Event-based Visible and Infrared Fusion via Multi-task Collaboration

Supplementary Material

The contents of this supplementary material are as follows: Sec. A1 provides additional experimentation details, including dataset statistics, network architecture specifications, and hardware implementation details. Sec. A2 conducts further ablation studies and gives further analysis. Sec. A3 presents additional qualitative results on both synthetic and real data.

A1. Additional Experimentation Details

Dataset Statistics. The KAIST [10] synthetic dataset used in our work consists of 12 sets with 41 clips, where the first 6 sets with 25 clips are split as training data and the rest as test data. Compared to other VIF datasets, KAIST provides consecutive visible and infrared frames, making it suitable for event data generation. After processing, we obtain 6,738 training pairs of blurry/sharp images and corresponding events within the simulated exposure window of each blurry frame. Additionally, from the 6,396 pairs in the test set, we randomly select 1,000 pairs for evaluation. For the real data, we capture 11 clips of infrared/visible frames and corresponding event data on various different scenes.

Network Architecture. For the event-based visible texture reconstruction network and event-guided infrared image deblurring networks, we use the official implementation of E2VID [39] and EFNet [42]. We perform cross-task event enhancement between the multi-scale encoder features of E2VID and the event encoder features of EFNet. Both two features have three scales and we apply enhancement on each scale separately. The output of ConvGRU has the same channel number as its input feature. The 1×1 convolution layers in Fig. 3 of the main paper are used to align the number of channels between the two features. For the fusion task, we adopt the last decoder block feature of E2VID and the SAM feature from EFNet as fusion input. The transformer blocks in the fusion network have 8 heads and the same input/output channel number. Given fused features, we adopt four consecutive convolution layers as the fusion decoder to recover the final fused image.

Hardware Implementation. The two cameras in our hardware system are connected using a customized rig to ensure horizontal alignment. The models of the two cameras are DAVIS 346 for event camera and HIK MV-CI003-GL-N15 for infrared camera. The lens focal length of the event and infrared camera are 16mm and 15mm, respectively, and the camera resolutions are 346×260 and 640×512, respectively.

We align the two cameras by first manually selecting key points from infrared frames and event-reconstructed visible frames. Then we warp the infrared frame to the visible frame by calculating the wrapping matrix. It is widely known that a homography warping matrix estimated from a single frame can be difficult to generalize to other frames. Additionally, the quality of night-time event-reconstructed images may not always be sufficient to support effective keypoint selection and frame-by-frame matching. To address these challenges, we try our best to keep the object distance and scene depth not varying too much when shooting a clip. In this way, a matrix computed for a frame can be applied to its neighboring few frames, which reduces the keypoint selection burden. When the keypoint of a particular frame and all its neighboring frames are too hard to be selected, we simply discard them. Note that our hardware is a experimental prototype that used to validate system design. In practice, advanced per-pixel alignment could be achieved through special VIS-IR beam splitter, such as the one implemented to capture the KAIST dataset [10]. After spatial alignment, all images have a spatial resolution of 346×260, and the events in each sample are recorded over a period of 20 ms corresponding to the exposure time of the infrared camera.

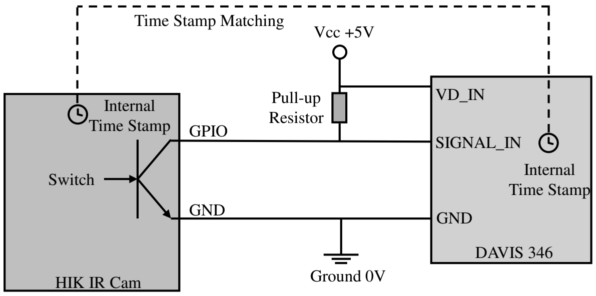


Figure A1. The temporal synchronization circuit between event and infrared cameras. The switch in the IR camera is turned on by default and turned off once start capturing. Therefore, the voltage between the SIGNAL_IN and GND of the event camera is nearly 0 by default, and goes up to +5V when IR Cam start capturing.

For temporal synchronization, we use a 5.0V voltage power supply and a 1KΩ pull-up resistor to form a synchronization circuit, as shown in Fig. A1. At the start of capturing, the infrared camera sends a high logic-level signal through the circuit and records its internal timestamp. At the same time, the signal is received by the event camera, which also records its internal timestamp upon receiving. In this way, we build a correspondence between the internal timestamps of both cameras and achieve synchronization.

Table A1. Ablation study results on the KAIST synthetic dataset. The best results are indicated in **bold**, and the second best are indicated with an underline.

	Model Configuration	EN↑	MI↑	SSIM↑	Q_{CB}↑
I	w/o Cross Task Event Enhance	**7.35**	1.96	1.26	0.33
II	Only shallow features as F_{vis} and F_{ir}	6.62	1.49	1.21	0.30
III	Only deep features as F_{vis} and F_{ir}	7.29	1.97	1.28	0.36
IV	w/o MI optimization	6.90	1.91	1.27	0.37
V	MI maximization only	7.16	1.93	1.26	0.36
VI	MI minimization only	6.58	1.55	1.24	0.32
	EVIF (ours)	7.33	**1.98**	**1.29**	**0.38**

Table A2. Comparison of various design choices of the cross-task event enhancement method.

	Design Choice	PSNR↑
I	w/o Cross-task Event Enhance	33.32
II	Bi-direction GRU → Convolution Layers	33.61
III	Bi-direction GRU → Uni-direction GRU	33.79
IV	Cross-task Channel Attention → Concat	33.55
V	Cross-task Channel Attention → Add	33.57
VI	Cross-task Channel Attention → Multiply	33.48
	Cross-task Event Enhancement	**33.94**

A2. Further Ablation Studies and Analysis

More Ablation Study Results. We conduct additional ablation experiments with multiple model configurations to verify the effectiveness of different components of EVIF on the synthetic KAIST dataset. The results are given in Table A1. The model configurations are as follows:

I. EVIF without using cross-task event enhancement between the first two tasks.

II. Only shallow features are used as inputs F_{vis} and F_{ir} for the fusion task. As depicted in Fig. 2 in the main paper, these shallow features are derived by processing the output images from previous tasks through a convolution layer.

III. Only deep features are used as inputs F_{vis} and F_{ir} for the fusion task. The deep features are extracted from the decoder features of the previous task networks.

IV. EVIF without using mutual information optimization in the fusion task.

V. EVIF with only mutual information maximization in the fusion task.

VI. EVIF with only mutual information minimization in the fusion task.

From the table, we can draw several conclusions: **First**, the cross-task event enhancement method effectively improves fusion results by enhancing the quality of deblurred infrared images. **Second**, utilizing both shallow and cross-task deep features for fusion proves more robust than using either type alone. Meanwhile, cross-task deep features perform much better than shallow features, highlighting the importance of cross-task feature sharing in EVIF. **Third**, the bi-level MI optimization strategy outperforms singular MI maximization or minimization approaches. Notably, performing MI minimization alone even decreases fusion performance due to severe information loss. These ablation study results further affirm the effectiveness of the design elements of EVIF. Moreover, these results position our framework as a pioneering and representative example of design methodology in this challenging new task.

Further Analysis on Cross-task Event Enhancement. The cross-task event enhancement method utilizes several key designs to effectively extract complementary texture information across tasks to facilitate infrared image deblurring, including bi-directional ConvGRU for texture feature extraction and cross-task channel attention (CTCA) for fusing cross-task features. To verify the effectiveness of these designs, we compare the deblurring performance of several alternative design choices on the KAIST dataset. As shown in Table A2, bi-directional GRU performs better than convolution layers for texture feature extraction. This is due to the continuous nature of event data, which demands temporal relation modeling. On the other hand, bi-directional temporal modeling also brings benefits compared to the uni-directional one, as it can capture context from both past and future data, thus gaining a more comprehensive understanding of the sequence. Apart from these, the cross-task channel attention for feature selection performs significantly better than directly concatenating, adding, or multiplying features. This indicates the importance of data-dependent feature selection in cross-task feature sharing. Finally, bringing all the designs together, our full method for cross-task event enhancement achieves the best deblurring performance and facilitates a higher fusion quality.

Further Analysis on bi-level MI optimization. As analyzed in the main manuscript, the bi-level MI optimization in EVIF can highlight the complementary information residing in both visible and infrared modalities while reducing information loss. To give a deeper understanding of such an effect, we visualize the input modality features F_{vis}, F_{ir} and filtered features F'_{vis}, F'_{ir} after MI optimization in Fig. A2. It is obvious that both F'_{vis}, F'_{ir} have highlighted particular areas that contain more information. For the visible modality, F'_{vis} further enhances texture-rich areas (*e.g.*, car, building windows, zebra crossing). On the other hand, we can see that compared with F_{ir}, F'_{ir} keeps the most salient objects in the infrared modality while suppressing background areas that are less informative. The visualization results demonstrate that bi-level MI optimization can successfully extract modality complementary information and provide a comprehensive description of the scene, which echoes our analysis in the main paper.

Impact of the balancing weights of different loss terms. The weights γ_1, γ_2 and γ_3 in Eq. 6 of the main paper are set to reach a balance between each loss terms, especially

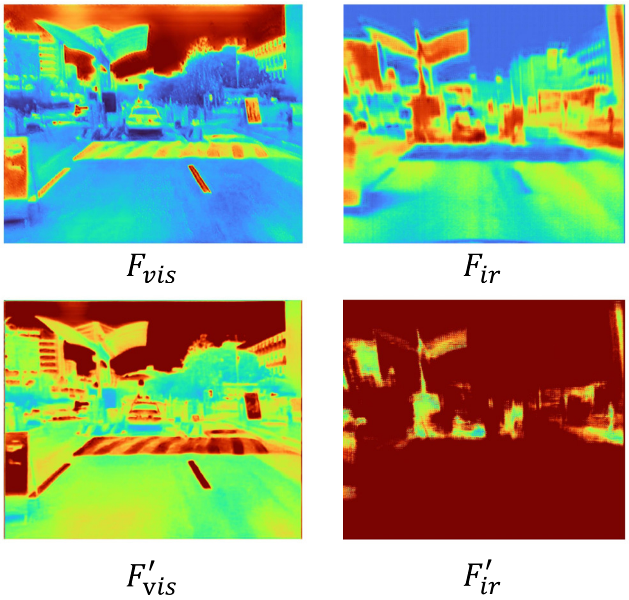


Figure A2. Visualization of visible and infrared modality features before and after MI optimization.

Table A3. Fusion performance of different loss weights on the KAIST synthetic dataset.

$\gamma_2 : \gamma_3$	EN↑	MI↑	SSIM↑	Q_{CB}↑
1:100	7.14	1.95	1.27	0.36
1:10	7.21	1.93	1.25	0.35
1:1	7.28	1.96	1.28	0.36
10:1	7.33	1.98	1.29	0.38
100:1	6.87	1.79	1.23	0.34

between the MI minimization and maximization. To study the impact of these loss terms, we keep γ_1, γ_2 to 1.0, 0.1 and adjust the ratio between γ_2, γ_3 to test the fusion quality. Table A3 gives the fusion results of different weight combinations. From the table, we can see that when $\gamma_2 : \gamma_3$ is less than 1:1, the performance slightly varies. The performance is best when we increase the ratio to 10:1. However, when keep increasing the ratio to 100:1, the performance starts to drop rapidly, this is due to the overly strong MI minimization, which tends to cause information loss.

Limitation of EVIF. While EVIF is a promising system for fusing visible and infrared modalities, it does have limitations due to the inherent sensing characteristics of event cameras, which respond only to changes in light, not to absolute pixel values. This limitation becomes evident in situations with little or no motion, where the event camera might fail to capture valid information. For instance, as shown in Fig. A3, when a car is stationary at a red light, only moving pedestrians are captured by the event camera, leading to a loss of static background information. This issue highlights that EVIF and conventional VIF systems can be cooperative rather than competitive, offering complemented strengths to handle a wide range of conditions.

A3. More Qualitative Comparision Results

Synthetic Data. Figs. A4 and A5 provide additional qualitative results on the synthetic KAIST dataset. The displayed images are marked as follows: "GT IR" and "GT VIS" are ground truth infrared and visible images, respectively. "Blurry IR" and "Blurry VIS" are corresponding simulated blurry infrared and visible images. "Events" denotes the event data. "EVIF Rec. VIS" and "EVIF Deblurred IR" are the results of the event-based visible texture reconstruction and event-guided infrared image deblurring tasks in EVIF. "EVIF Fusion (ours)" is the final fusion result of EVIF. "NAFNet Deblurred IR" and "NAFNet Deblurred VIS" are the frame-based deblurring results obtained by processing "Blurry IR" and "Blurry VIS" with two NAFNet models trained on KAIST infrared and visible data, respectively. The rest images are fusion results of different methods. Fusion methods without $*$ directly apply "Blurry IR" and "Blurry VIS" as VIF input. Fusion results with $*$ adopt "NAFNet Deblurred IR" and "NAFNet Deblurred VIS" as VIF input. Compared to other methods, EVIF excels in recovering visible textures from events and restoring thermal radiation details from blurry infrared images, while conventional VIF methods suffer from blurry effects even with NAFNet deburring as data pre-processing.

Real Data. In addition to the synthetic data, more qualitative results on real-captured data are provided in Figs. A6 to A10. Since no ground truth exists, we only use "IR" and "VIS" to denote the captured infrared and visible images. All comparative methods either directly take "IR" and "VIS" as input, or pre-process them using NAFNet before fusion (marked with $*$). The results show that EVIF offers clearer scene descriptions under extreme lighting and rapid motion, while conventional VIF methods tend to produce inferior results due to reduced input data quality.

Despite the advantage of EVIF, the figures also reveal that real scenes often exhibit blurriness and exposure issues simultaneously. Such a complex degradation significantly increases the difficulty for any pre-processing algorithm to restore satisfactory inputs for conventional VIF methods. Therefore, relying on input pre-processing to handle extreme lighting and high-dynamic motion scenes remains an arduous task. This is also evidenced by the lower effectiveness of NAFNet on the real data. In contrast, EVIF directly uses high-quality event data, bypassing these challenges and offering a more direct approach to scene representation. As a result, our method has the potential to pave the way for future advancements in handling even more diverse and challenging scenarios.

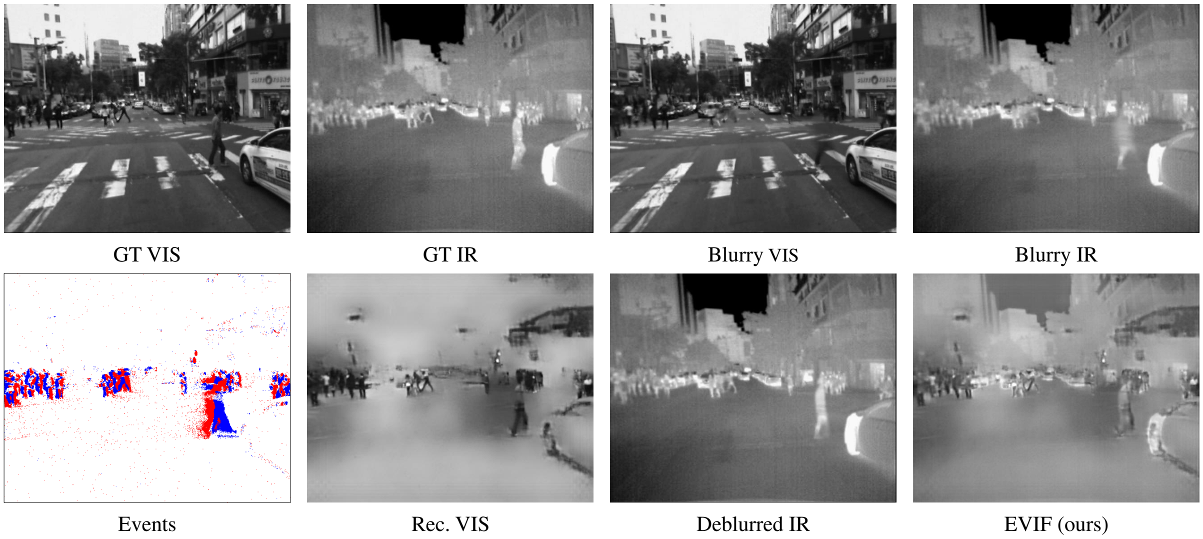
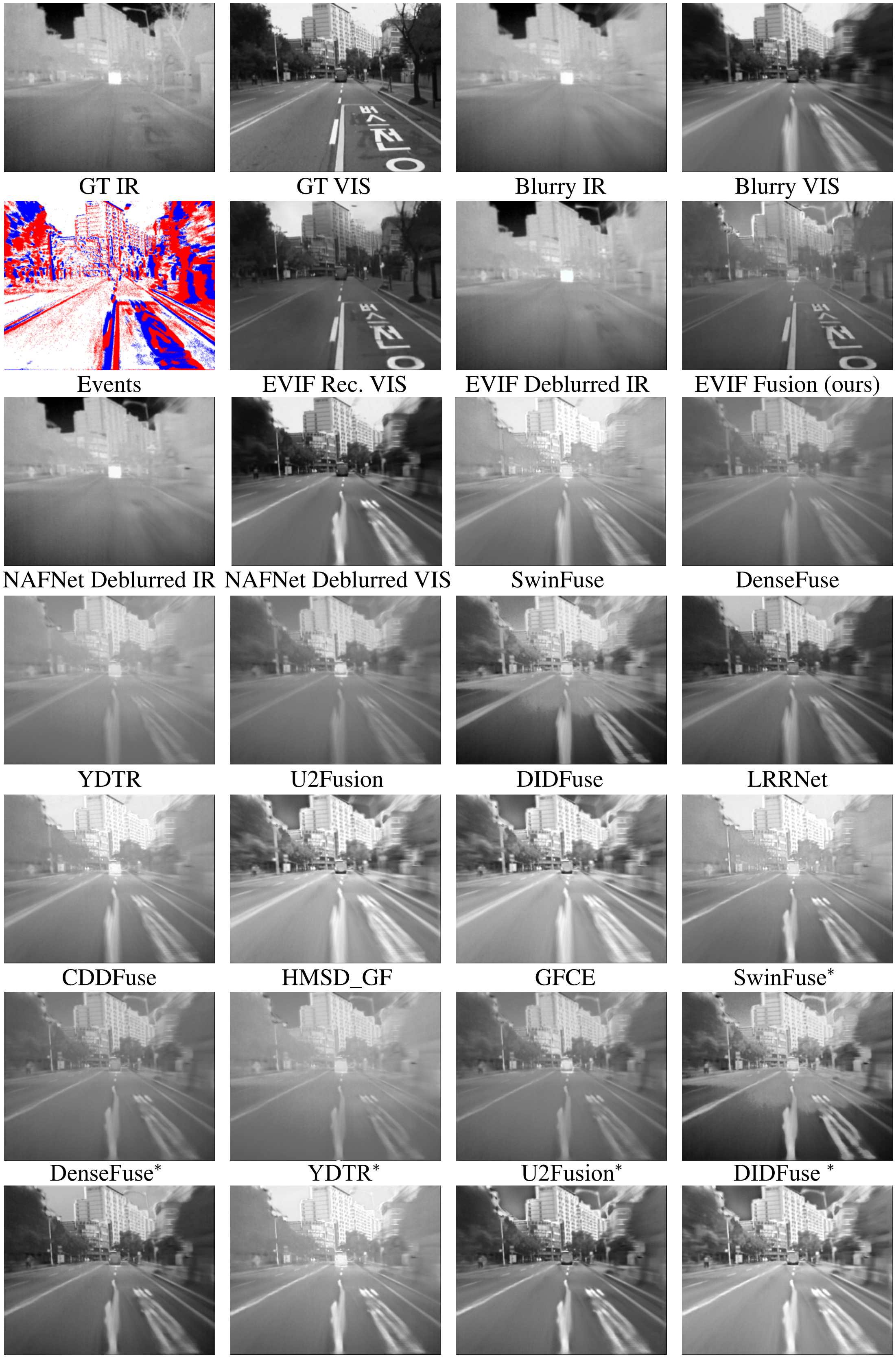


Figure A3. Illustration of EVIF's limitation in a camera-fixed scenario. The image showcases a scene where the car carrying cameras is stopped at a red light, with only moving pedestrians being captured by the event camera. This results in the loss of static background visible information, demonstrating a complementary rather than competitive relationship between EVIF and conventional VIF systems.



GT IR | GT VIS | Blurry IR | Blurry VIS

Events | EVIF Rec. VIS | EVIF Deblurred IR | EVIF Fusion (ours)

NAFNet Deblurred IR | NAFNet Deblurred VIS | SwinFuse | DenseFuse

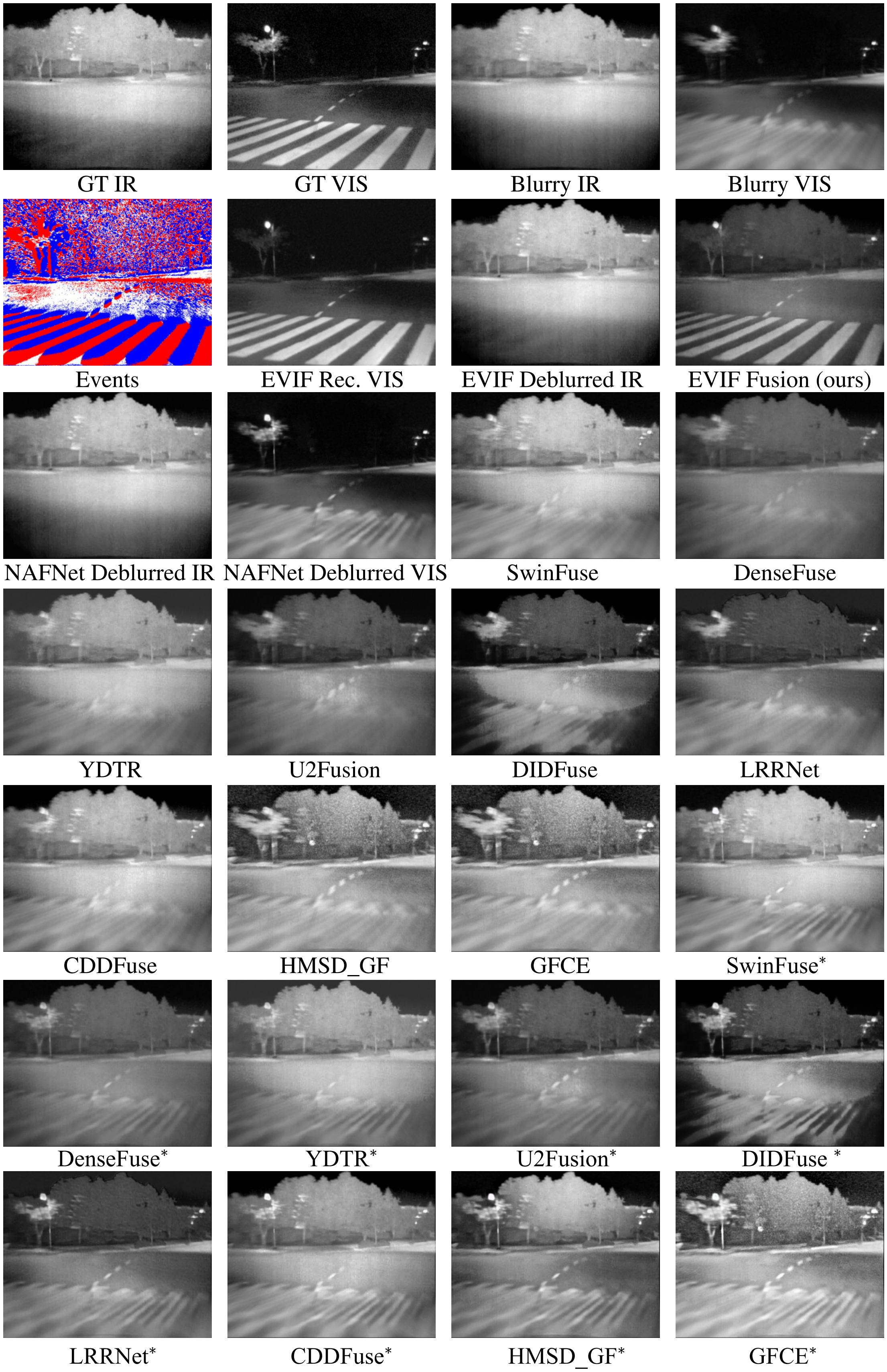
YDTR | U2Fusion | DIDFuse | LRRNet

CDDFuse | HMSD_GF | GFCE | SwinFuse*

DenseFuse* | YDTR* | U2Fusion* | DIDFuse *

LRRNet* | CDDFuse* | HMSD_GF* | GFCE*

Figure A4. Qualitative comparison on KAIST synthetic data. Please refer to the blue text in Sec. A3 for the meaning of each image.



GT IR

GT VIS

Blurry IR

Blurry VIS

Events

EVIF Rec. VIS

EVIF Deblurred IR

EVIF Fusion (ours)

NAFNet Deblurred IR

NAFNet Deblurred VIS

SwinFuse

DenseFuse

YDTR

U2Fusion

DIDFuse

LRRNet

CDDFuse

HMSD_GF

GFCE

SwinFuse*

DenseFuse*

YDTR*

U2Fusion*

DIDFuse *

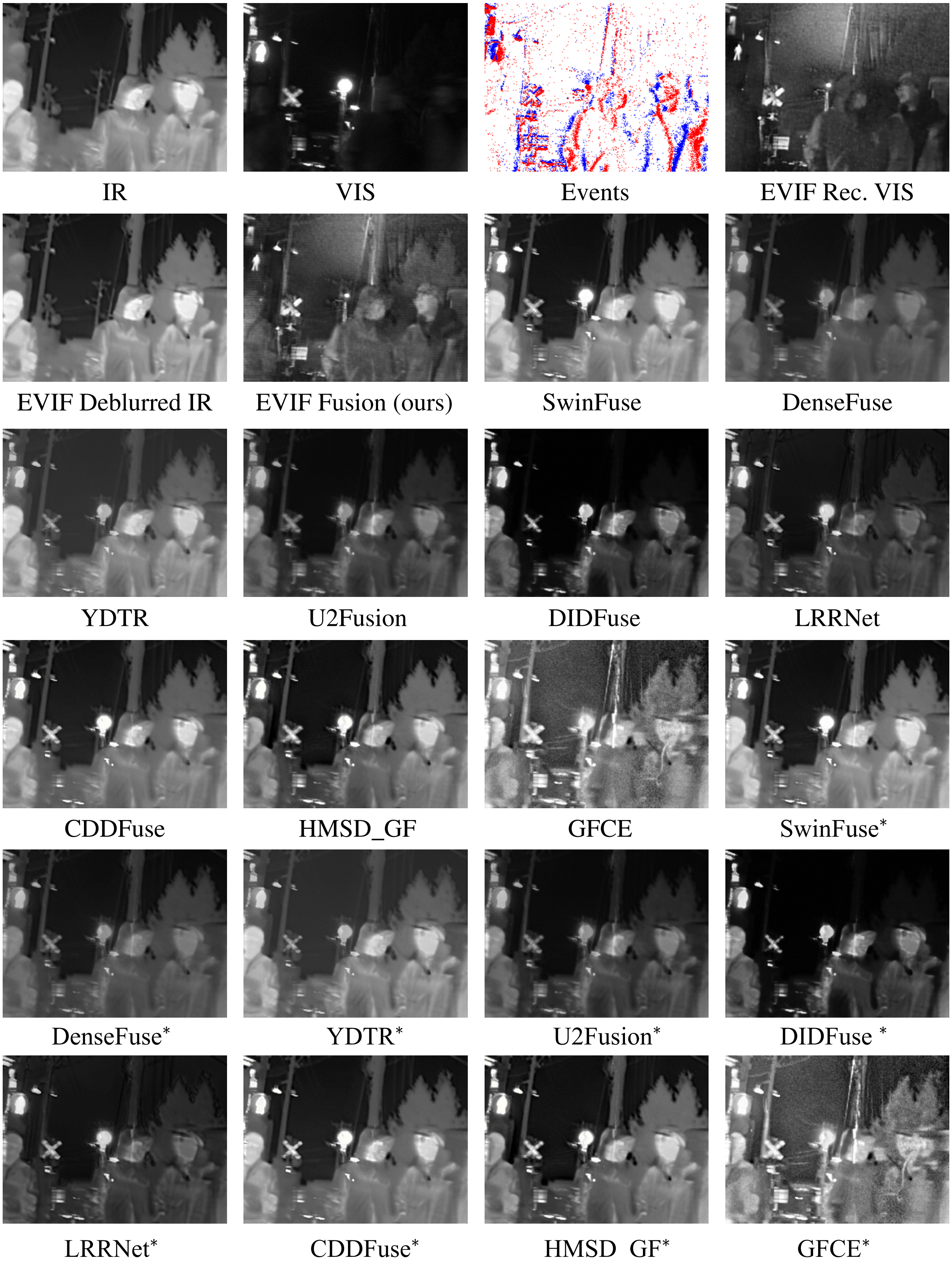
LRRNet*

CDDFuse*

HMSD_GF*

GFCE*

Figure A5. Qualitative comparison on KAIST synthetic data. Please refer to the blue text in Sec. A3 for the meaning of each image.



IR VIS Events EVIF Rec. VIS

EVIF Deblurred IR EVIF Fusion (ours) SwinFuse DenseFuse

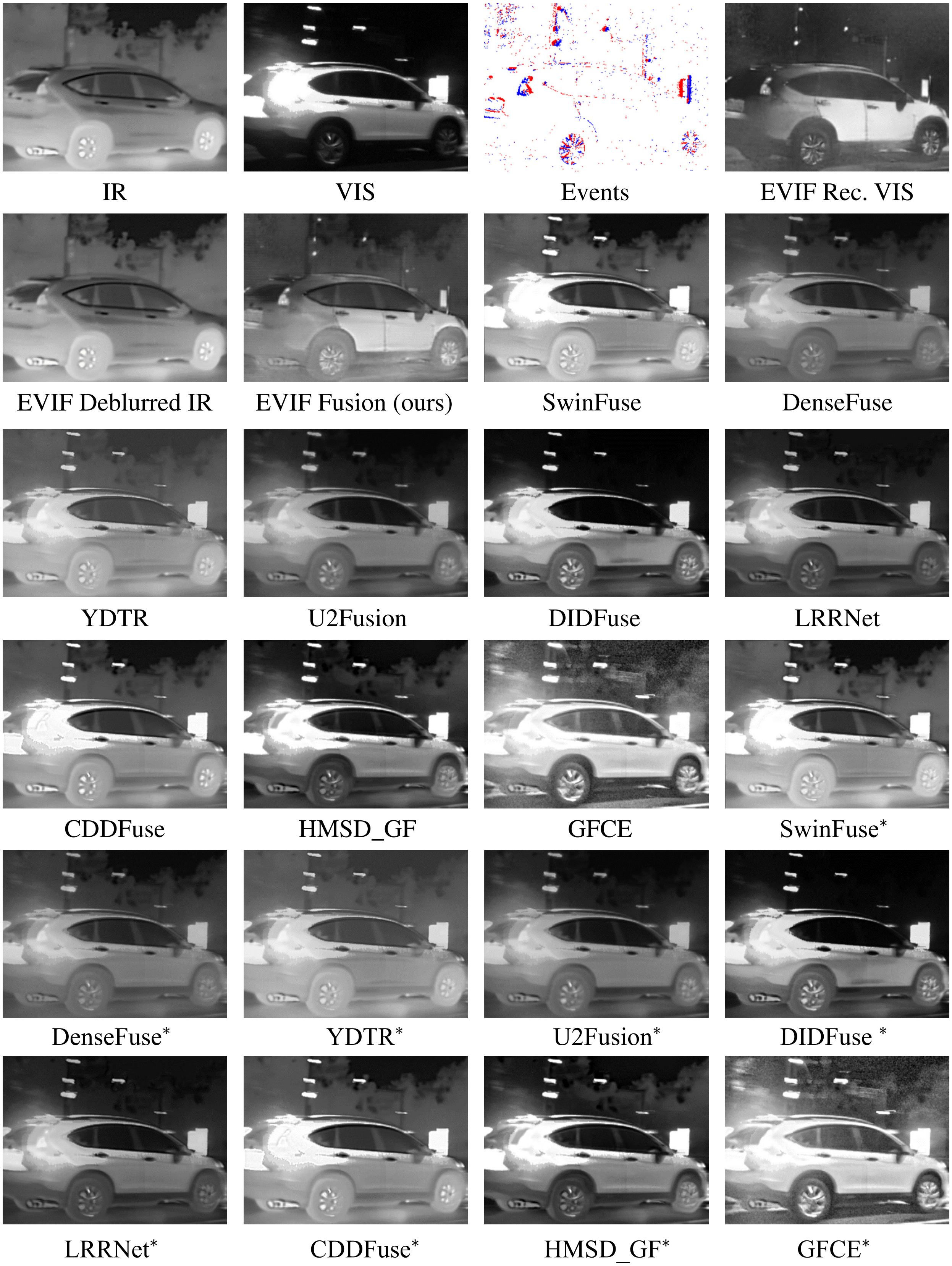
YDTR U2Fusion DIDFuse LRRNet

CDDFuse HMSD_GF GFCE SwinFuse*

DenseFuse* YDTR* U2Fusion* DIDFuse *

LRRNet* CDDFuse* HMSD_GF* GFCE*

Figure A6. Qualitative comparison of EVIF on real-captured data. The detailed meanings of each image are similar to those in Fig. A4 and A5. One exception is that real data has no ground truth, so we only use “IR” and “VIS” to denote the captured infrared and visible images.



IR VIS Events EVIF Rec. VIS

EVIF Deblurred IR EVIF Fusion (ours) SwinFuse DenseFuse

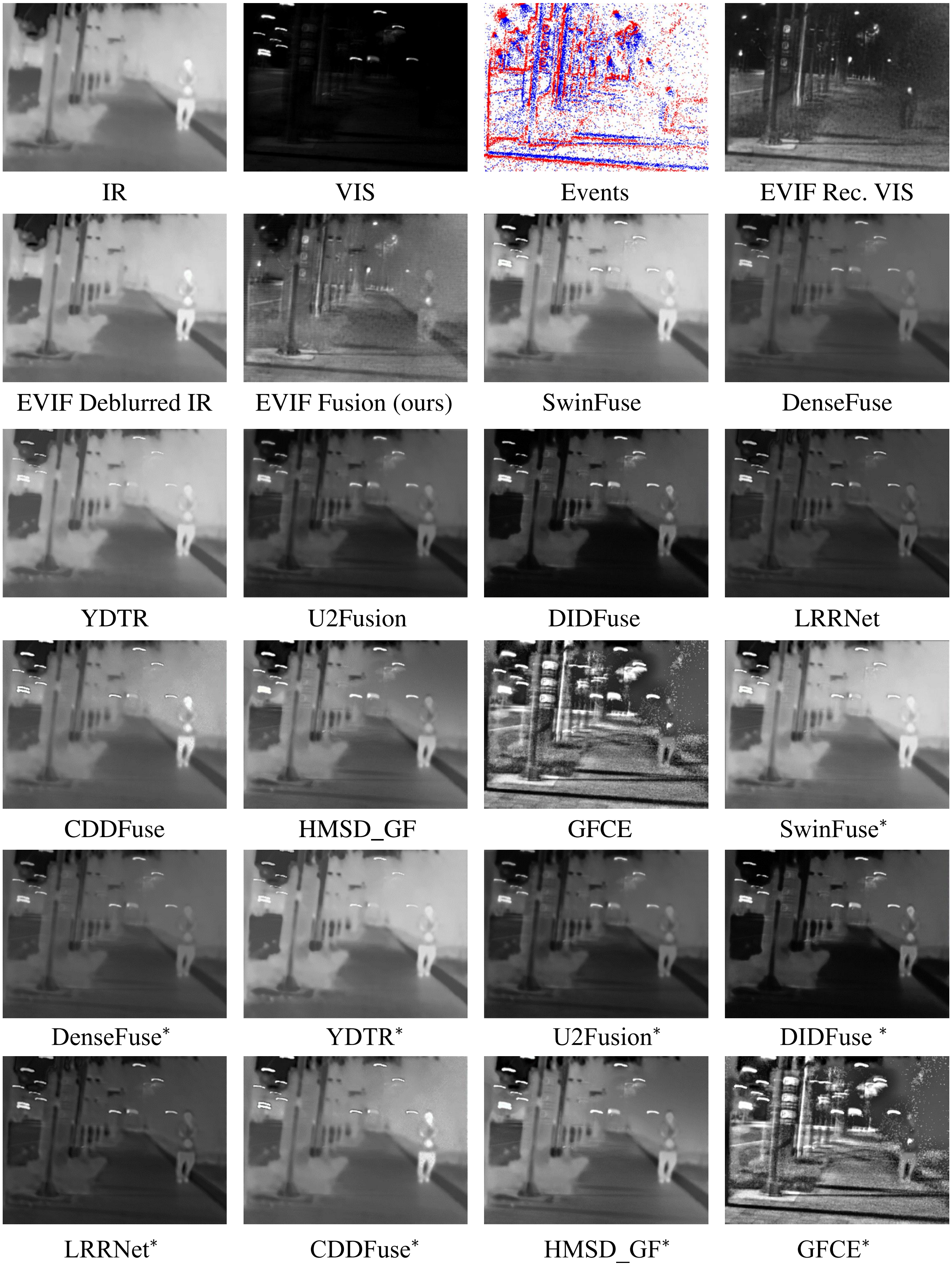
YDTR U2Fusion DIDFuse LRRNet

CDDFuse HMSD_GF GFCE SwinFuse*

DenseFuse* YDTR* U2Fusion* DIDFuse *

LRRNet* CDDFuse* HMSD_GF* GFCE*

Figure A7. Qualitative comparison of EVIF on real-captured data. The detailed meanings of each image are similar to those in Fig. A4 and A5. One exception is that real data has no ground truth, so we only use "IR" and "VIS" to denote the captured infrared and visible images.



IR

VIS

Events

EVIF Rec. VIS

EVIF Deblurred IR

EVIF Fusion (ours)

SwinFuse

DenseFuse

YDTR

U2Fusion

DIDFuse

LRRNet

CDDFuse

HMSD_GF

GFCE

SwinFuse*

DenseFuse*

YDTR*

U2Fusion*

DIDFuse *

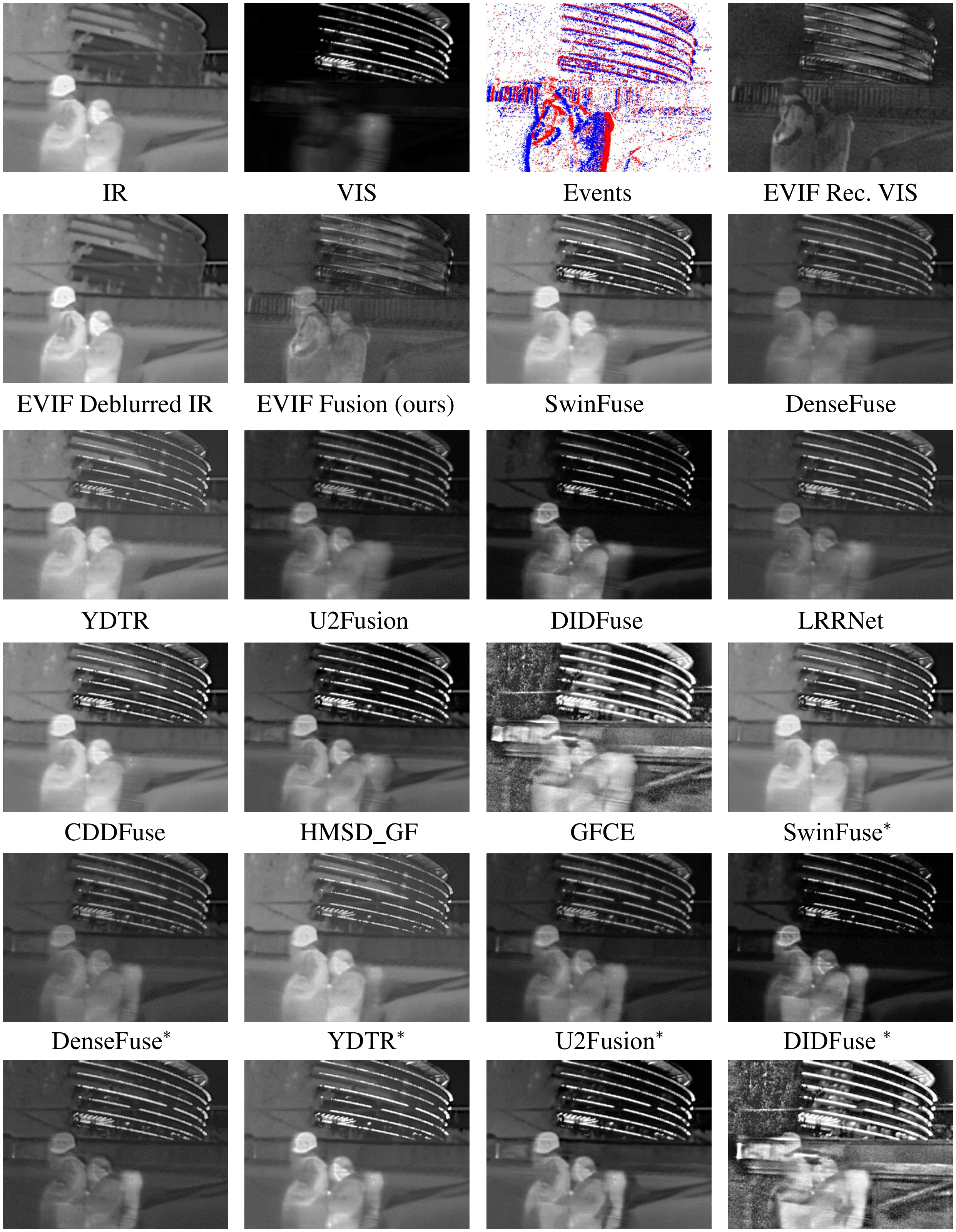
LRRNet*

CDDFuse*

HMSD_GF*

GFCE*

Figure A8. Qualitative comparison of EVIF on real-captured data. The detailed meanings of each image are similar to those in Fig. A4 and A5. One exception is that real data has no ground truth, so we only use "IR" and "VIS" to denote the captured infrared and visible images.



IR VIS Events EVIF Rec. VIS

EVIF Deblurred IR EVIF Fusion (ours) SwinFuse DenseFuse

YDTR U2Fusion DIDFuse LRRNet

CDDFuse HMSD_GF GFCE SwinFuse*

DenseFuse* YDTR* U2Fusion* DIDFuse *

LRRNet* CDDFuse* HMSD_GF* GFCE*

Figure A9. Qualitative comparison of EVIF on real-captured data. The detailed meanings of each image are similar to those in Fig. A4 and A5. One exception is that real data has no ground truth, so we only use "IR" and "VIS" to denote the captured infrared and visible images.

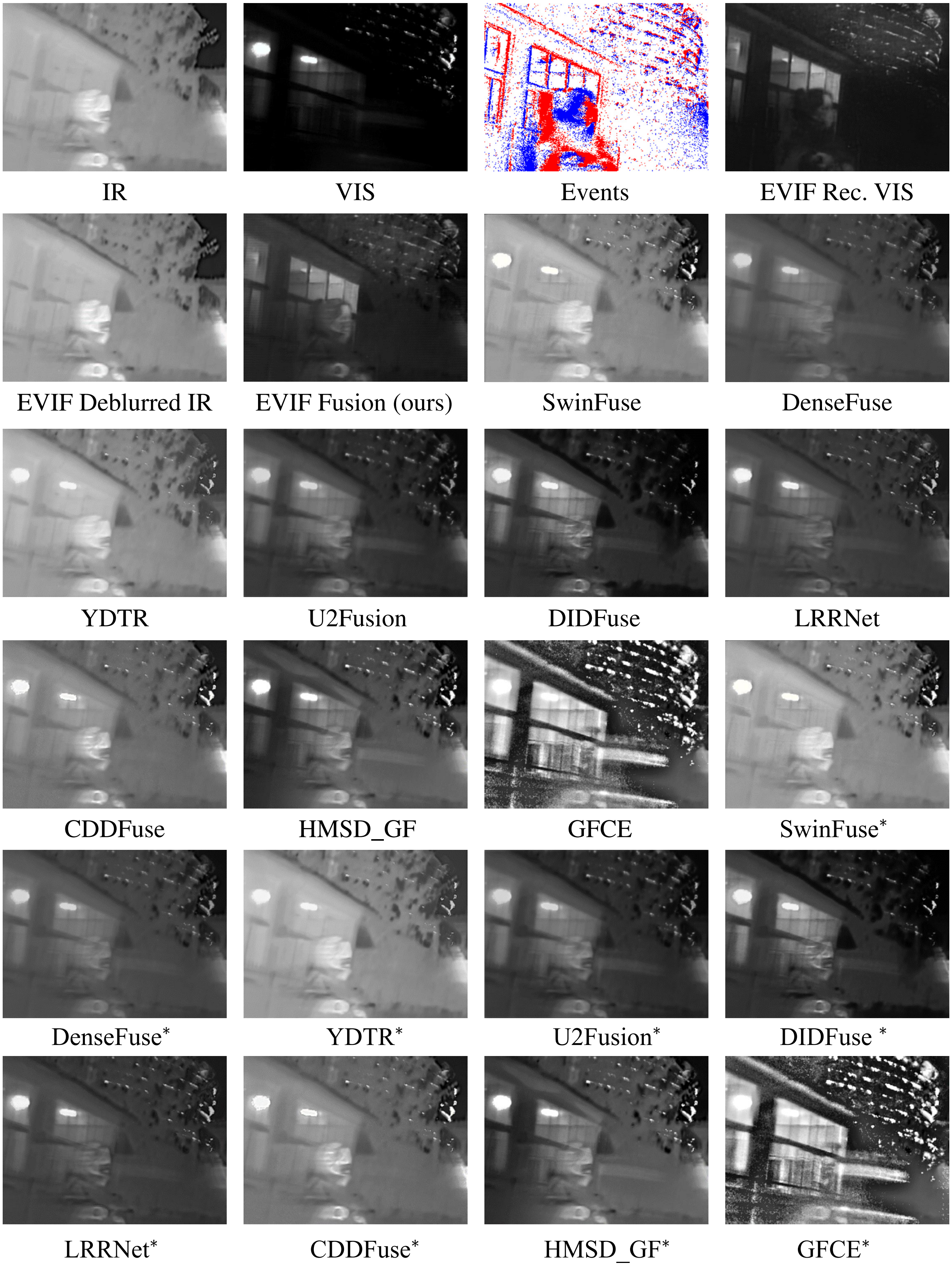


Figure A10. Qualitative comparison of EVIF on real-captured data. The detailed meanings of each image are similar to those in Fig. A4 and A5. One exception is that real data has no ground truth, so we only use "IR" and "VIS" to denote the captured infrared and visible images.