Video Frame Interpolation via Direct Synthesis with the Event-based Reference - Supplementary Material -

Yuhan Liu[1] Yongjian Deng[1*] Hao Chen[2] Zhen Yang[1]
[1]College of Computer Science, Beijing University of Technology
[2]Key Lab of Computer Network and Information Integration, Southeast University
{liuyuhan@emails., yjdeng@, yangzhen@}bjut.edu.cn, haochen303@seu.edu.cn

Due to the limited space in the main text, we provide more details in the Supplementary Material. This supplementary material is comprised of the following sections.

Sec. 1 presents the detailed network structure of our proposed approach.

Sec. 2 includes descriptions of different datasets used for model evaluation.

Sec. 3 conducts supplementary experiments both in qualitative and quantitative views. We also attach a video clip for obtaining more intuitive comparison between SOTA methods.

1. Method Details

1.1. Reconstruction stage

We give a more detailed network structure diagram in the method section of the main text, as shown in Fig. 1.

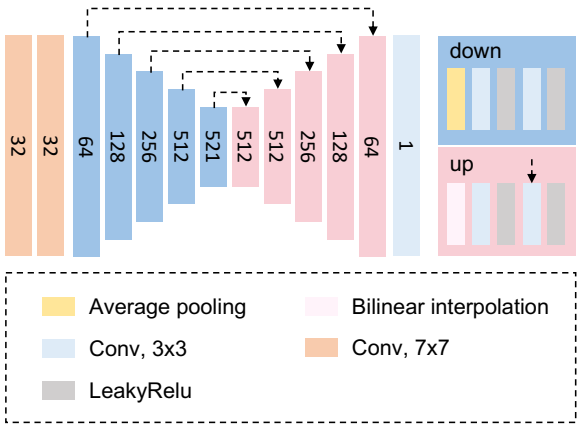

Figure 1. Detailed presentation of the reconstruction network for obtaining the event-based reference.

We adopt an U-shaped network [6] that utilizes pure events for the synthesis of the event-based reference at the interpolated frame position, circumventing the effect of poor keyframe quality on the interpolation results.

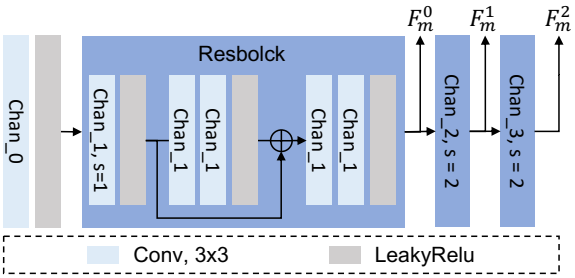

Figure 2. Detailed presentation of the feature *Encoders* adopted in the synthesis and refinement stages.

1.2. Synthesis stage

The structure of feature encoders ($\mathcal{E}_I, \mathcal{E}_E, \mathcal{E}_R$) in the synthesis stage is shown in Fig. 2, where we take $\mathcal{E}_I$ as an example. The symbol $F_m^{0,1,2}$ denotes features obtained from the network at different scales that are supposed to be fed into the E-PCD module. The $\{Chan_i | i = 0, 1, 2, 3\}$ in the figure represents output channels of the convolutional layer, and s denotes the stride. In specific, output channels are [16, 32, 64, 128], [16, 32, 64, 96] and [8, 16, 32, 64] for $\mathcal{E}_I$, $\mathcal{E}_E$ and $\mathcal{E}_R$, respectively.

Besides, we include four convolutional blocks after E-PCD modules to fuse the bidirectional interpolating features, where the first three consists of a normal 3×3 $Conv$ followed by a $LeakyRelu$ activation function with output channel as 64. The last one holds the similar architecture as the first three blocks but with output channel of 3.

1.3. Refinement stage

We describe detailed network structure of the refinement stage in this section. As illustrated in the main manuscript, the refinement stage comprises of two parts denoted as "*Three Different Encoders*" and "*Transformer Decoder*".

*: Corresponding author

Specifically, "*Three Different Encoders*" represent feature encoders that take $\{I_0, I_1, E_{0\to1}, \hat{I}_\tau\}$ as input. We formulate this process in Eq. (1).

$$\begin{aligned} F^{0,1,2}_{m,*} &= \mathcal{E}^*_I(I_m), m \in \{0,1\}, \\ G^{0,1,2}_{*} &= \mathcal{E}^*_E(E_{0\to1}), \\ V^{0,1,2}_{*} &= \mathcal{E}^*_{pred}(\hat{I}_\tau)), \end{aligned} \tag{1}$$

where the structure of these encoders ($\{\mathcal{E}^*_I, \mathcal{E}^*_E, \mathcal{E}^*_{pred}\}$) are similar to feature encoders utilized in the synthesis stage and output channels of their inside layers are set as [8, 16, 32, 64].

After obtaining encoded features from various views ($F^{0,1,2}_{m,*}, G^{0,1,2}_{*}, V^{0,1,2}_{*}$), we input them into an interactive Transformer-based decoder (Inspired by [4]). To gain feature refinement at multi-scale, we convey a multi-level Transformer-based decoding to achieve the final interpolation. We illustrate the specific structure of the decoding block at i-th level in Fig. 3 and formulate the input construction of this block in Eq. (2).

$$\begin{aligned} \mathbb{F}^i_F &= \mathcal{C}(F^i_{m,*}, G^i_*), \\ \mathbb{F}^i_R &= \mathcal{C}(V^i_*, F^i_{m,*}, G^i_*, \chi^{i-1}), \\ \mathbb{F}^i_E &= \mathcal{C}(V^i_*, G^i_*), \end{aligned} \tag{2}$$

where the input feature $\mathbb{F}_R$ at i-th level is obtained by fusing the output of previous layer (χ^{i-1}). A total of three different scales of I_τ can be obtained for training, and the generation and computation of Q, K, V follows the standard attentive learning method proposed in [2].

2. Datasets Details

Synthesis Datasets. As mentioned in the main paper, we generate events on four publicly available VFI datasets. The first dataset is Vimeo90K-Septuplet dataset [9], which contains 62450 8-frame scenes (only training scenarios) with a size of 448×256. The second dataset is Vimeo90k-Triplet dataset [9], which contains 3781 3-frame scenes (only test scenarios) with a size of 448×256. The third dataset is GoPro dataset [5], which contains 22 different scenes with a size of 1280×720. The third dataset is SNU-FILM [1], which consists of four levels representing the hardness of performing interpolation: easy, medium, hard, and extreme, representing different numbers of skipped frames from 1 to 15. It consists of 31 different scenarios.

Real Event Datasets We conduct evaluations on two real-event datasets. The first dataset is the High Quality Frames (HQF) dataset [7] of 14 different scenes captured by the DAVIS-240C event camera, which has a resolution of 240×180. The second dataset is the HS-ERGB dataset [8] of 15 different scenes captured by the Prophesee Gen4 event camera, which has a larger resolution 800×856.

Table 1. Ablation study without using grayscale ground truth on Vimeo90K-Triplet and HS-ERGB (far) datasets.

Grayscale Supervision	Vimeo90K		HS-ERGB (far)	
	PSNR	SSIM	PSNR	SSIM
Not Used	37.65	0.965	32.87	0.907
Used	**39.17**	**0.977**	**33.56**	**0.921**

3. Additional Experiments

3.1. Video Demos

We generate demo videos on the proposed real datasets, such as HS-ERGB [8] and HQF [7]. Demo videos named as Video.mp4 include the qualitative comparison with other SOTA VFI methods.

3.2. Is grayscale ground truth supervision required?

Due to space constraints, we use the grayscale ground truth values directly in the main paper for supervision and did not verify their effectiveness, which we add here, and the experimental results are shown in Tab. 1. We can observe from the table that the grayscale supervision provide a large enhancement $w.r.t$ the optimization effects.

3.3. Threshold settings for erosion operation

We visualize the erosion operation for different threshold settings, as shown in Fig. 4. We can see that the customized erosion operation is able to get rid of the effect of event noise as much as possible and preventing the significant loss of event data compared to the original erosion operation ($\delta = 1$). In this work, we select δ as 6 for all evaluations.

3.4. Stress test

An increase in the number of inter-frame skips will exacerbate the problem of motion blurring and occlusion, thus exacerbating the challenge of generating accurate results. A large number of frame skips will greatly test the robustness of the model to complex motion. With this in mind, we conducted a challenging stress test of our model using the HS-ERGB dataset containing 31 frame-hopping interpolations, and the results are shown in Fig. 5. It can be concluded that our method performs better in such extreme interpolation situations.

3.5. Additional Visual Results

3.5.1 More Visual Results on SNU-FILM dataset

In the Figs. 6 to 8, we show additional qualitative results of interpolated frames on the SNU-FILM (extreme) dataset. In the figure, we compare with state-of-the-art frame-based video frame interpolation methods, UPR-Net-L [3], event-based video frame interpolation method, TimeLens [8] and

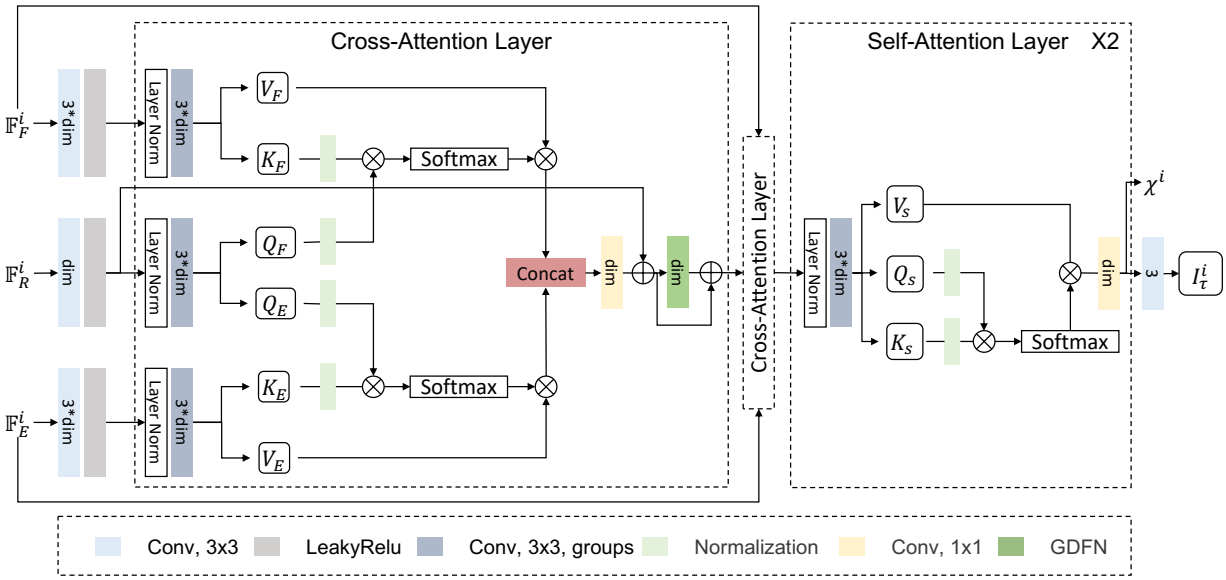

Figure 3. Detailed presentation of the $Transformer\ Decoder$ in the synthesis stage. GDFN represents Gated Deconvolutional Feed-Forward Network designed by [10]. In our works, "dim" is defined as 64 for all decoder scales (*e.g.*, $\{i = 0, 1, 2\}$)

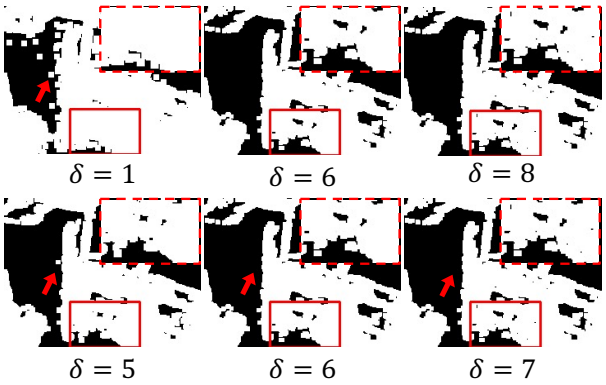

Figure 4. Visualization of event masks generated with different erosion thresholds.

CBMNet-L [4]. We confirm that our method significantly outperforms other frame and event-based video frame interpolation methods.

3.5.2 More Visual Results on HQF dataset

We aim to further validate the advantage in the generalizability of our proposed method with various scenarios in the real world. In the Figs. 9 and 10, we show more qualitative results on the HQF dataset.

3.5.3 More Visual Results on HS-ERGB dataset

In the Figs. 11 to 14, we show additional qualitative results on the HS-ERGB dataset.

References

[1] Myungsub Choi, Heewon Kim, Bohyung Han, Ning Xu, and Kyoung Mu Lee. Channel attention is all you need for video frame interpolation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10663–10671, 2020. 2

[2] 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. 2

[3] Xin Jin, Longhai Wu, Jie Chen, Youxin Chen, Jayoon Koo, and Cheul-hee Hahm. A unified pyramid recurrent network for video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1578–1587, 2023. 2

[4] Taewoo Kim, Yujeong Chae, Hyun-Kurl Jang, and Kuk-Jin Yoon. Event-based video frame interpolation with cross-modal asymmetric bidirectional motion fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18032–18042, 2023. 2, 3

[5] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017. 2

[6] 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: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 1

[7] T. Stoffregen, C. Scheerlinck, D. Scaramuzza, T. Drummond, L. Kleeman N. Barnes, and R. Mahoney. Reducing

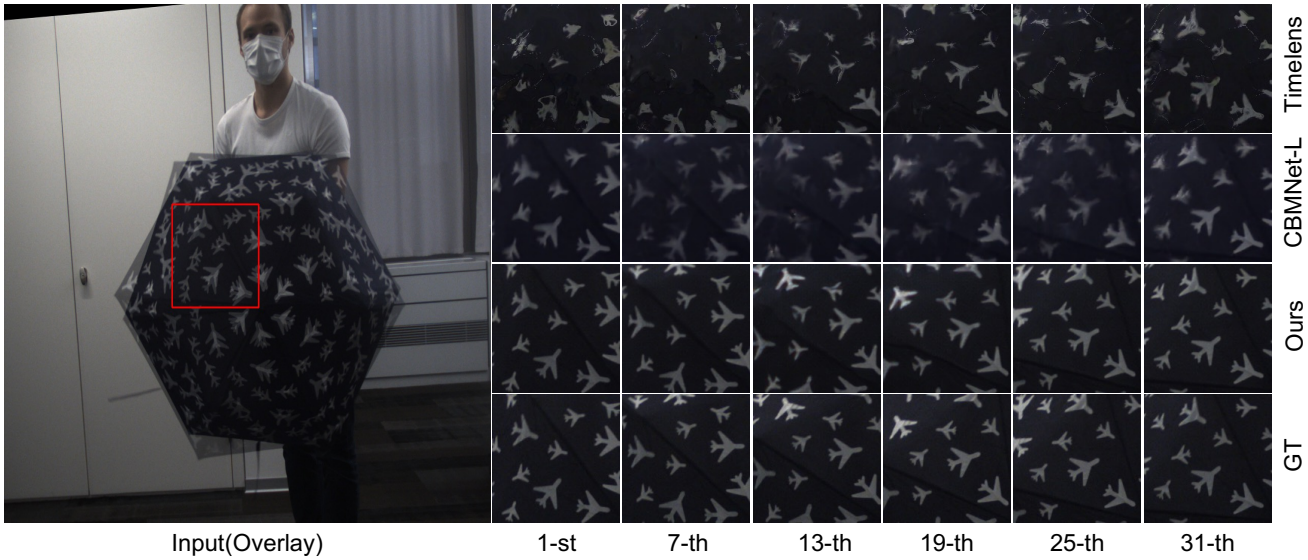

Figure 5. Qualitative comparison of the stress test, notes that the umbrella actually rotates more than one turn.

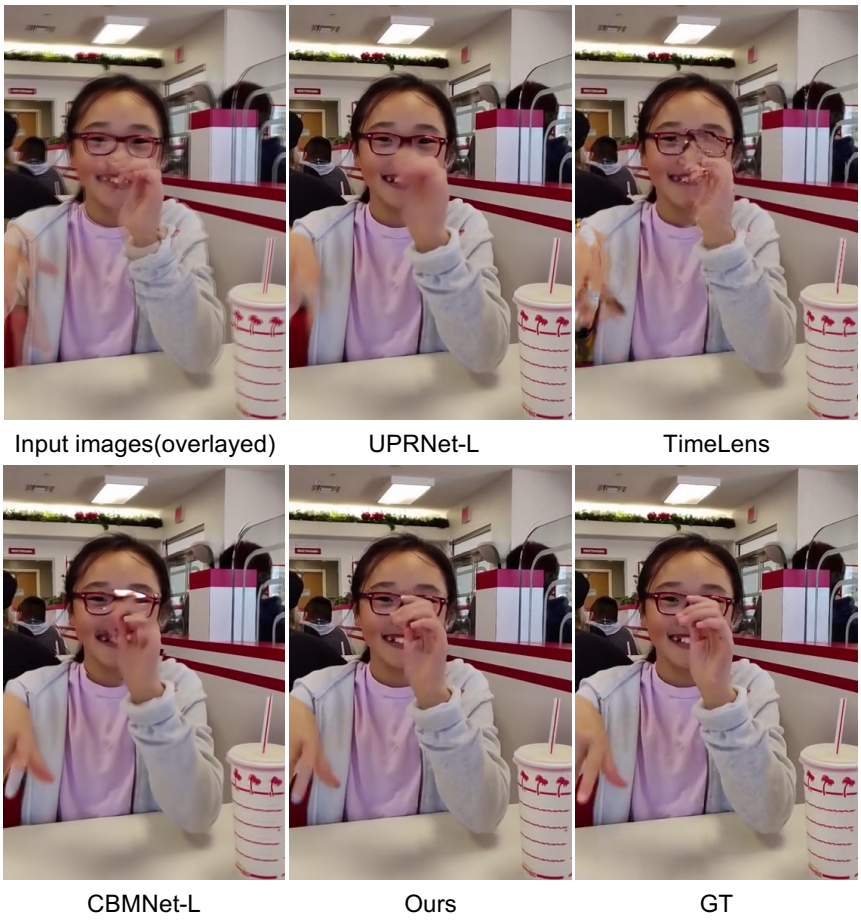

Figure 6. Visual results on the SNU-FILM dataset. (Best viewed when zoomed in.)

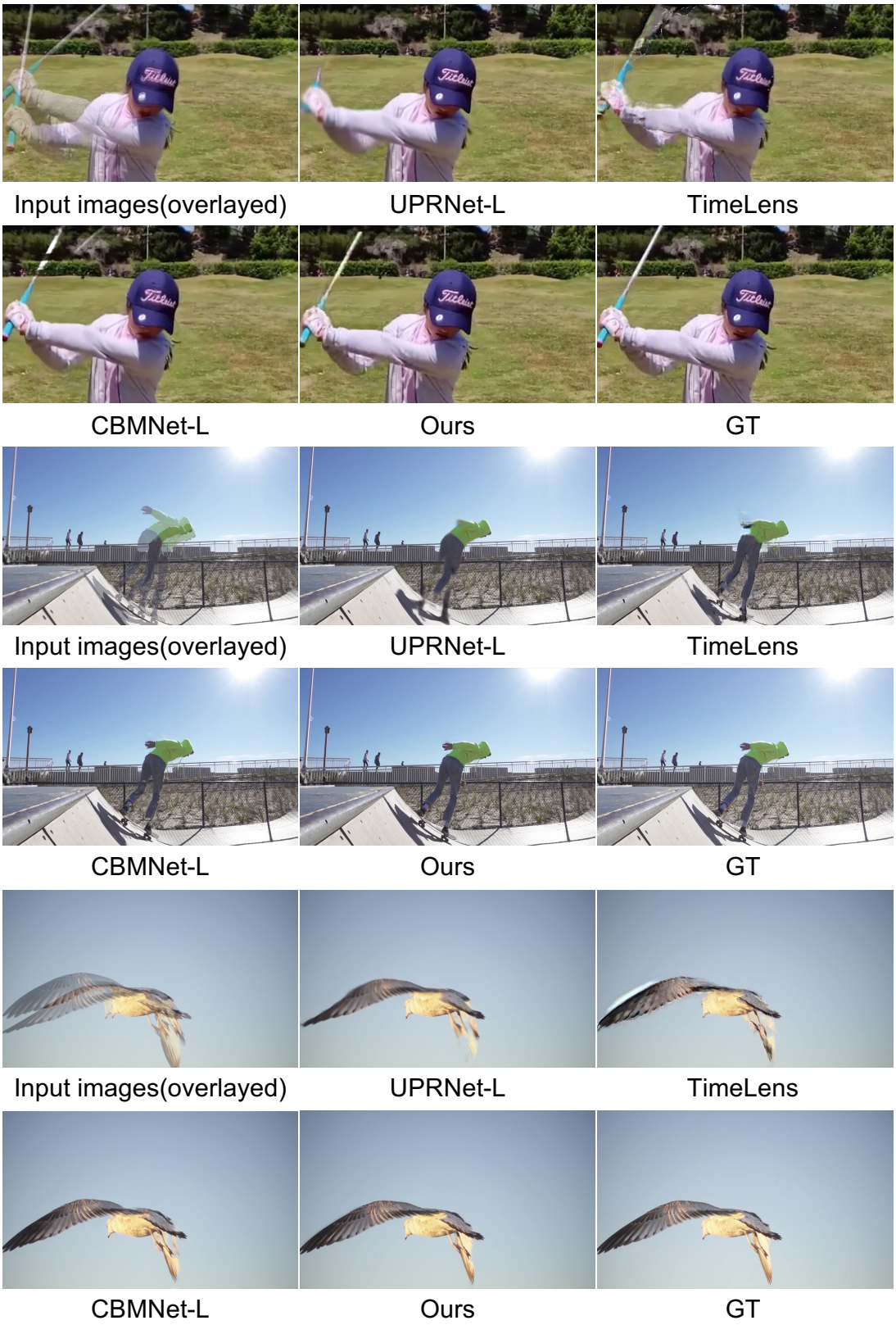

Figure 7. Visual results on the SNU-FILM dataset. (Best viewed when zoomed in.)

the sim-to-real gap for event cameras. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[8] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time lens: Event-based video frame interpolation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16155–16164, 2021. 2

[9] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision (IJCV)*, 127 (8):1106–1125, 2019. 2

[10] 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. 3

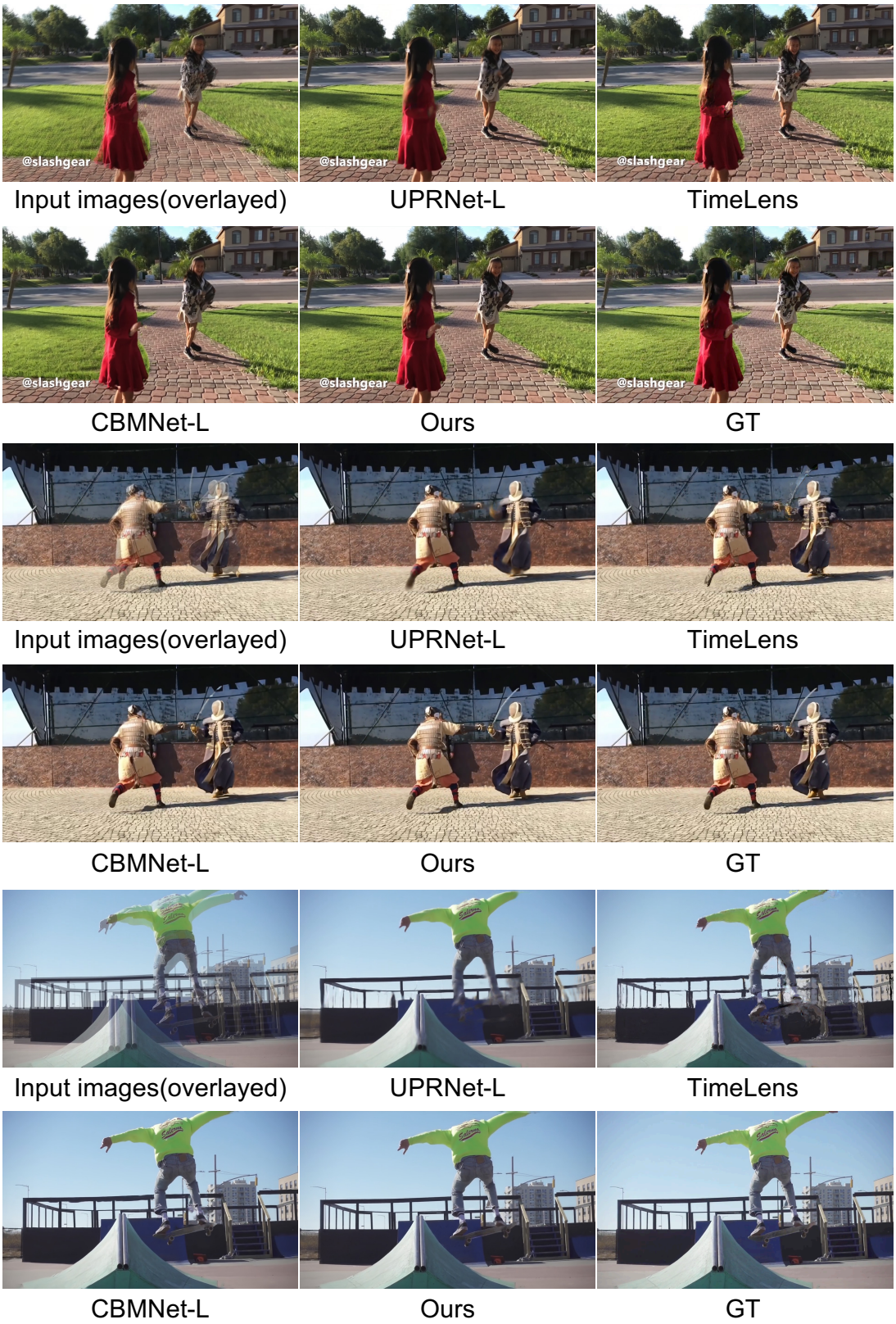

Figure 8. Visual results on the SNU-FILM dataset. (Best viewed when zoomed in.)

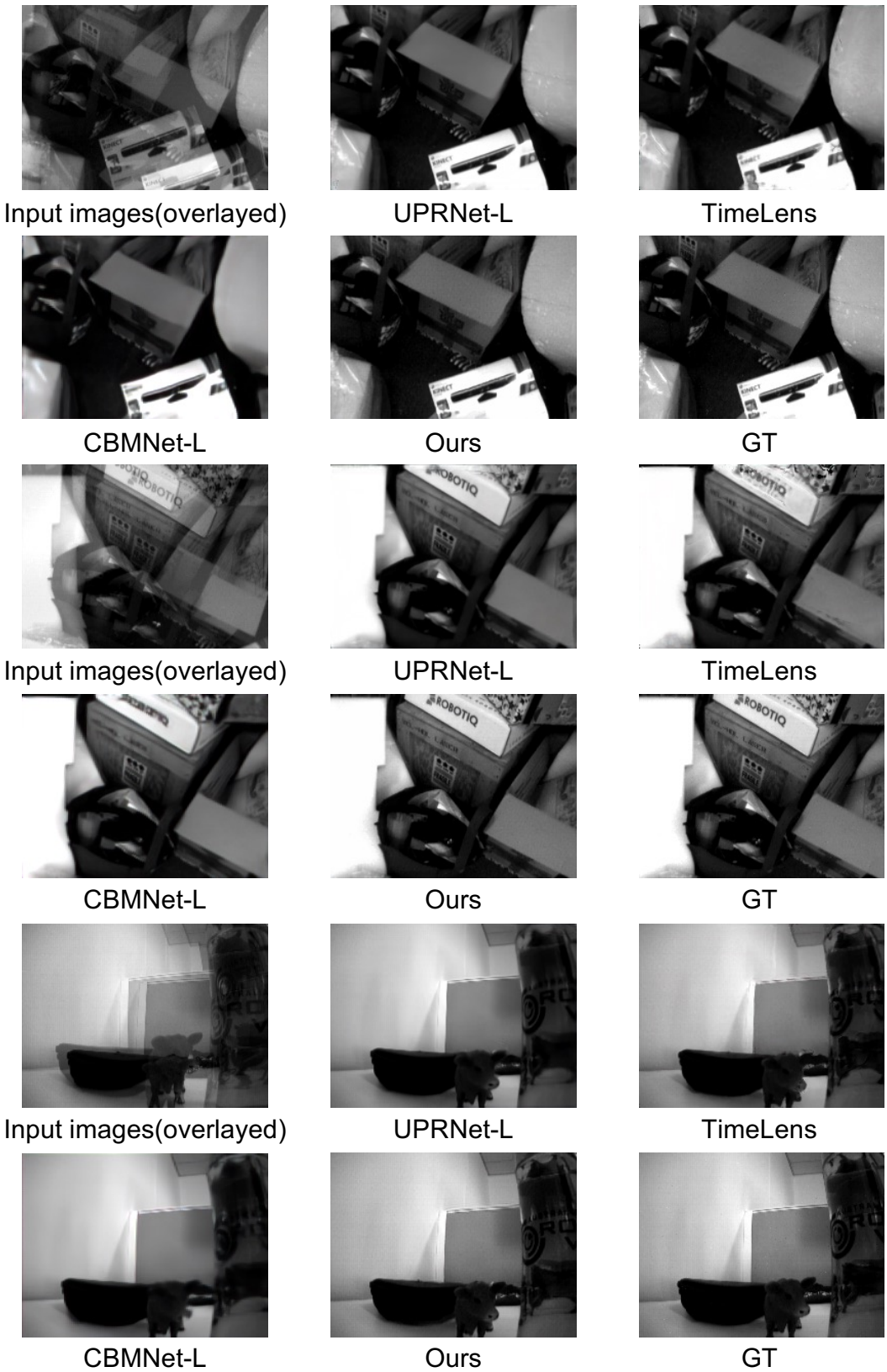

Figure 9. Visual results on the HQF dataset. (Best viewed when zoomed in.)

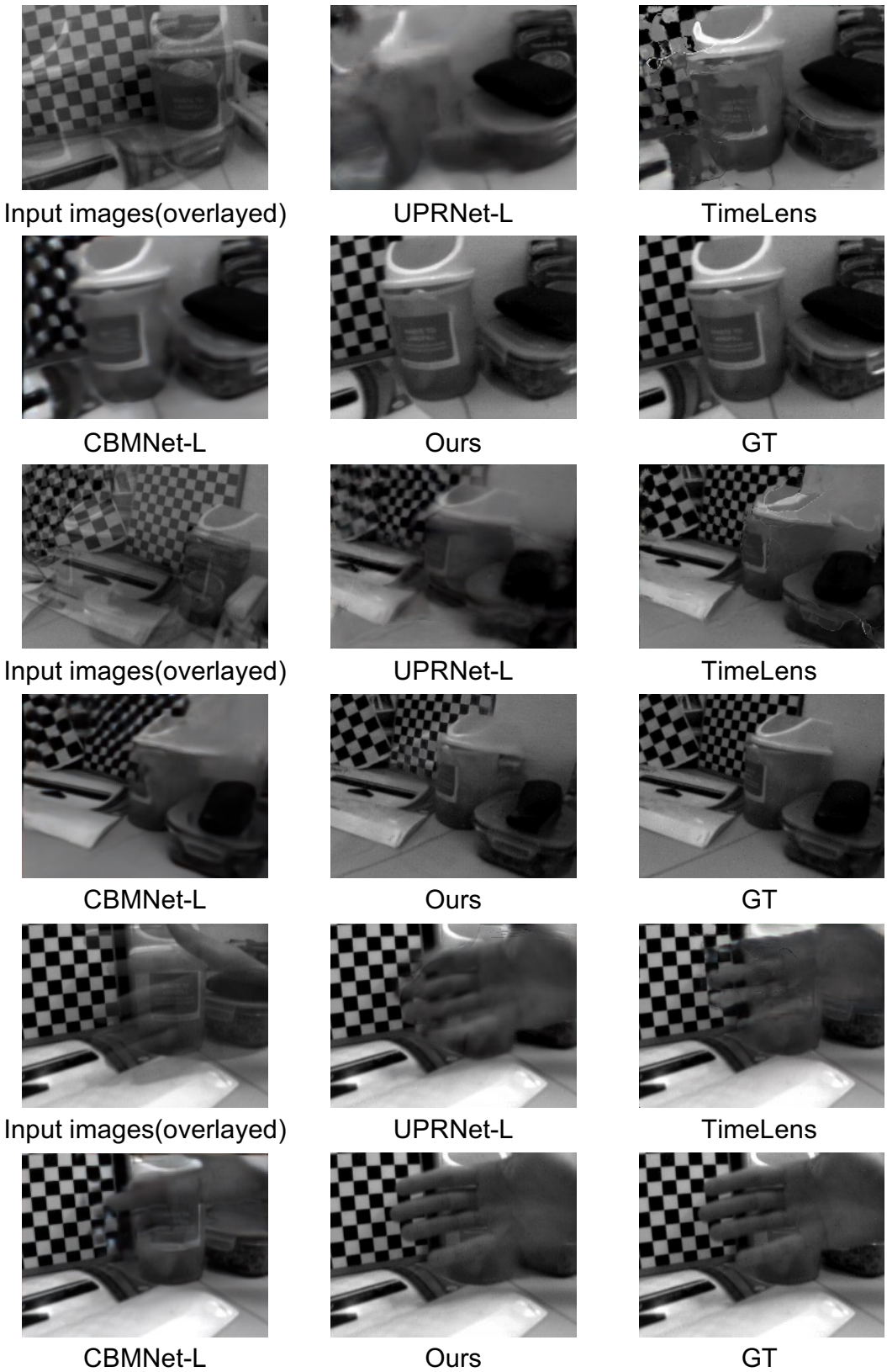

Figure 10. Visual results on the HQF dataset. (Best viewed when zoomed in.)

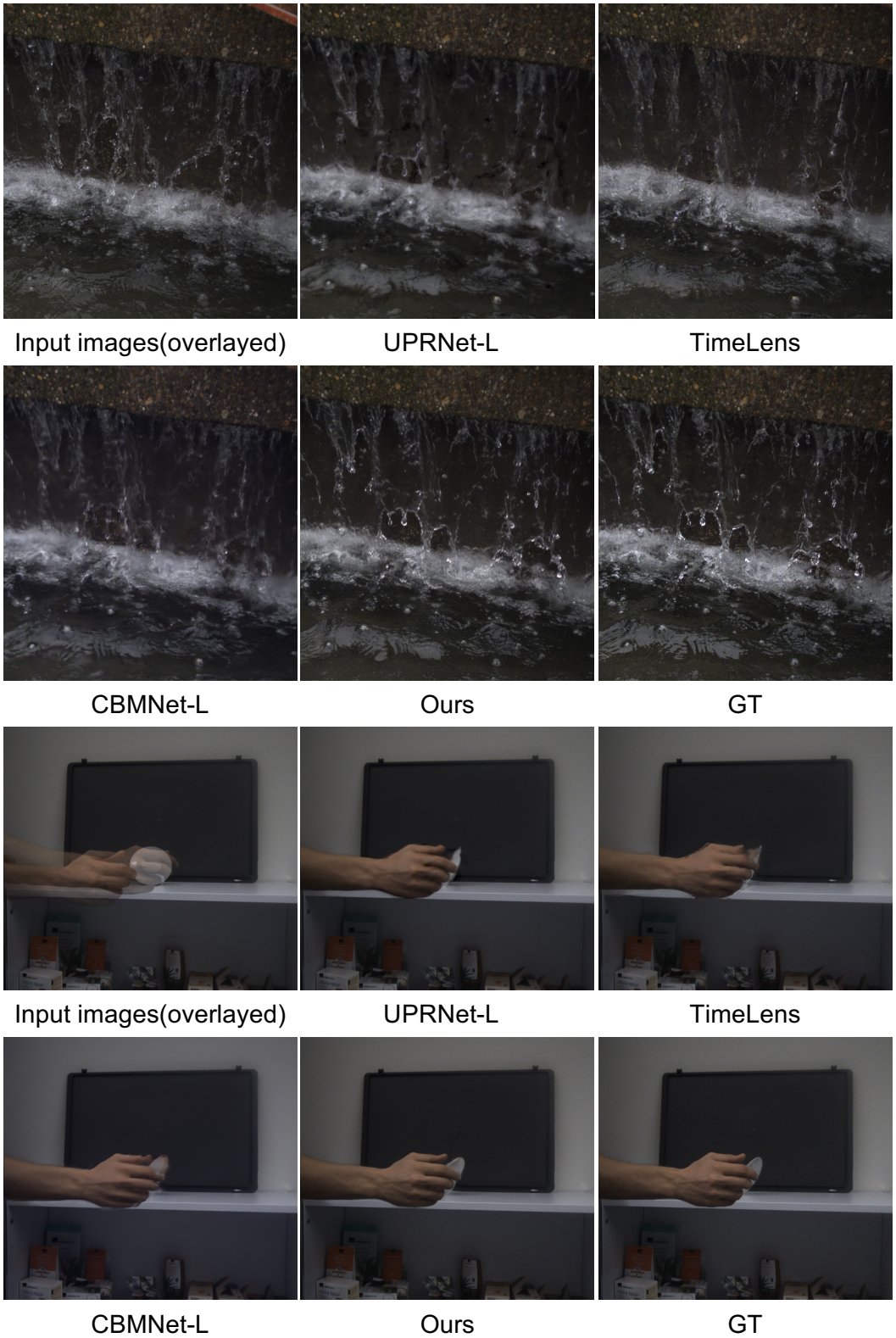

Figure 11. Visual results on the HS-ERGB dataset. (Best viewed when zoomed in.)

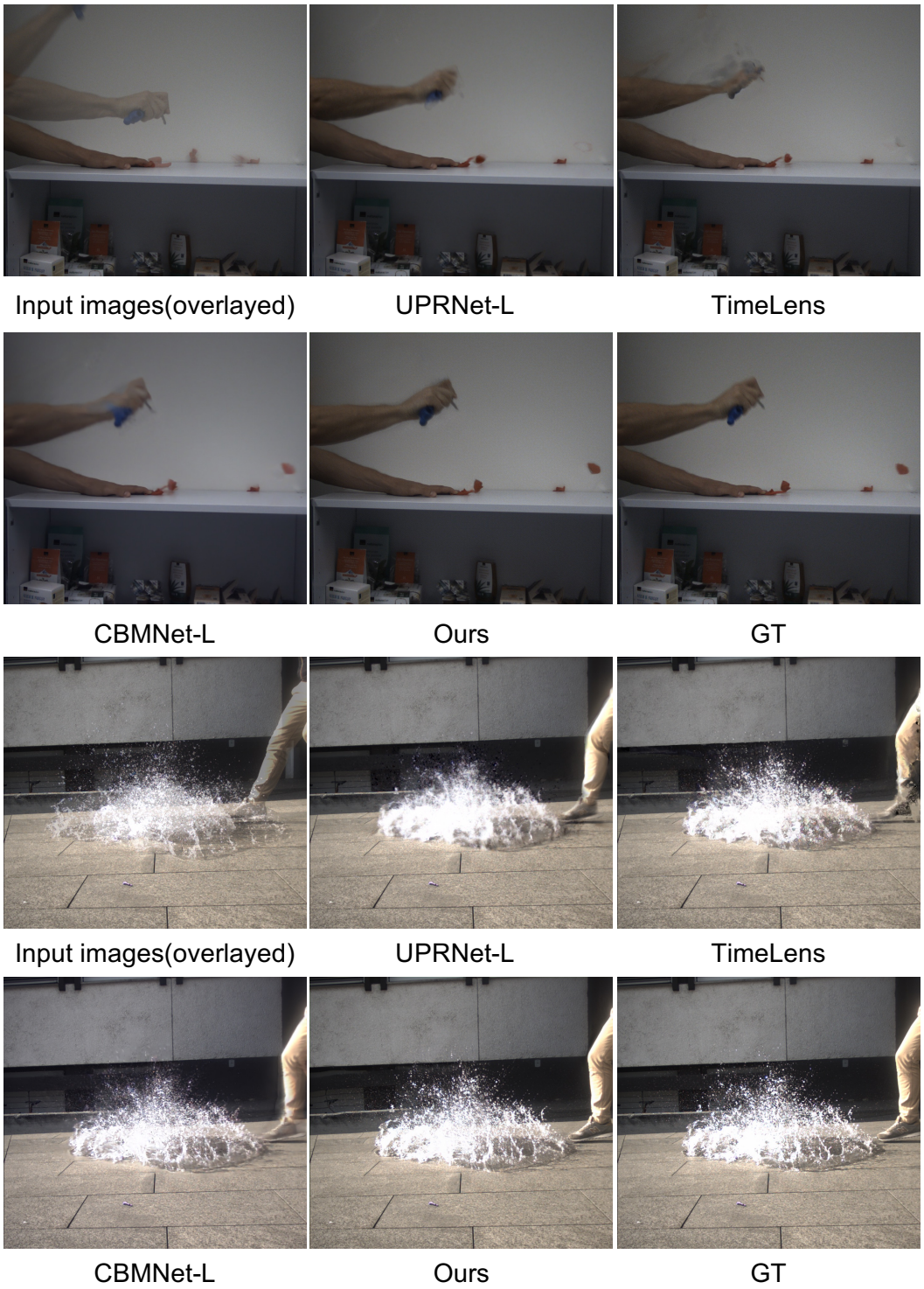

Figure 12. Visual results on the HS-ERGB dataset. (Best viewed when zoomed in.)

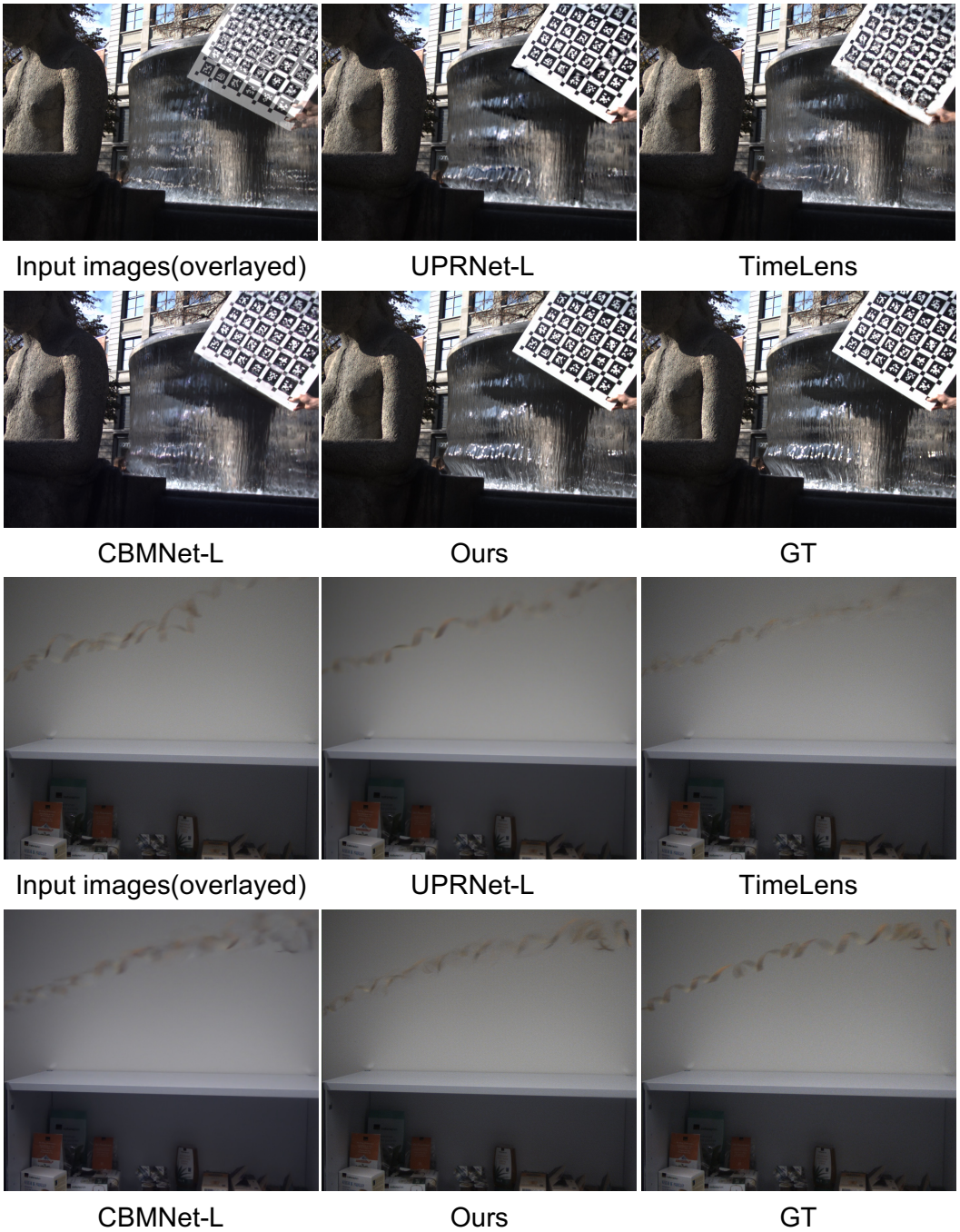

Figure 13. Visual results on the HS-ERGB dataset. (Best viewed when zoomed in.)

Figure 14. Visual results on the HS-ERGB dataset. (Best viewed when zoomed in.)