

Event-based Video Frame Interpolation with Cross-modal Asymmetric Bidirectional Motion Fields

–Supplementary materials–

Taewoo Kim, Yujeong Chae, Hyun-Kurl Jang, Kuk-Jin Yoon

Korea Advanced Institute of Science and Technology

{intelpro,yujeong,jhg0001,kjyoon}@kaist.ac.kr

Abstract

Due to the lack of the space in the main paper, we provide more details of the proposed ERF-X170FPS datasets and experimental results in the supplementary materials.

1. ERF-X170FPS Dataset

1.1. Camera Setup Details

Beam-splitter-based camera setup A beamsplitter is an optical device for splitting incident light into two beams according to a specified ratio. Therefore, a beamsplitter enables two different cameras to capture the same scenes by the split light source. The beam-splitter-based camera setup for photographing the ERF-X170FPS dataset is shown in Fig.1. For beam-splitter selection, we choose a non-polarized cube beam-splitter rather than the plate-based beam-splitter to alleviate beam-shifting issues. We select *BS-CUBE-NON-POL-VIS-50MM-TS* beam-splitter with a size of 50cm^3 , which can capture a large field of view of the scenes. The beamsplitter splits the incident light into non-polarized light in a ratio of 50:50. After that, we designed a 3D-CAD model for a rigid camera rig that can completely immobilize two cameras and a beam-splitter, as shown in Fig.3. For the RGB camera, we select *FLIR BFS-U3-16S2C-CS*. The camera can shoot videos at the resolution of 1440×1080 and up to 226FPS and support an external trigger interface. Also, we selected *EVK4 HD Prohesee Gen4.1 HD* event camera. The event camera can capture videos with a resolution of 1280×720 . We then fixed these two cameras and a beam-splitter to the designed camera rig. As a result, two cameras can receive the incoming co-axis light source at a fixed position.

Camera synchronization *In practice, we can't obtain the accurate timestamps of two cameras without interfacing with the external trigger.* For this reason, we designed a micro-controller(ATmega328) as an external trigger for hardware level synchronization of the event and RGB cam-

era. The event and RGB cameras are connected to the microcontroller through a trigger cable, as shown in Fig.1. Therefore, the generated signals of the microcontroller are simultaneously transmitted to the event camera and RGB camera, respectively. After that, we create recording software using provided C++ SDK of each camera product to control these two cameras by receiving signals from the microcontroller. Each camera receives the falling edge and the rising edge of trigger signals and performs synchronization with the period of the signal. Through this external trigger, we can control the RGB camera's frame rate and exposure time with synchronized signals. Also, we obtain accurate timestamps of the events between two consecutive standard frames. As a result, we can obtain two different modality data with precise timestamp information as the two cameras are synchronized at the hardware level.

Calibration Two cameras receive a co-axis light source due to beam-splitter camera setup. As a result, two cameras have the minimal baselines. However, they have different fields of view due to the different sensor sizes of each camera. To this end, we calibrate the event and RGB camera for spatially aligning two different modality data. For intrinsic and extrinsic calibration, we use a blinking checkerboard pattern. After the calibration process, we transform the spatial pixel position of the events using the estimated homography matrix. We then crop the standard frames whose fields of view do not overlap with the event camera. As a result, we simultaneously record spatially aligned event and frame data with a resolution of 1440×975 .

1.2. Photographing Dataset

To properly evaluate VFI performance in diverse circumstances, it is essential to shoot various scenes, such as multiple camera motions and objects. Specifically, in event-based VFI, scenes for synthesis-based interpolation and warping-based interpolation should be distributed harmoniously. As mentioned in the main paper, synthesis-based interpolation is effective in regions where motion

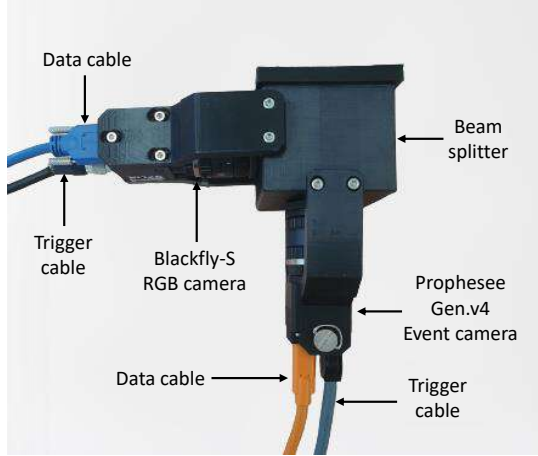


Figure 1. Our beam-splitter-based camera setup.

fields are invalid. The synthesis-based interpolation works well in situations such as flooding water, rotating objects, fire, and occlusion of scenes. However, in many typical situations, there are many areas where the motion fields are valid due to camera or object movement, except for the above cases. In the case of the previous event-based VFI datasets [13], the test scenes of the BS-ERGB dataset mainly photographed the scenes where motion fields are invalid. (e.g., flooding water tank, fire, popping eggs, fast rotating objects, thin objects **with static camera movement**). For these reasons, it is hard to evaluate the performance of motion-based frame interpolation methods in the previous event-based VFI dataset. To alleviate the deficiency, we photographed our ERF-X170FPS dataset with the following multiple categories:

- (i) We capture moving objects on a fast-moving car with diverse vehicle speeds.
- (ii) We move the camera with irregular directions and speed to shoot static scenes and moving objects. (e.g., flowers, lake, fountain, road, windmill, traffic signs, building, animals, crowds, etc.)
- (iii) We photographed the situation of the dynamic motion of people and animals with fast camera movement. (e.g., dancers, soccer, tennis player, rotating men, etc)
- (iv) We photographed the fast deformable objects (e.g., water, exploding cola, rapidly falling and rotating objects, etc.).

In the case of (i), (ii), warping-based interpolation mainly works, and synthesis-based interpolation generally works well in (iv). In the case of (iii), the interpolation result is the sum of the combination of two methods. Based on this analysis, we photograph four possible situations in balanced numbers.

1.3. ERF-X170FPS-Split

We manually selected 36 scenes for the test set of ERF-X170FPS in consideration of the degree of occlusion and

motion speed. As in the Tab.1, we have balanced the distribution of the above four situations. Compared to the BS-ERGB [13] dataset, the proposed ERF-X170FPS dataset is well distributed in four categories of situations to allow better evaluation of motion-based VFI methods. As mentioned in the main paper, we evaluated the {3, 7, 11} skips of original videos to compare the other VFI methods for diverse motion ranges. The examples of our test set of ERF-X170FPS are shown in Fig.2. We divided the remaining scenes into validation and train sets.

2. Additional Experimental Results and Details

2.1. Video Demos

We generated demo videos on the proposed ERF-X170FPS and GoPro/HQF datasets. Demo videos named as [Video_demo.mp4](#) include the qualitative comparison of videos with other SoTA VFI methods.

2.2. Quantitative evaluation results of multi-frame interpolation on GoPro [6] dataset

In the main paper, we report the evaluation results for the middle frame of the skipped video frames on the synthetic event datasets in the Tab. 2. We additionally perform comparison on the whole frames of the skipped video frames (7skips in GoPro datasets). As with the main paper, we significantly outperform frame-based and event-based video frame interpolation methods.

2.3. Datasets Details

Real Event Datasets As mentioned in the main paper, we conducted in two publicly available real-event datasets. The first dataset is High Quality Frame (HQF) [11] dataset captured by the DAVIS-240C event camera with 14 different scenes. This dataset provides the synchronized events and frames (240×180 resolution) with 14 different scenes. In addition, we conduct the experiments on the BS-ERGB [13] dataset. Following the evaluation protocol with the previous methods [2, 13, 14, 16], we evaluate whole skipped frames within {1, 3} frame skips for both datasets.

2.4. Implementation Details

We implemented our framework using PyTorch [10]. To train our networks, we use batch size 6 for the all datasets and AdamW [5] optimizer to update network weight using initial learning rate $1e^{-4}$ and decay rate 0.5. We apply random cropping to the frame and events for the same pixel position. For the quantitative evaluation, we use the standard evaluation metrics, PSNR and SSIM [15].

Table 1. The overview of test set of ERF-X170FPS.

Seq.Name	Camera settings	Explanations	Scene class
Building 01	170FPS, 990 frames	capturing with fast camera movement of building.	(ii)
Traffic load 01	170FPS, 990 frames	capturing fast car and bicycles with zig-zag camera motion.	(ii)
Fountain water pump 01	170FPS, 990 frames	capturing fountain with non-linear camera motion.	(ii)
Fountain water pump 02	170FPS, 990 frames	capturing fountain with up-down camera motion.	(ii)
Flowers 01	170FPS, 990 frames	capturing flowers with rotations.	(ii)
Geese and lake 01	170FPS, 990 frames	capturing with moving geese in the lake.	(ii)
Traffic road 01	170FPS, 990 frames	capturing fast moving cars and traffic signs.	(ii)
Dancer 01	170FPS, 990 frames	capturing fast moving dancer with non-linear camera motion.	(iii)
Dancer 02	170FPS, 990 frames	capturing dancer with close distance.	(iii)
Geese swarm 01	170FPS, 990 frames	capturing geese swarm.	(iii)
Windmill 01	170FPS, 990 frames	capturing fast rotating windmill.	(iii)
Bicycle road 01	170FPS, 990 frames	capturing bicycle road with fast up-down camera movement.	(ii)
Traffic road 02	170FPS, 990 frames	capturing fast moving car with fast camera movement.	(ii)
Soccer players 01	170FPS, 990 frames	capturing fast soccer players.	(iii)
Soccer players 02	170FPS, 990 frames	capturing fast soccer players.	(iii)
Soccer players 03	170FPS, 990 frames	capturing fast dribbling soccer players.	(iii)
Driving forest 01	170FPS, 990 frames	capturing the forest on the car.	(i)
Driving u-turn 01	170FPS, 990 frames	capturing trees with car u-turn.	(i)
Driving forest 02	170FPS, 990 frames	capturing the forest on the car.	(i)
Driving bicycle stand 01	170FPS, 990 frames	capturing the bicycle stands on fast car.	(i)
Tennis 01	170FPS, 990 frames	capturing tennis players with irregular left-right camera motion.	(iii)
Tennis 02	170FPS, 990 frames	capturing tennis players.	(iii)
Driving u-turn 02	170FPS, 990 frames	capturing the scenes on the fast u-turning car.	(i)
Driving urban 01	170FPS, 990 frames	capturing the urban scenes on the fast moving car.	(i)
Driving left-turn 01	170FPS, 990 frames	capturing with a left-turn on an intersection and shooting cars and trees.	(i)
Driving bridge 01	170FPS, 990 frames	capturing the bridge on the fast moving car.	(i)
Driving department store 01	170FPS, 990 frames	capturing the department store on the fast moving car.	(i)
Driving road 01	170FPS, 990 frames	capturing the trees and national police agency on the fast moving car.	(i)
Falling pop-corn 01	170FPS, 990 frames	capturing falling pop-corn.	(iv)
Climbing people 01	170FPS, 990 frames	capturing person quickly climbing stairs.	(iii)
Fast rotating man 01	170FPS, 990 frames	capturing fast rotating man.	(iii)
Exploding cola 01	170FPS, 990 frames	capturing exploding cola with mentos.	(iv)
Exploding cola 02	170FPS, 990 frames	capturing exploding cola with mentos.	(iv)
lakelet 01	170FPS, 990 frames	capturing lakelet with non-linear camera motion.	(iv)
Waterfall 01	170FPS, 990 frames	capturing water fall in the lakelet.	(iv)
Handwash 01	170FPS, 990 frames	capturing a person washing his hands.	(iv)

2.5. Qualitative comparison of the warped frame on GoPro [6] dataset

Due to the lack of the main paper, we only report the quantitative results of the warped frame of inter-frame motion fields in the main paper. In addition to the quantitative results, we perform qualitative comparison of estimated inter-frame motion fields in the Fig.4 As shown in the figure, our EIF-BiOFNet more reliably estimate bidirectional inter-frame motion fields than state-of-the-art inter-frame motion field estimation methods [3, 4, 8, 9, 14].

2.6. How does EIF-BiOFNet operate w/o events?

If there is no motion (with no events), the anchor and the boundary frames are the same, and the OF comes out as nearly zero. The other case is that relative motion exists, but events are unavailable (anchor and boundary frames differ). For the second case, we could only train the I-BiOFNet, and the ablation results are shown in Tab. 3. We can see that the anchor feature synthesis may not perform well compared to using events. However, even without events, ours shows

comparable performance to that of image-based VFI SoTA method ABME [9]

2.7. Implementation details of [14]

For finetuning [14] on the ERF-X170FPS datasets, we followed the original paper’s approach by training each stage individually and adhering to the original training strategy. Tab. 4 demonstrates that the finetuned model outperforms the official pretrained model.

In the case of TimeLens-Flow, the pretrained model is trained on not only synthetic event datasets but also real-world event datasets. Therefore, the performance will be degraded if we directly apply the pre-trained model to the synthetic datasets. For a fair comparison with ours, we re-trained the TimeLens-flow only on the GoPro dataset.

2.8. Additional Visual Results

2.8.1 More Visual Results on ERF-X170FPS dataset

In Fig. 6~Fig. 10, we show more qualitative results of interpolated frames on the ERF-X170FPS dataset. In the figure,

Table 2. Quantitative evaluation of multi-frame interpolation (whole skipped frames of 7skips) on the GoPro [6] dataset.

Methods	GoPro								
	SuperSloMo [4]	SepConv [7]	DAIN [1]	BMBC [8]	TimeLens [†] [14]	TimeReplayer [†] [2]	A ² OF [†] [16]	Ours	Ours-Large
PSNR	28.95	29.13	28.81	29.08	34.81	34.02	36.61	<u>37.77</u>	38.03
SSIM	0.876	0.876	0.876	0.875	0.959	0.960	0.971	<u>0.974</u>	0.975

Table 3. Ablation study of inter-frame motion fields without events on the GoPro datasets.

	ABME [30]	I-BiOF (w/o ev.)	I-BiOF (w/ ev.)	E-BiOF	EIF-BiOF
PSNR	22.1	21.9	28.9	27.8	30.4

Table 4. Quantitative evaluation results of TimeLens [14] on the ERF-X170FPS dataset. [14]-P and [14]-F represent the pre-trained and finetuned model of TimeLens [14], respectively.

	3skips		7skips		11skips		Avg.	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
[14]-P	23.08	0.724	20.76	0.666	19.44	0.633	21.09	0.674
[14]-F	25.34	0.807	21.99	0.729	20.18	0.685	22.50	0.740

we compare with state-of-the-art frame-based video frame interpolation methods, ABME [9], RIFE [3], event-based video frame interpolation method, TimeLens [14]. We confirm that our method significantly outperforms other frame- and event-based video frame interpolation methods.

2.8.2 More Visual Results on GoPro [6] dataset

In the Fig. 11, we show more qualitative results on the GoPro dataset.

2.8.3 More Visual Results on Adobe240fps [12] dataset

In the Fig. 12, we show more qualitative results on the Adobe240fps dataset.

References

- [1] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 4
- [2] Weihua He, Kaichao You, Zhendong Qiao, Xu Jia, Ziyang Zhang, Wenhui Wang, Huchuan Lu, Yaoyuan Wang, and Jianxing Liao. Timereplayer: Unlocking the potential of event cameras for video interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17804–17813, 2022. 2, 4
- [3] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 3, 4, 8
- [4] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slo-mo: High quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9000–9008, 2018. 3, 4, 8
- [5] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 2
- [6] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3883–3891, 2017. 2, 3, 4
- [7] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive convolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 4
- [8] Junheum Park, Keunsoo Ko, Chul Lee, and Chang-Su Kim. Bmbc: Bilateral motion estimation with bilateral cost volume for video interpolation. In *European Conference on Computer Vision*, 2020. 3, 4, 8
- [9] Junheum Park, Chul Lee, and Chang-Su Kim. Asymmetric bilateral motion estimation for video frame interpolation. In *International Conference on Computer Vision*, 2021. 3, 4, 8
- [10] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 2
- [11] Timo Stoffregen, Cedric Scheerlinck, Davide Scaramuzza, Tom Drummond, Nick Barnes, Lindsay Kleeman, and Robert Mahony. Reducing the sim-to-real gap for event cameras. In *European Conference on Computer Vision*, pages 534–549. Springer, 2020. 2
- [12] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1279–1288, 2017. 4
- [13] Stepan Tulyakov, Alfredo Bochicchio, Daniel Gehrig, Stamatios Georgoulis, Yuanyou Li, and Davide Scaramuzza. Time Lens++: Event-based frame interpolation with non-linear parametric flow and multi-scale fusion. *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2
- [14] 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, 3, 4, 8
- [15] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 2

- [16] Song Wu, Kaichao You, Weihua He, Chen Yang, Yang Tian, Yaoyuan Wang, Ziyang Zhang, and Jianxing Liao. Video interpolation by event-driven anisotropic adjustment of optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2, 4



Figure 2. The examples of our ERF-X170FPS test dataset. Our dataset contains diverse scenes and motion speed at 170FPS, which consists of HR video frames and temporally synchronized HR event data.

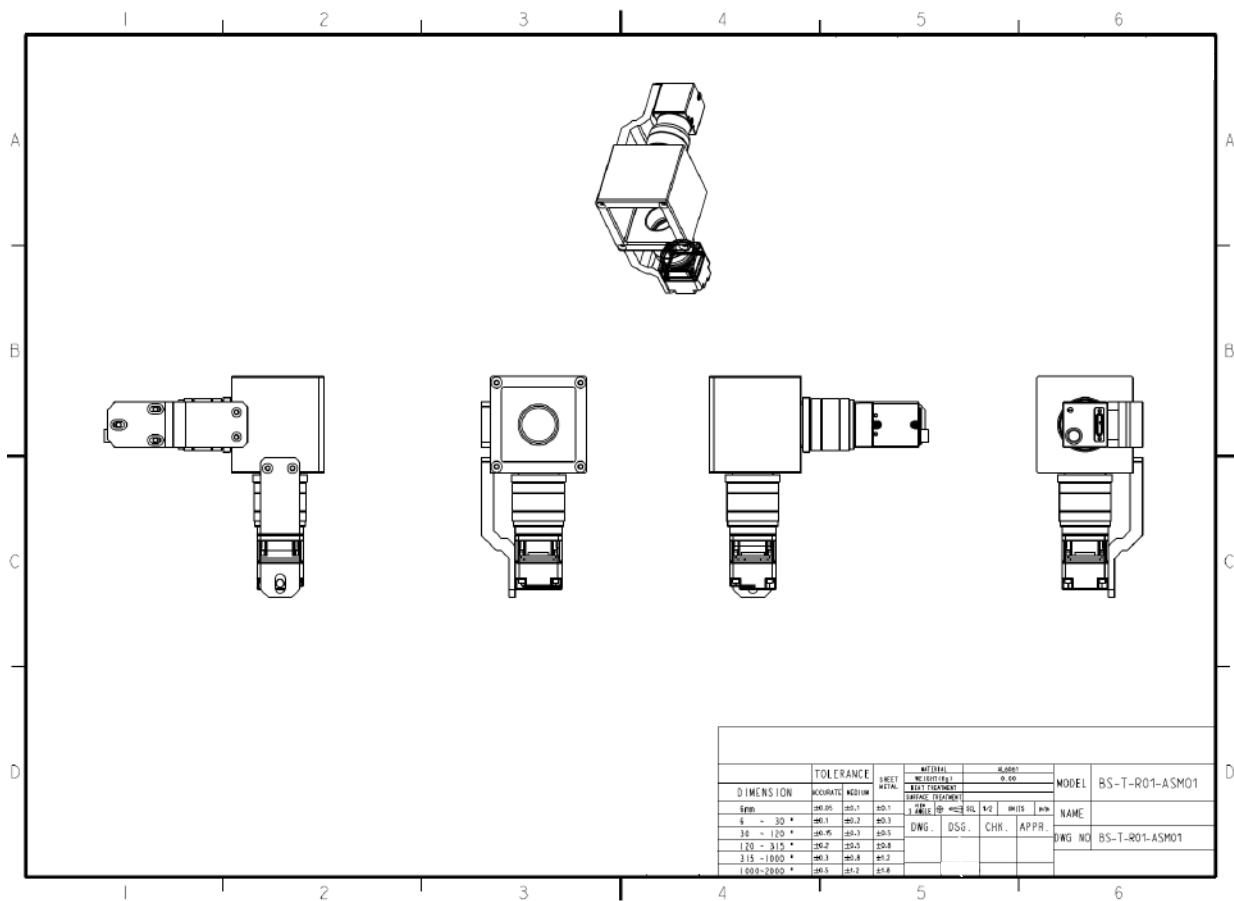


Figure 3. Beam-splitter-based camera rig 3D CAD drawing.



Figure 4. The qualitative comparison on the warped frame of estimated inter-frame motion fields on GoPro dataset. In order, (a) GT frame, (b) SuperSloMo [4] (c) BMBC [8] (d) ABME [9] (e) RIFE [3] (f) TimeLens [14] (g) Ours. As in the results, we confirm that our method produces more accurate warped frames than state-of-the-art inter-frame motion fields estimation methods. **Please zoom for better visualization.**

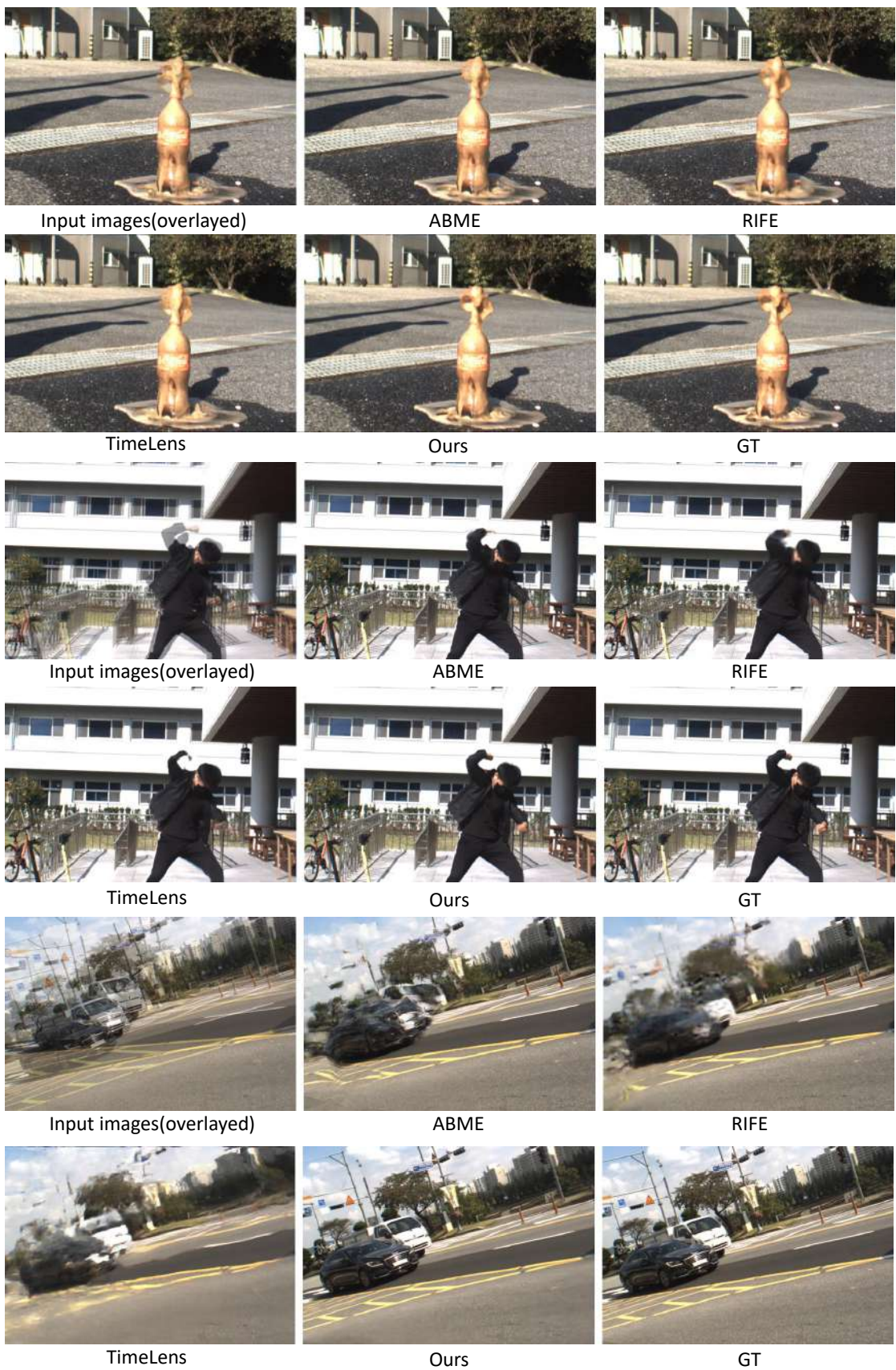


Figure 5. Visual results on the ERF-X170FPS dataset. (Best viewed when zoomed in.)

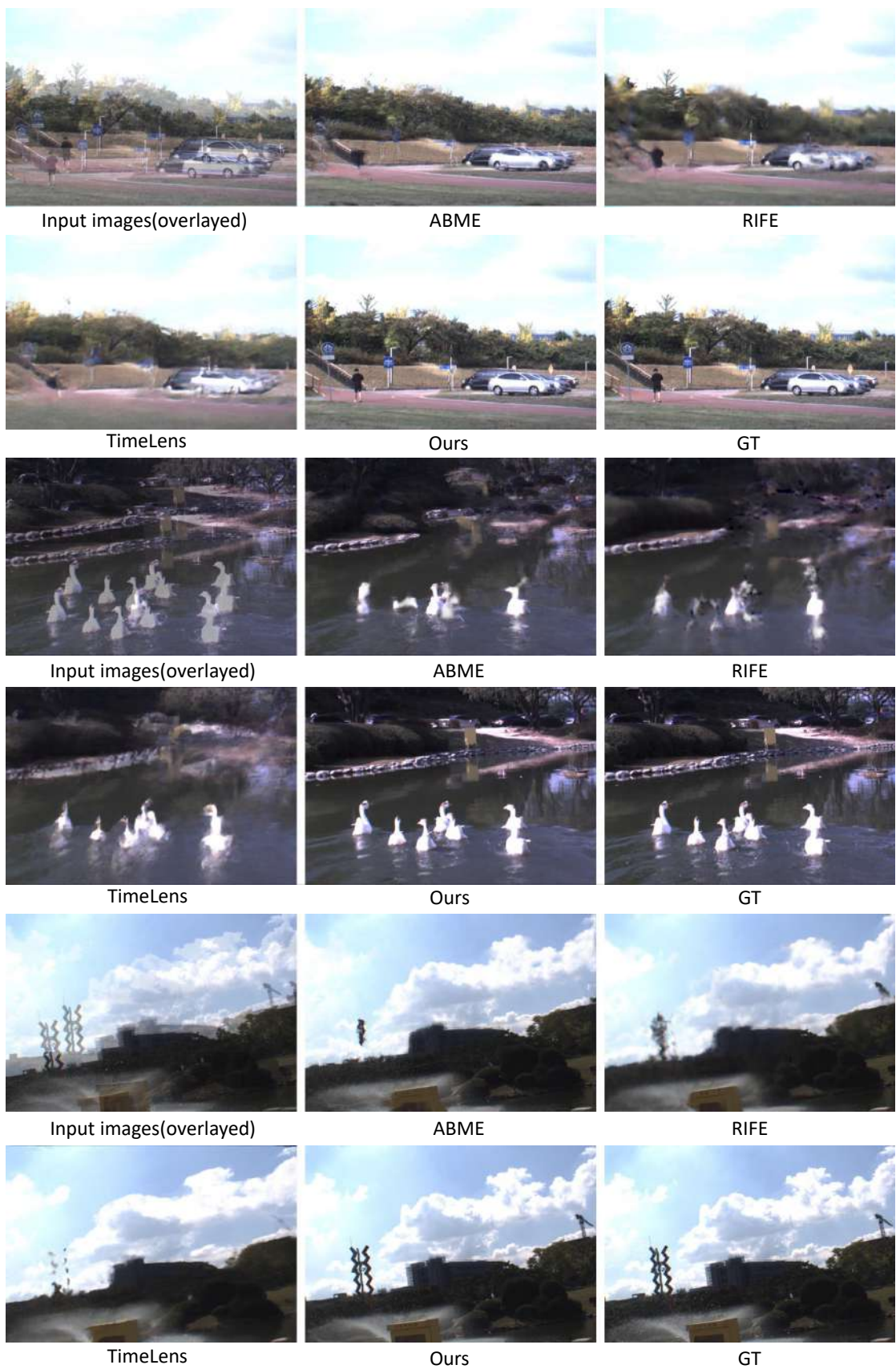


Figure 6. Visual results on the ERF-X170FPS dataset. (Best viewed when zoomed in.)

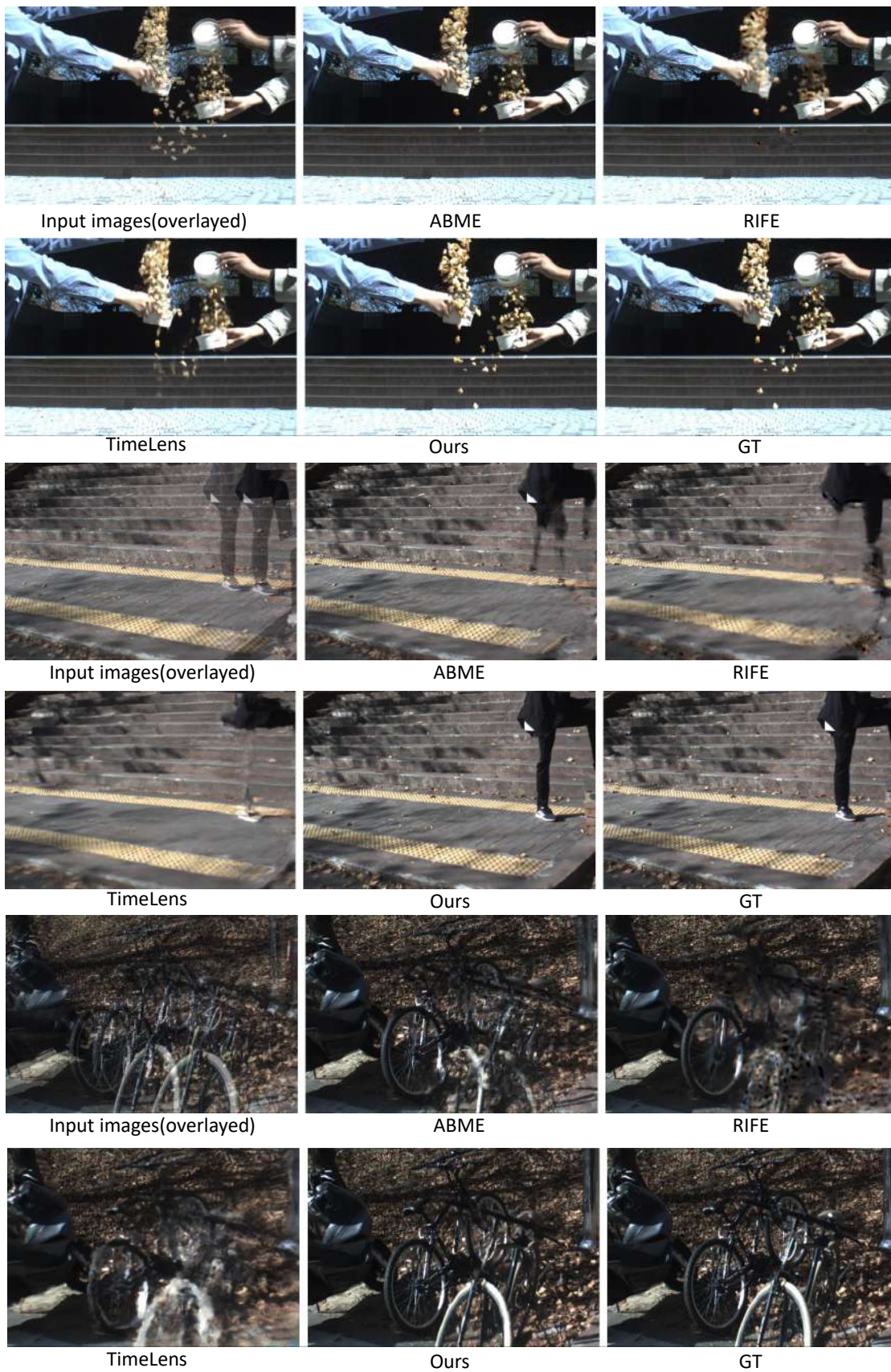


Figure 7. Visual results on the ERF-X170FPS dataset. (Best viewed when zoomed in.)

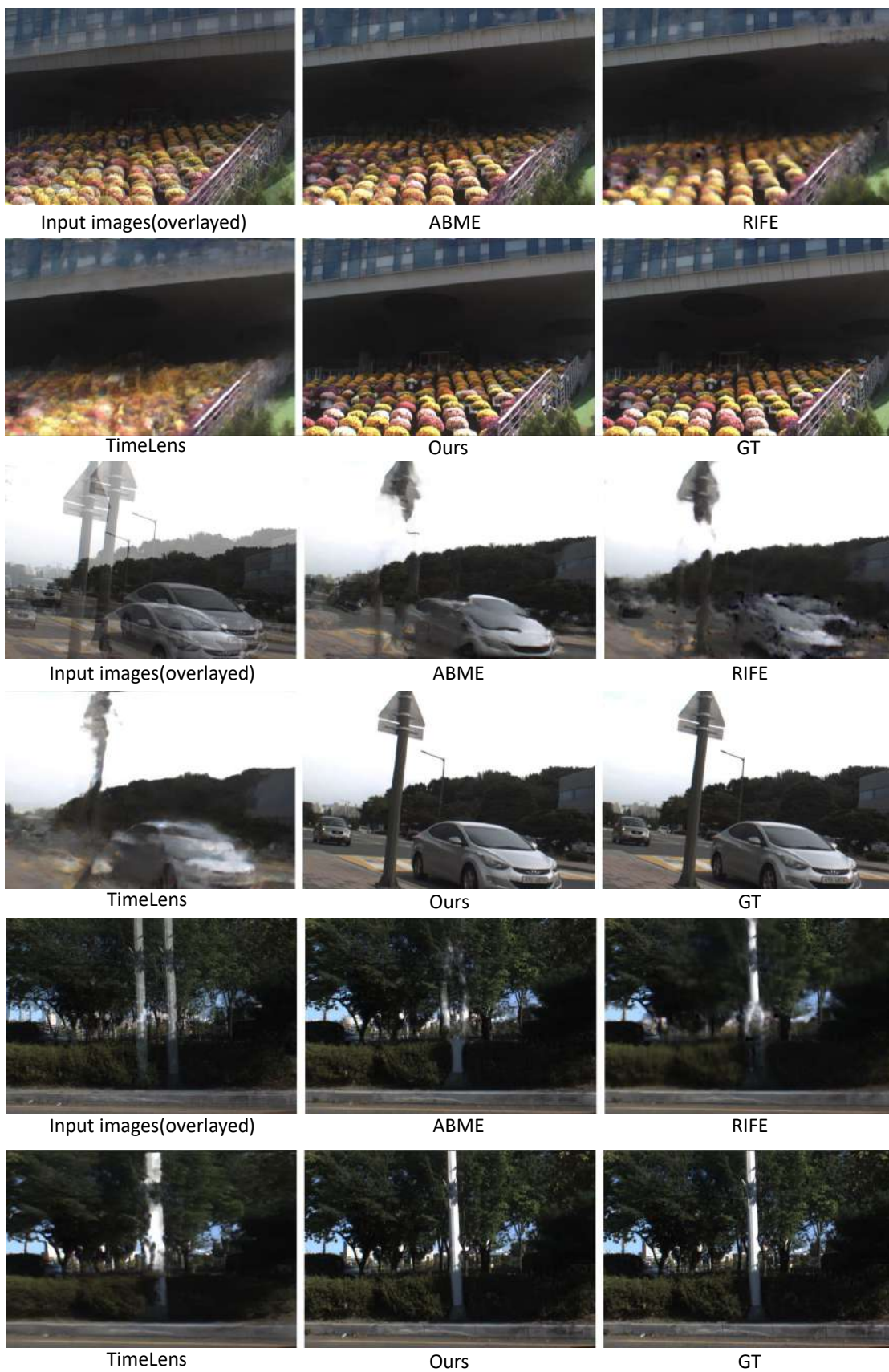


Figure 8. Visual results on the ERF-X170FPS dataset. (Best viewed when zoomed in.)



Figure 9. Visual results on the ERF-X170FPS dataset. (Best viewed when zoomed in.)



Figure 10. Visual results on the ERF-X170FPS dataset. (Best viewed when zoomed in.)

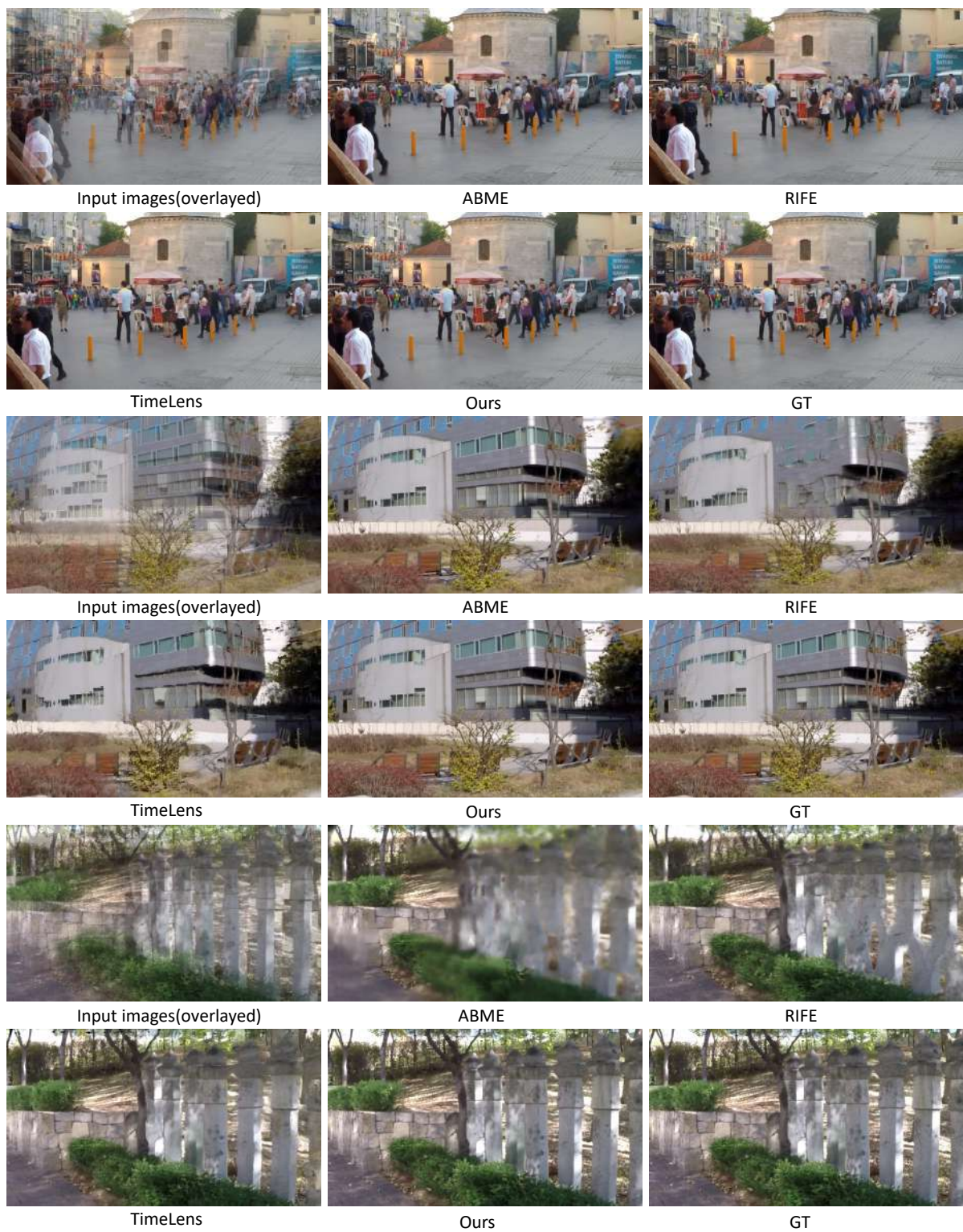


Figure 11. Visual results on the GoPro dataset. (Best viewed when zoomed in.)

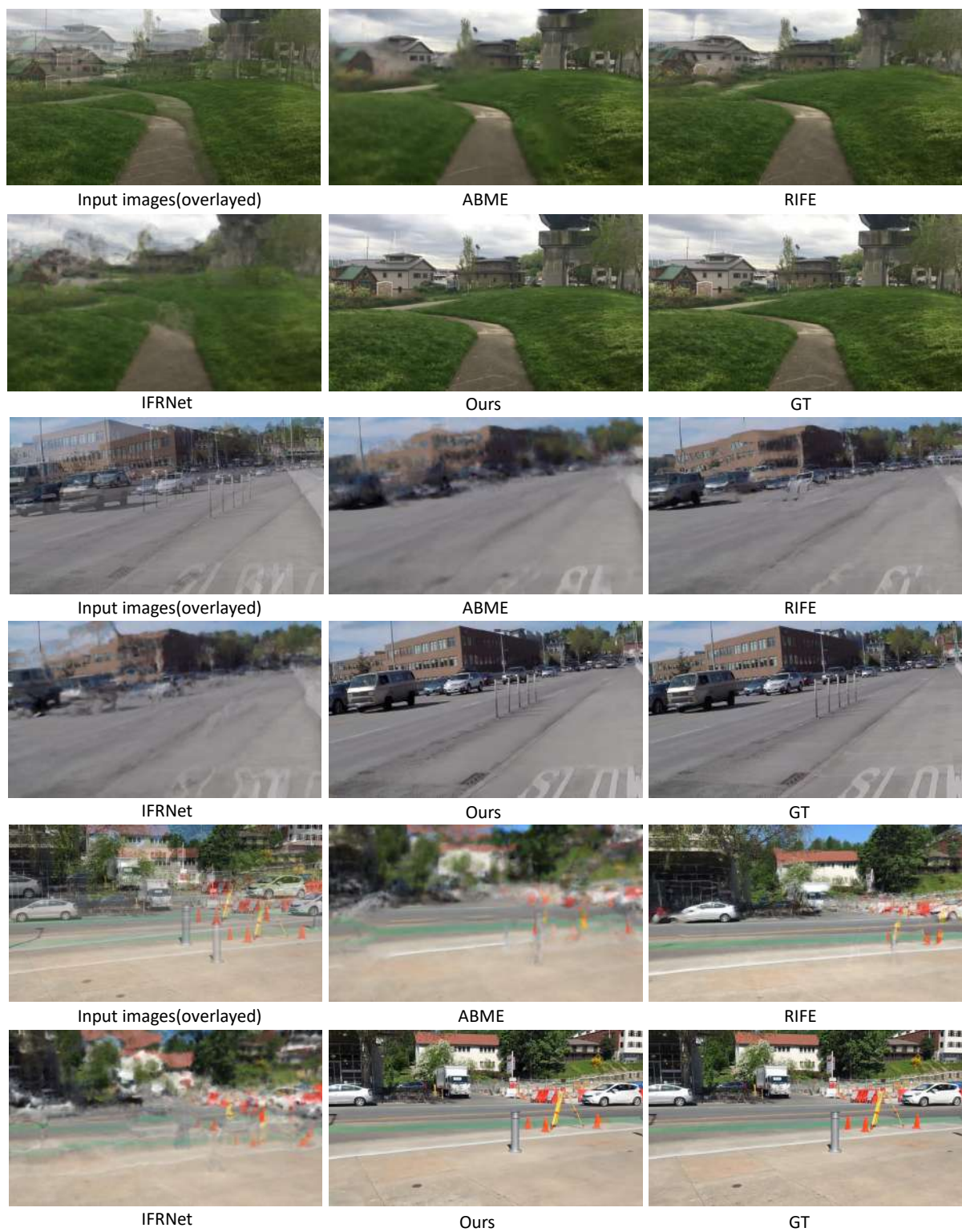


Figure 12. Visual results on the Adobe240fps dataset. (Best viewed when zoomed in.)