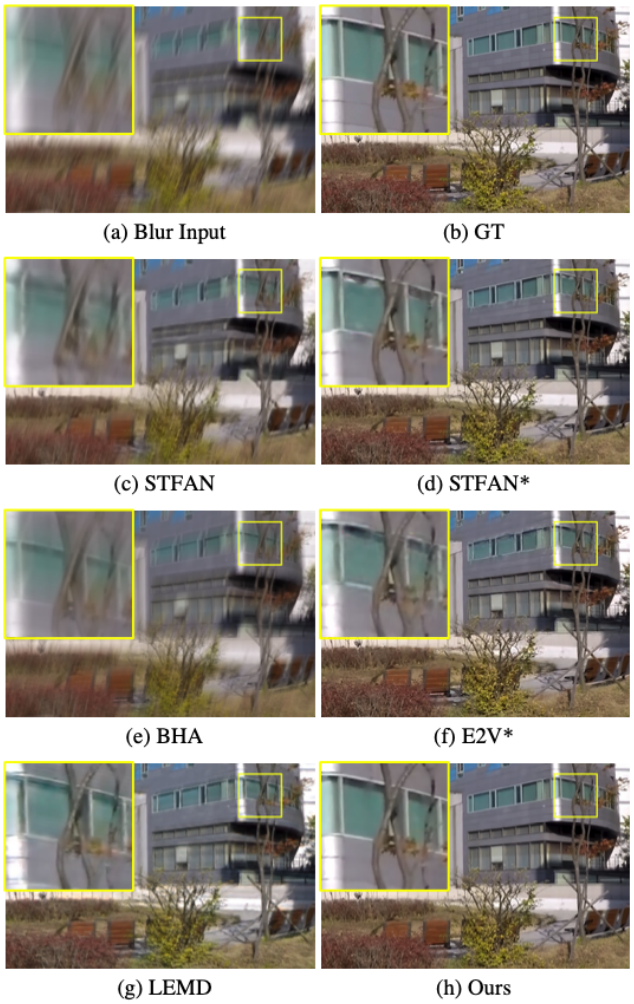


Fig. 7. Visual comparisons on video deblurring on the GoPro [3] dataset.

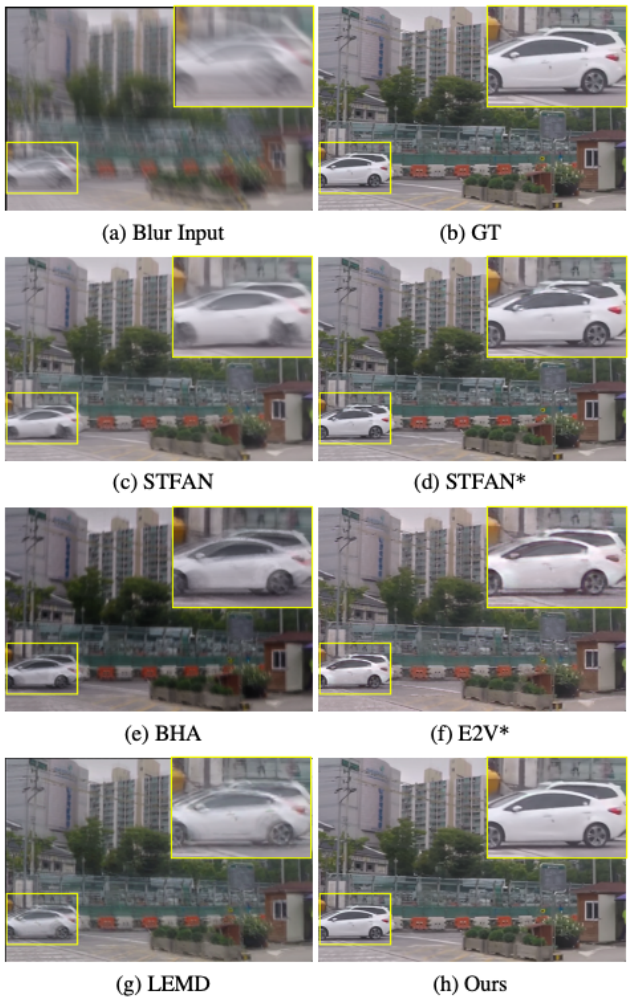


Fig. 8. Visual comparisons on video deblurring on the GoPro [3] dataset.

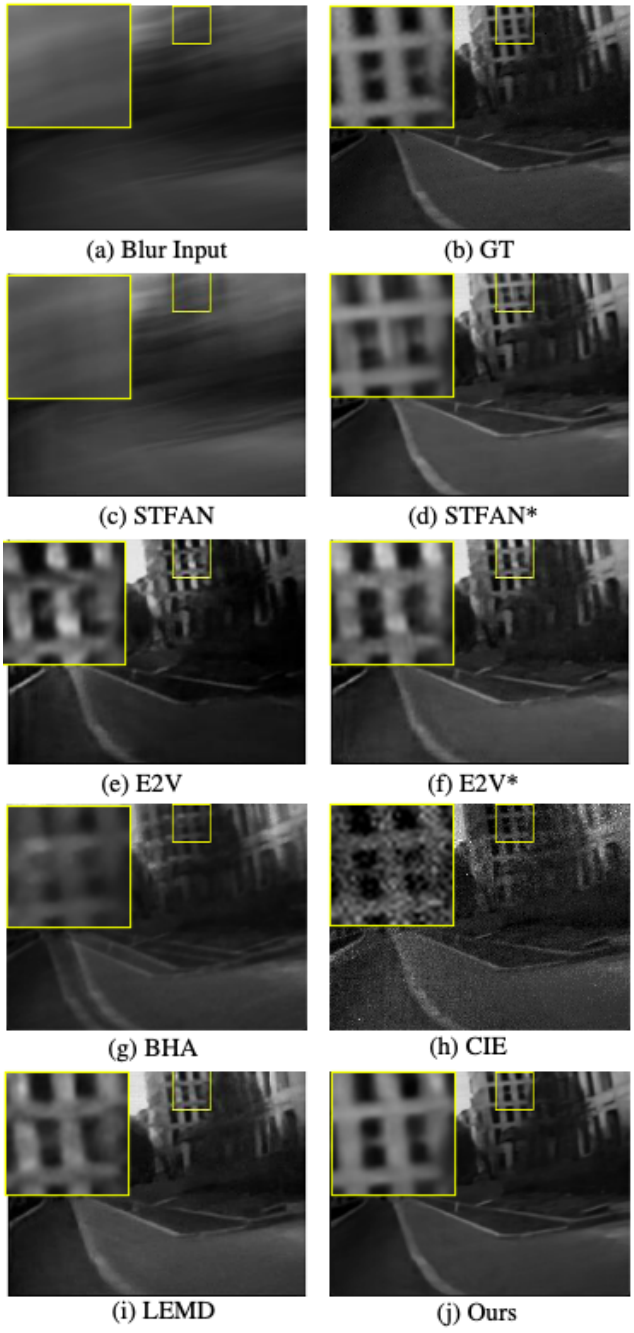


Fig. 9. Visual comparisons on video deblurring on the synthetic subset of Blur-DVS [1].

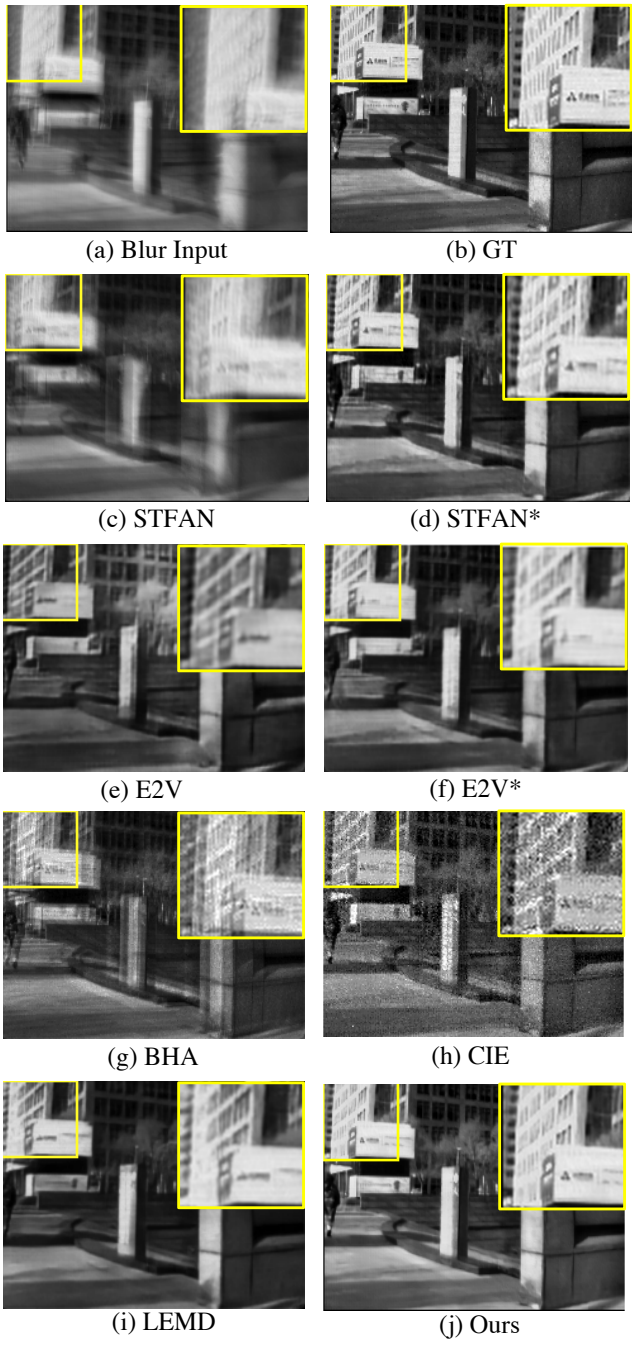


Fig. 10. Visual comparisons on video deblurring on the synthetic subset of Blur-DVS [1].

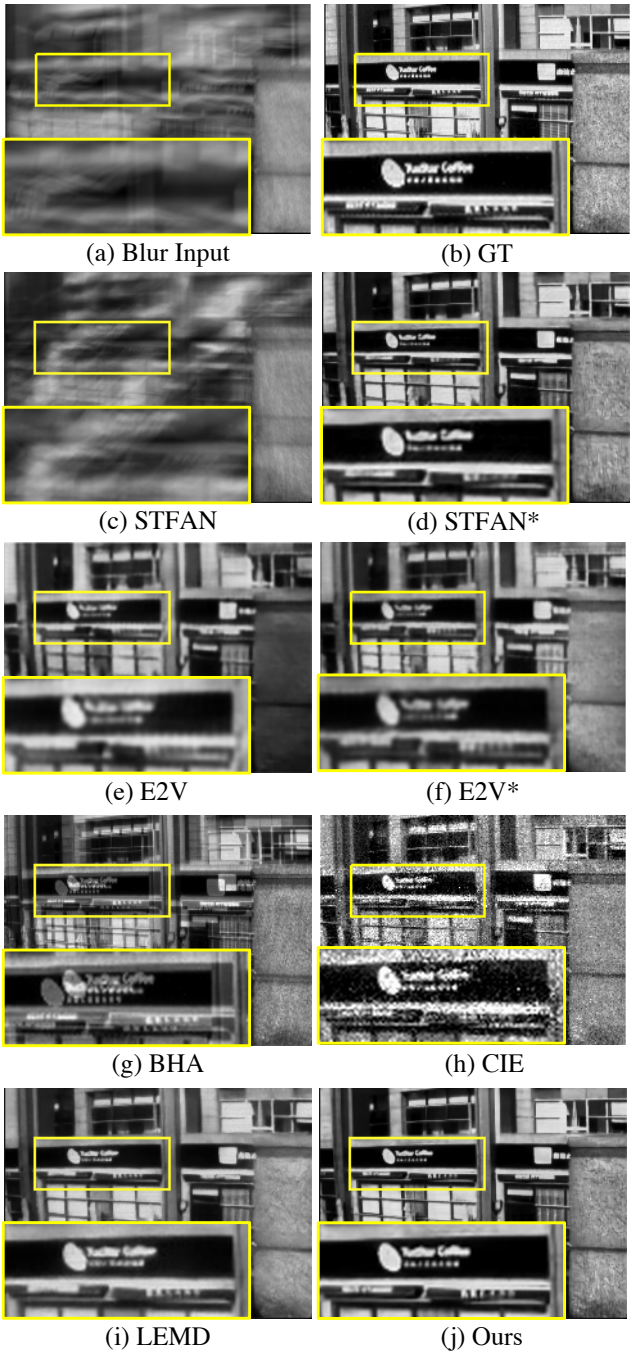


Fig. 11. Visual comparisons on video deblurring on the synthetic subset of Blur-DVS [1].

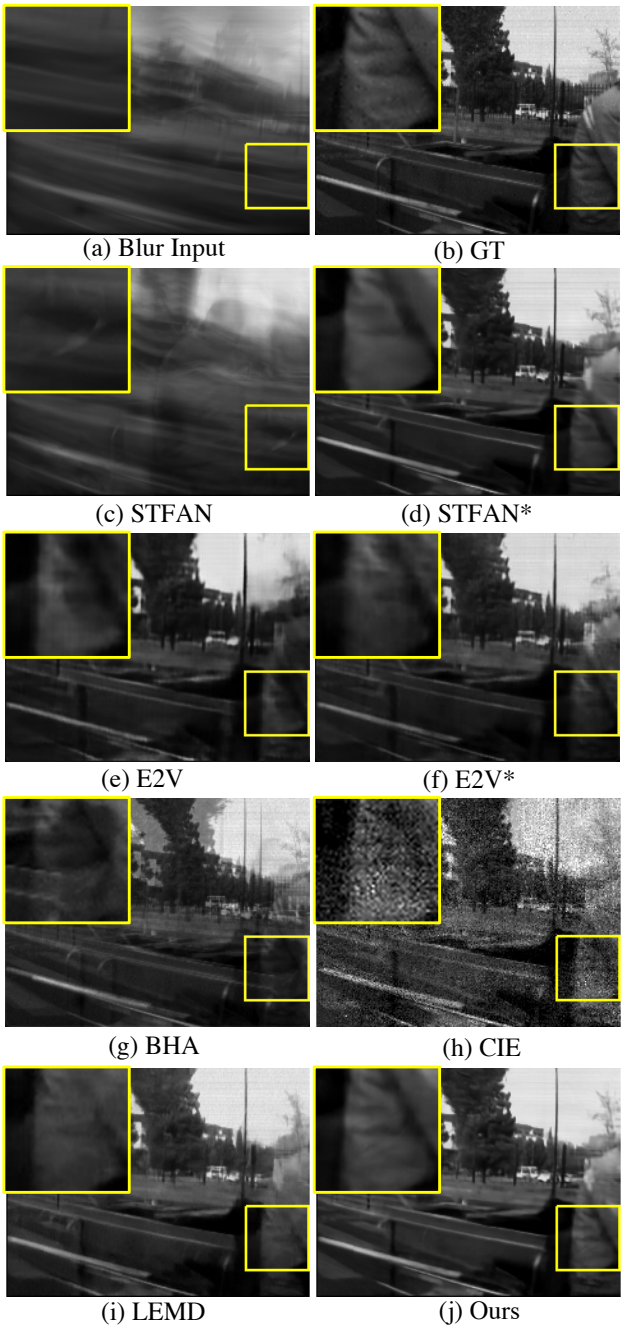


Fig. 12. Visual comparisons on video deblurring on the synthetic subset of Blur-DVS [1].

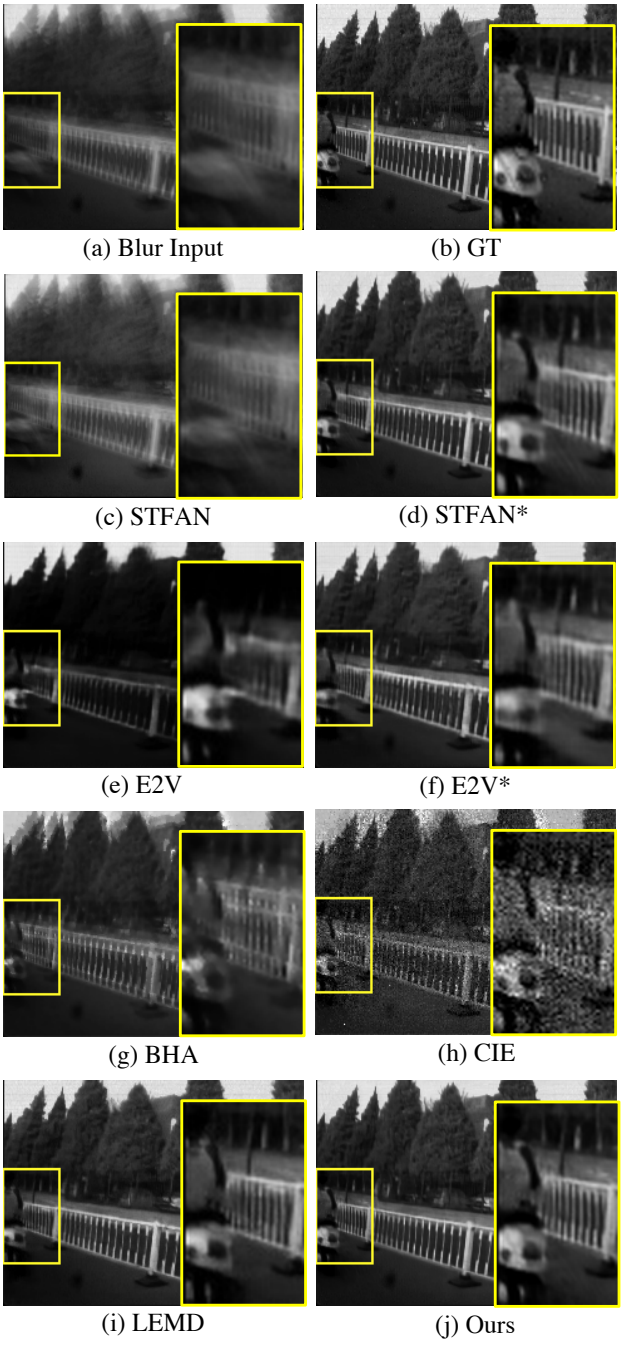


Fig. 13. Visual comparisons on video deblurring on the synthetic subset of Blur-DVS [1].

6 More Results on Real-World Blurry Videos

In this section, we present additional comparisons with the state-of-the-art algorithms, including STFAN [7], E2V [5], BHA [4] and LEMD [1], on real-world blurry videos [1,4]. To demonstrate the effectiveness of the proposed framework, we further compare the enhanced versions of the single-sensor algorithms. STFAN* feeds additional event data into the spatio-temporal filter adaptive module of STFAN to assist frame alignment and deblurring. E2V* feeds events together with intensity frames into E2V for each of its recurrent reconstruction step. The compared methods as well as the enhanced variants are trained with hybrid inputs following the corresponding released procedures. The proposed method generates much cleaner results with more details and fewer noises and artifacts.

Fig. 14. Visual comparisons on video deblurring on the real subset of Blur-DVS [1].

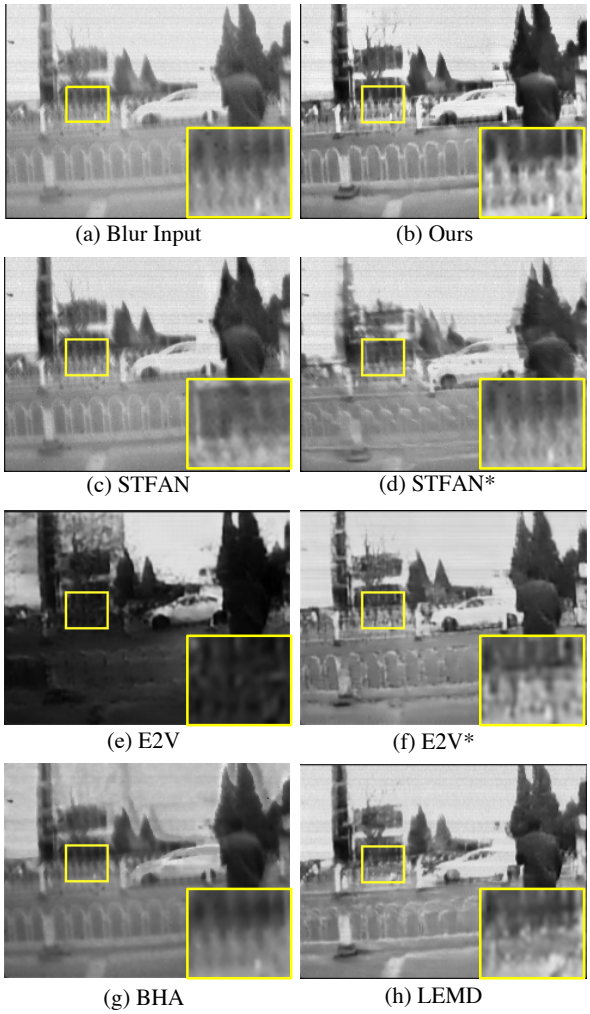


Fig. 15. Visual comparisons on video deblurring on the real subset of Blur-DVS [1].

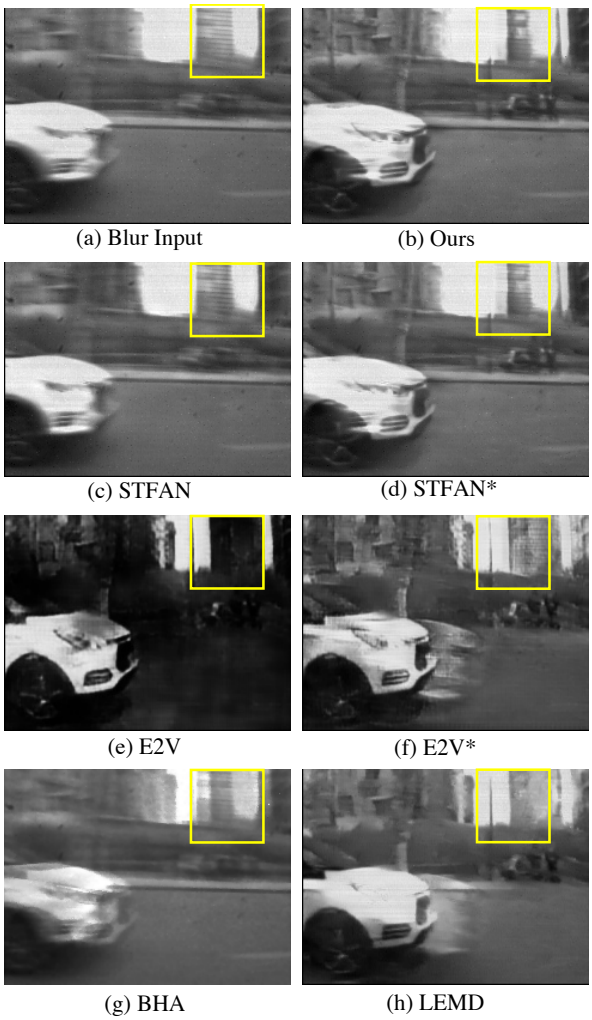


Fig. 16. Visual comparisons on video deblurring on the real subset of Blur-DVS [1].

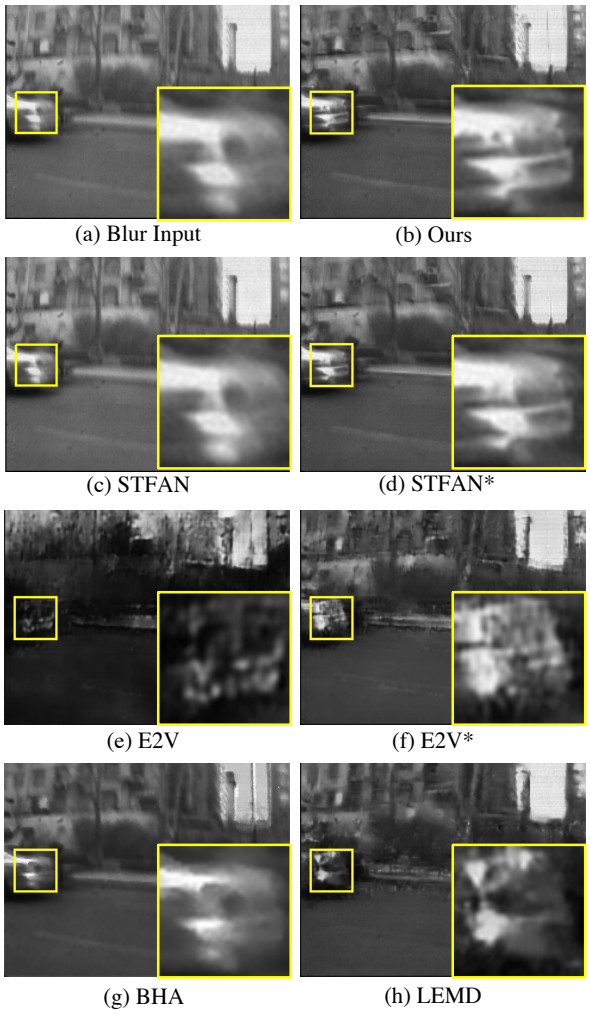


Fig. 17. Visual comparisons on video deblurring on the real subset of Blur-DVS [1].

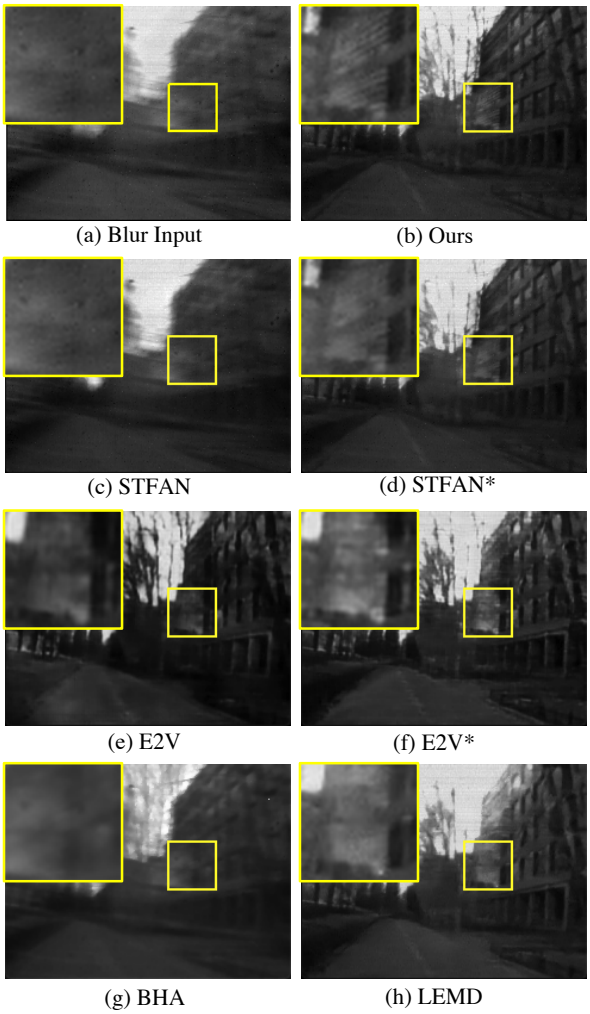


Fig. 18. Visual comparisons on video deblurring on the real subset of Blur-DVS [1].

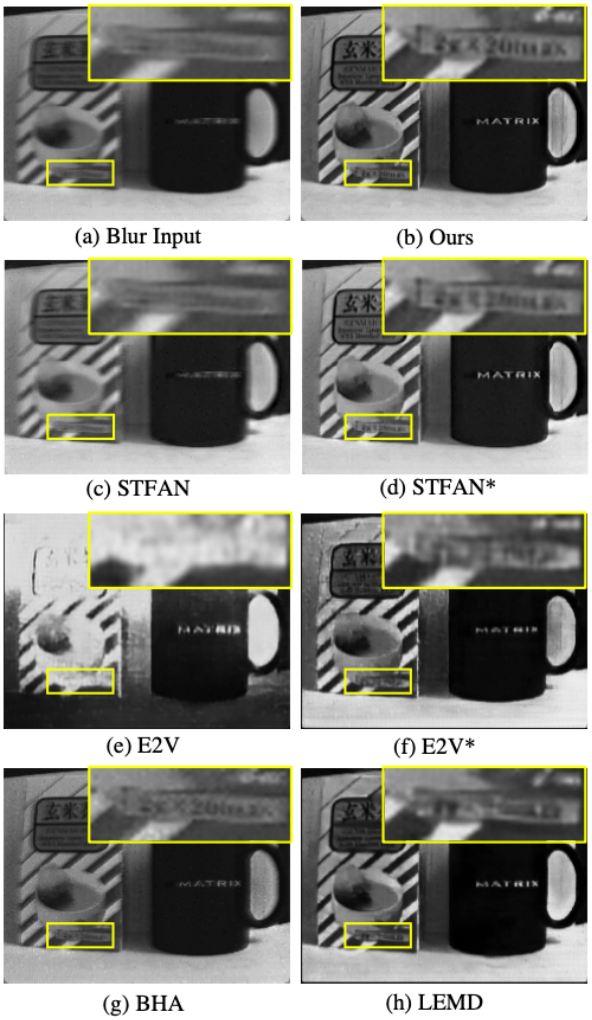


Fig. 19. Visual comparisons on video deblurring on the real-world blurry videos [4] with camera shaking.

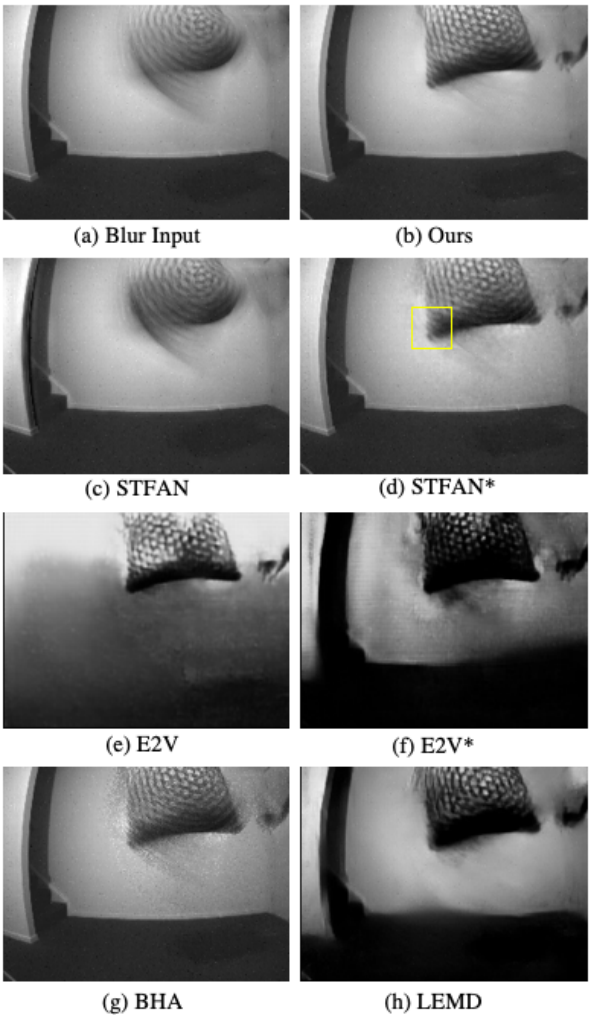


Fig. 20. Visual comparisons on video deblurring on the real-world blurry videos [4] with high-speed object motion.

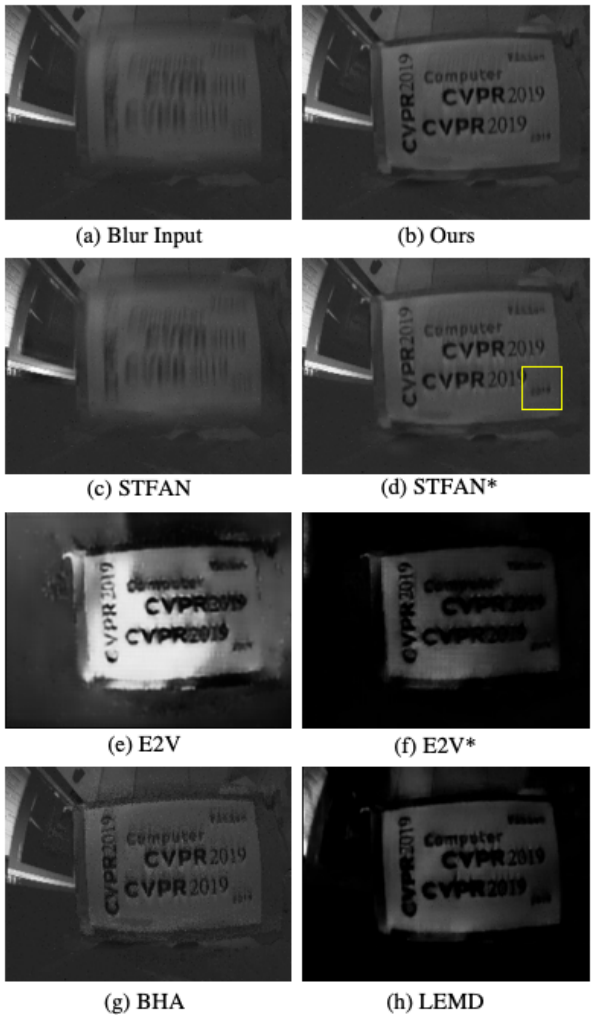


Fig. 21. Visual comparisons on video deblurring on the real-world blurry videos [4] in a low lighting condition.

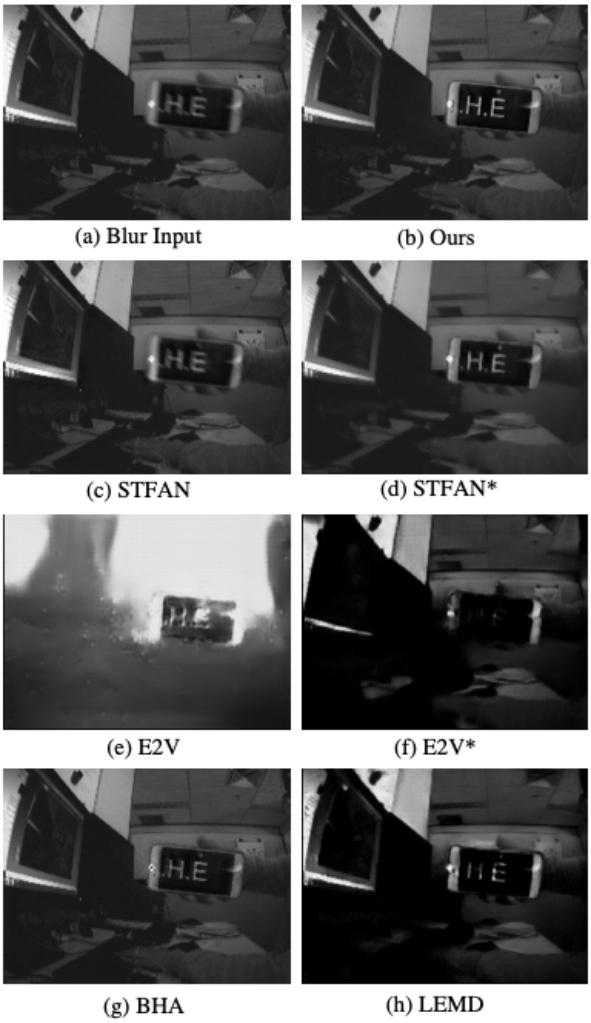


Fig. 22. Visual comparisons on video deblurring on the real-world blurry videos [4] in a complex dynamic condition.

7 More Results on Video Reconstruction

In this section, we present additional comparisons on simultaneous video deblurring and interpolation with the existing video reconstruction algorithms, including BHA [4] and LEMD [1], on the synthetic and real-world blurry videos. Moreover, to demonstrate the effectiveness of the proposed framework, we further compare the enhanced versions of an image-based video reconstruction method [2] (denoted as TNTT*). TNTT* feeds events and blurry intensity frames into both keyframe deblurring network and frame interpolation network of TNTT. The deep learning-based methods are trained with hybrid inputs following the corresponding released procedures. The proposed method generates much cleaner results with more details and fewer noises and artifacts.

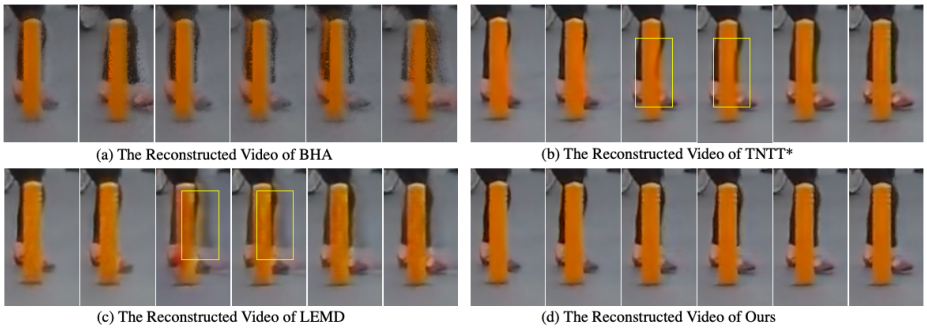


Fig. 23. Visual comparisons on video reconstruction on the GoPro [3] dataset.

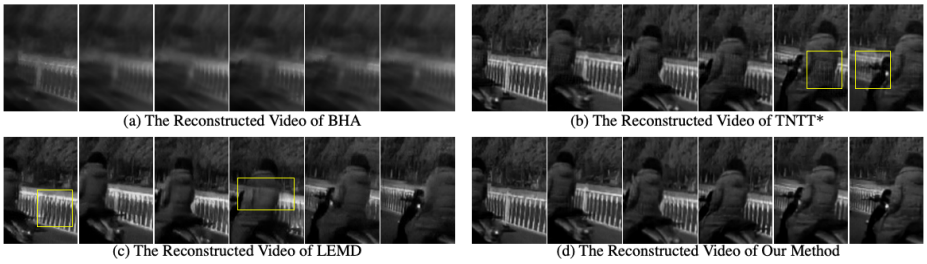


Fig. 24. Visual comparisons on video reconstruction on the synthetic subset of Blur-DVS [1].

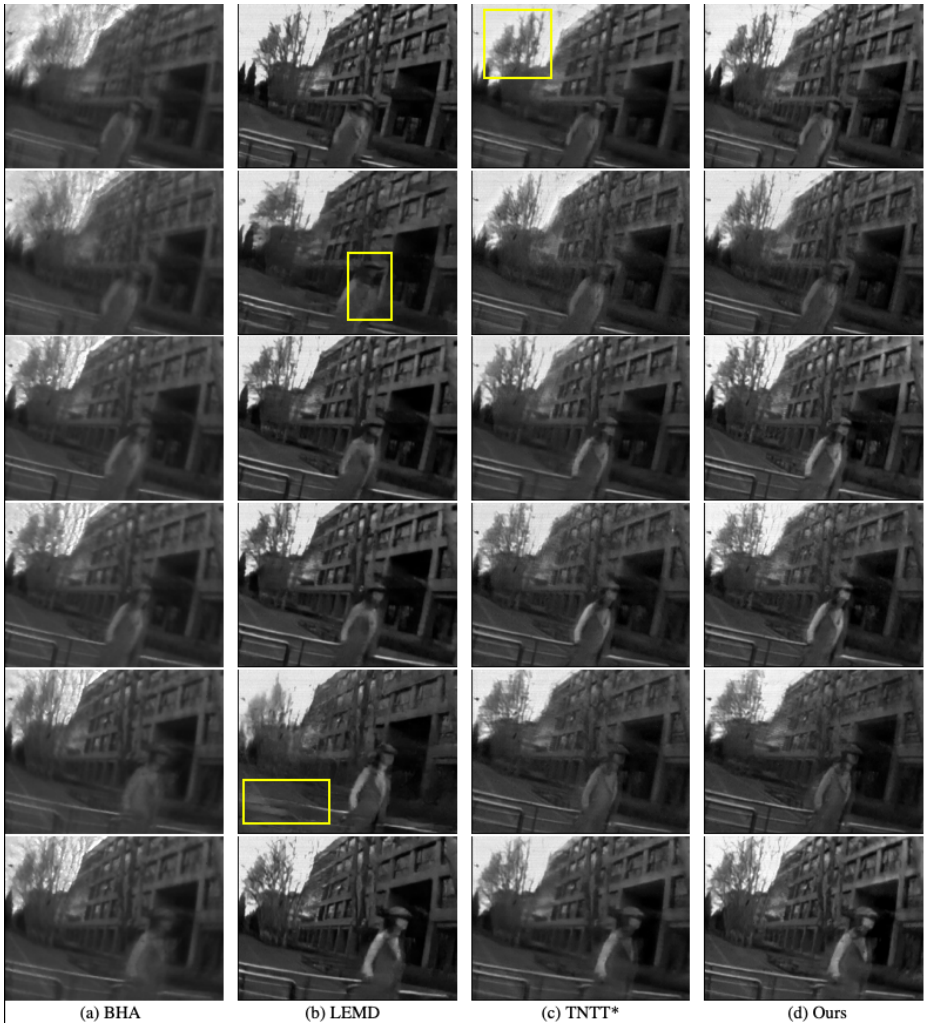


Fig. 25. Visual comparisons on video reconstruction on the real subset of Blur-DVS [1].

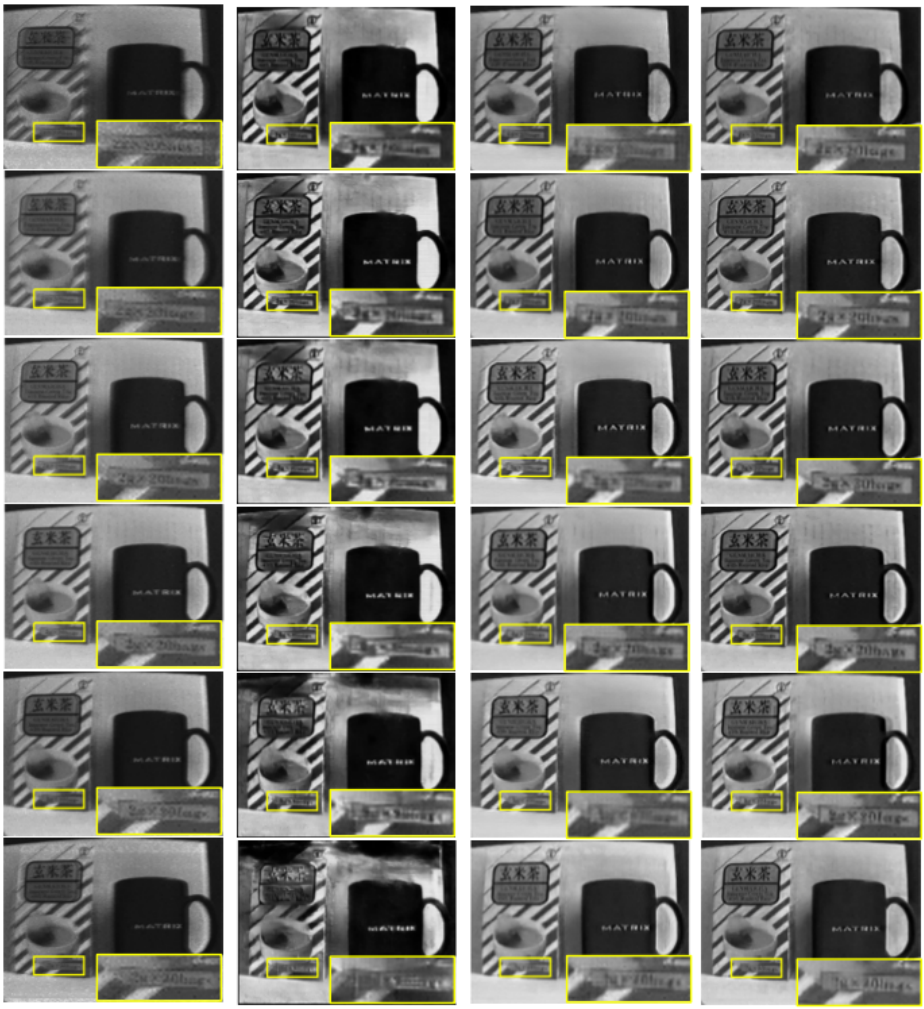


Fig. 26. Visual comparisons on video reconstruction on the real-world blurry videos [4].

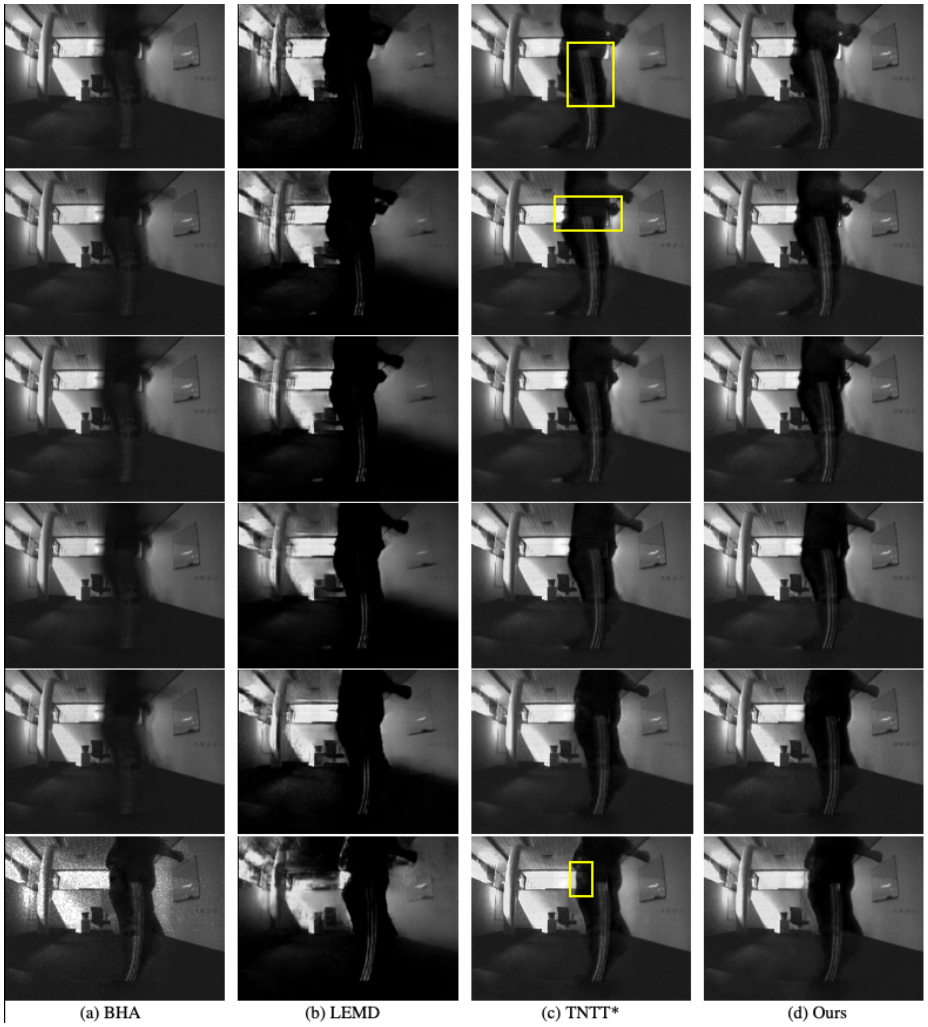
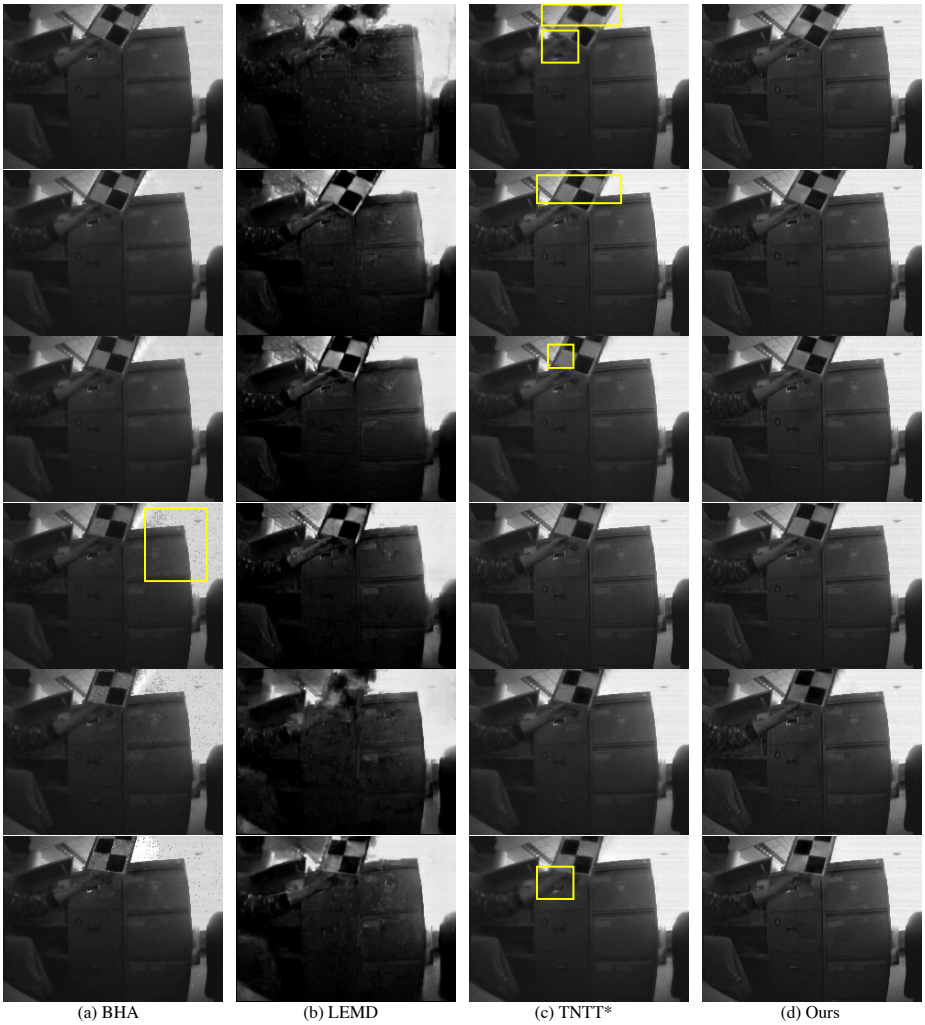
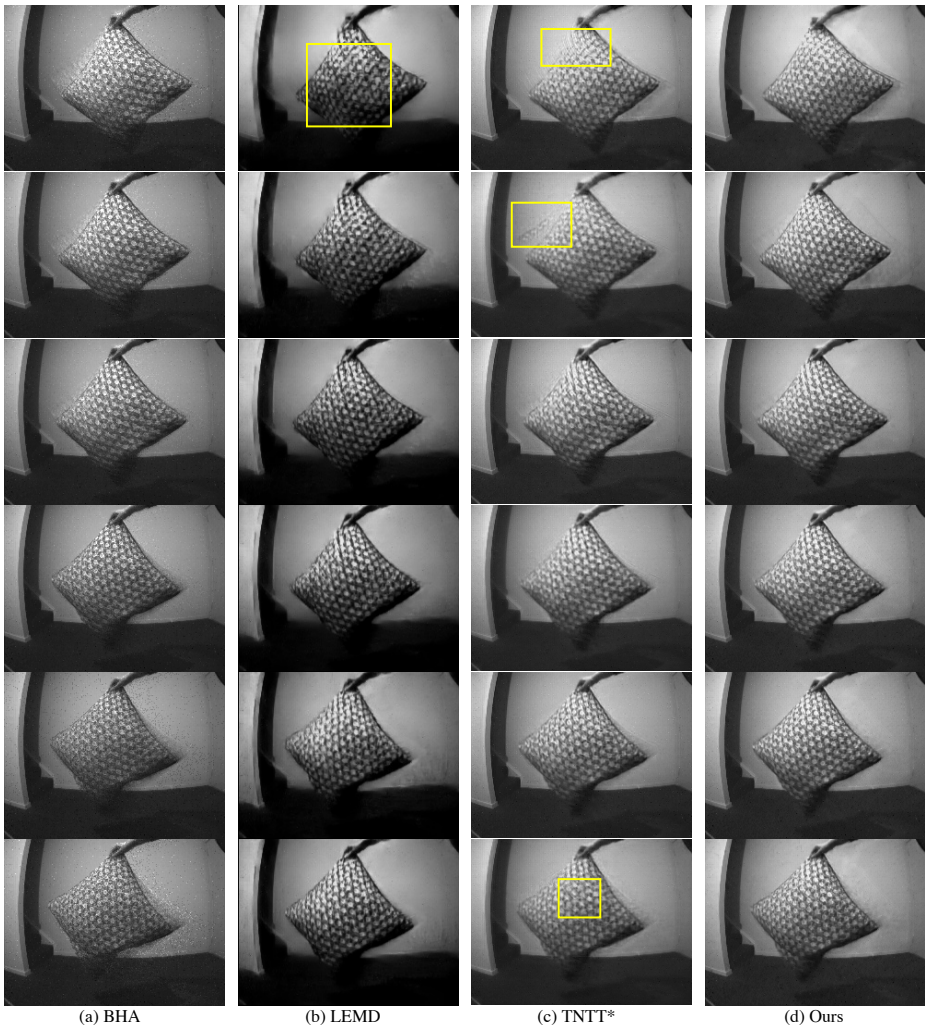


Fig. 27. Visual comparisons on video reconstruction on the real-world blurry videos [4].



(a) BHA (b) LEMD (c) TNTT* (d) Ours

Fig. 28. Visual comparisons on video reconstruction on the real-world blurry videos [4].



(a) BHA (b) LEMD (c) TNTT* (d) Ours

Fig. 29. Visual comparisons on video reconstruction on the real-world blurry videos [4].

(a) BHA (b) LEMD (c) TNTT* (d) Ours

Fig. 30. Visual comparisons on video reconstruction on the real-world blurry videos [4].

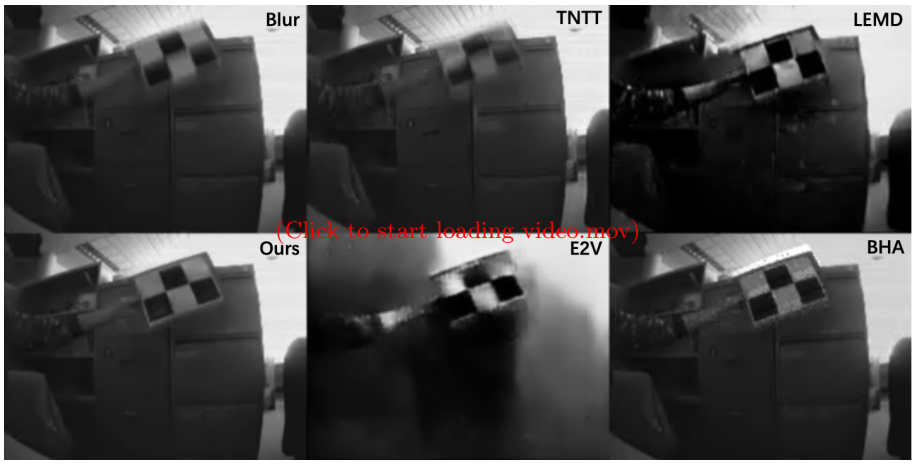


Fig. 31. Visual comparisons on video reconstruction on the real-world blurry videos [4].

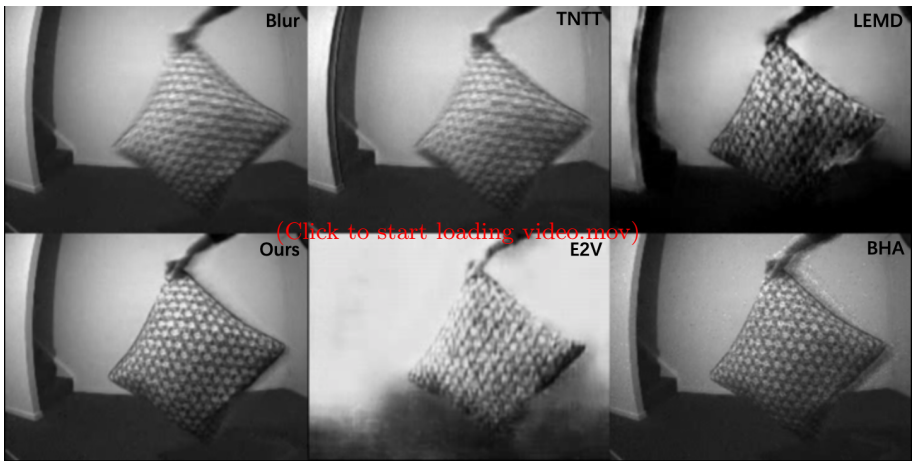


Fig. 32. Visual comparisons on video reconstruction on the real-world blurry videos [4].

References

1. Jiang, Z., Zhang, Y., Zou, D., Ren, J., Lv, J., Liu, Y.: Learning event-based motion deblurring. In: CVPR (2020) 6, 8, 9, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 28, 29
2. Jin, M., Hu, Z., Favaro, P.: Learning to extract flawless slow motion from blurry videos. In: CVPR (2019) 8, 28
3. Nah, S., Hyun Kim, T., Mu Lee, K.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: CVPR (2017) 9, 10, 11, 12, 28
4. Pan, L., Scheerlinck, C., Yu, X., Hartley, R., Liu, M., Dai, Y.: Bringing a blurry frame alive at high frame-rate with an event camera. In: CVPR (2019) 6, 8, 9, 18, 24, 25, 26, 27, 28, 30, 31, 32, 33, 34, 35
5. Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D.: Events-to-video: Bringing modern computer vision to event cameras. In: CVPR. pp. 3857–3866 (2019) 8, 9, 18
6. Scheerlinck, C., Barnes, N., Mahony, R.: Continuous-time intensity estimation using event cameras. In: ACCV (2018) 9
7. Zhou, S., Zhang, J., Pan, J., Xie, H., Zuo, W., Ren, J.: Spatio-temporal filter adaptive network for video deblurring. In: ICCV (2019) 8, 9, 18