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

Mengyue Geng[1], Lin Zhu[2] *, Lizhi Wang[2], Wei Zhang[4], Ruiqin Xiong[1], Yonghong Tian[1,3,4] *

[1]School of Computer Science, Peking University
[2]Beijing Institute of Technology
[3]School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University
[4]Peng Cheng Laboratory

{mygeng, rqxiong, yhtian}@pku.edu.cn, {linzhu, wanglizhi}@bit.edu.cn, zhangwei1213052@126.com

Abstract

Visible and Infrared image Fusion (VIF) offers a comprehensive scene description by combining thermal infrared images with the rich textures from visible cameras. However, conventional VIF systems may capture over/under exposure or blurry images in extreme lighting and high dynamic motion scenarios, leading to degraded fusion results. To address these problems, we propose a novel Event-based Visible and Infrared Fusion (EVIF) system that employs a visible event camera as an alternative to traditional frame-based cameras for the VIF task. With extremely low latency and high dynamic range, event cameras can effectively address blurriness and are robust against diverse luminous ranges. To produce high-quality fused images, we develop a multi-task collaborative framework that simultaneously performs event-based visible texture reconstruction, event-guided infrared image deblurring, and visible-infrared fusion. Rather than independently learning these tasks, our framework capitalizes on their synergy, leveraging cross-task event enhancement for efficient deblurring and bi-level min-max mutual information optimization to achieve higher fusion quality. Experiments on both synthetic and real data show that EVIF achieves remarkable performance in dealing with extreme lighting conditions and high-dynamic scenes, ensuring high-quality fused images across a broad range of practical scenarios.

1. Introduction

Visible and Infrared image Fusion (VIF) has attracted sustained research interest due to its wide range of applications such as robotic vision [21], surveillance [45] and remote sensing [6]. Visible images contain abundant scene textures. However, they are sensitive to illumination changes.

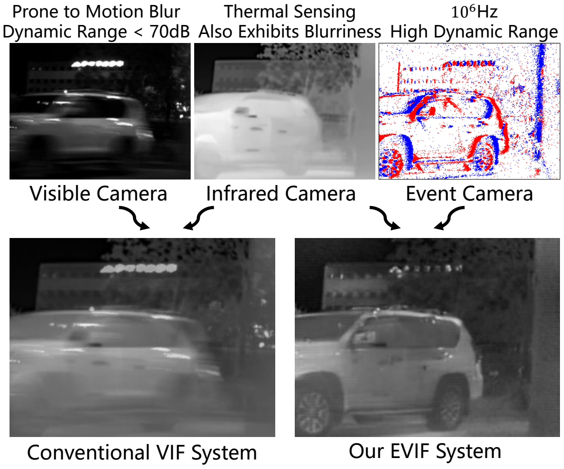

Figure 1. Comparision between conventional VIF system and our EVIF system. In challenging scenes with extreme lighting and rapid motion, conventional VIF systems often produce over/underexposed or blurry images, leading to poor fusion quality. Leveraging the high dynamic range and low latency of event cameras, EVIF excels in capturing visible textures and can effectively mitigating issues of blurriness in such scenarios.

Infrared images, on the other hand, capture thermal information that is not affected by light but may lose texture details. Recognizing this complementarity, the goal of VIF is to take the best of both modalities and produce a fused image that allows a comprehensive description of the scene.

Studies over the past decades have provided remarkable progress in the field of VIF [20, 28, 44]. Traditional approaches typically fall into several categories such as multiscale transforms-based and sparse representation-based ones [29]. With the rise of deep learning techniques, a plethora of deep learning-based methods have also been developed [59]. Alongside algorithmic advancements, there have been significant data contributions, marked by the release of increasingly comprehensive and valuable

*Corresponding authors.

datasets [2, 3, 16, 17] in various target scenes.

Despite significant achievements, current VIF systems can sometimes deliver subpar results in scenarios with extreme lighting or high dynamic motion. A primary constraint is the input data quality. Traditional frame-based visible cameras usually possess a relatively low dynamic range (*e.g.*, about 60 dB). As shown in Fig. 1, this can lead to overexposure or underexposure when scenes have a vast range of ambient lighting. Moreover, in scenes with rapid motion, these frame-based cameras might experience motion blur due to their limited frame rates. Meanwhile, infrared cameras are not immune to motion blur either, especially for uncooled microbolometers prevalent in consumer applications and research domains [14, 35]. Given the compromised input quality, the output from VIF algorithms is inevitably affected. As a result, the capabilities of current VIF methods are often inadequate for these challenging scenarios.

In this paper, we address the aforementioned problems by presenting a novel Event-based Visible and Infrared Fusion (EVIF) system. Unlike traditional VIF systems, EVIF pairs an infrared camera with a visible event camera [24]. Event cameras are biologically inspired sensors that record rapid changes in light intensity with high dynamic range ($>$ 120 dB) and low-latency (in the order of μs) event signals. In EVIF, the recorded events serve a dual purpose: 1) to unearth visible textures, and 2) to provide motion clues that aid in transforming blurry infrared images into sharp ones. Thanks to the high dynamic range and low latency of the event camera, the extracted visible textures are free from blurriness and remain robust even under extreme lighting conditions. Concurrently, the fine-grained motion perception ability of events significantly bolsters the quality of sharp infrared image recovery.

With the EVIF system, a pivotal question emerges: *How can we effectively harmonize events with infrared images to produce high-quality fused images?* Examining the system design, we delineate three core tasks within EVIF: 1) event-based visible texture reconstruction, 2) event-guided infrared image deblurring, and 3) visible and infrared fusion. An intuitive approach might be to tackle each task separately. However, considering that these tasks cater to different facets of the very same scene, there is an inherent notion that they could be complementary.

Building upon the above analysis, we introduce a novel multi-task collaborative framework designed to synergistically tackle the three delineated tasks, thereby optimizing the visible and infrared fusion quality. Within our framework, three networks are assigned to the three tasks. Rather than allowing each network to learn in isolation, we emphasize their interplay and mutual reinforcement. To realize this, we first propose a cross-task event enhancement method aimed at the efficient deblurring of infrared images. Intuitively, event features for texture reconstruction mainly capture fine scene details, while those for motion deblurring tend to focus on edge motion. In light of this, the cross-task event enhancement module extracts the appearance details in the former with a bi-directional recurrent adaptor and fuses with the latter to compensate for the appearance information loss, yielding improved results for more precise infrared image recovery.

Another collaborative aspect of our framework lies in the fusion mechanism, where we diverge from existing solutions and embrace a bi-level min-max mutual information optimization approach. Specifically, after reconstructing the visible image and deblurring the infrared one, we fuse them in the decoded feature space, enriching the representation of both modalities. We minimize mutual information between filtered modality features to decrease redundancy and increase complementarity, while maximizing it between the fused feature and original modality feature before filtering to prevent information loss. This min-max optimization balances feature distinctiveness and completeness, improving fusion quality.

Extensive experiments on both synthetic data and real data captured by a prototype hardware system verify the effectiveness of our approach. To summarize, the contributions of this paper are as follows:

- We propose an Event-based Visible and Infrared Fusion (EVIF) system, which leverages an event camera to address the limitations of traditional VIF systems in extreme lighting and high dynamic motion scenarios. To the best of our knowledge, this is the first work that utilize events to address VIF tasks under such challenging conditions.
- We design a multi-task collaborative framework for EVIF to obtain high-quality fused images. By exploiting cross-task event enhancement and bi-level mutual information optimization, the performance of individual tasks in EVIF can be elevated to achieve better fusion quality.
- We build a prototype system to verify the effectiveness of our method and contribute real data to promote further research.

2. Related Work

Visible and Infrared Image Fusion. Visible and Infrared image Fusion (VIF) techniques have garnered significant attention over the past decades [29]. Traditional methods typically extract suitable representations from both images [5, 13, 18, 26, 52]. The fusion process then involves combining these representations using well-designed fusion strategies. With the rise of deep learning in various computer vision domains, many deep learning-based VIF methods have also emerged, promising enhanced fusion quality [59]. Most of these methods leverage CNNs [20, 22, 27, 31]. More recently, GAN-based [25, 30, 57] and Transformer-based [36, 50] methods have also gained traction. While these advancements have pushed the bound-

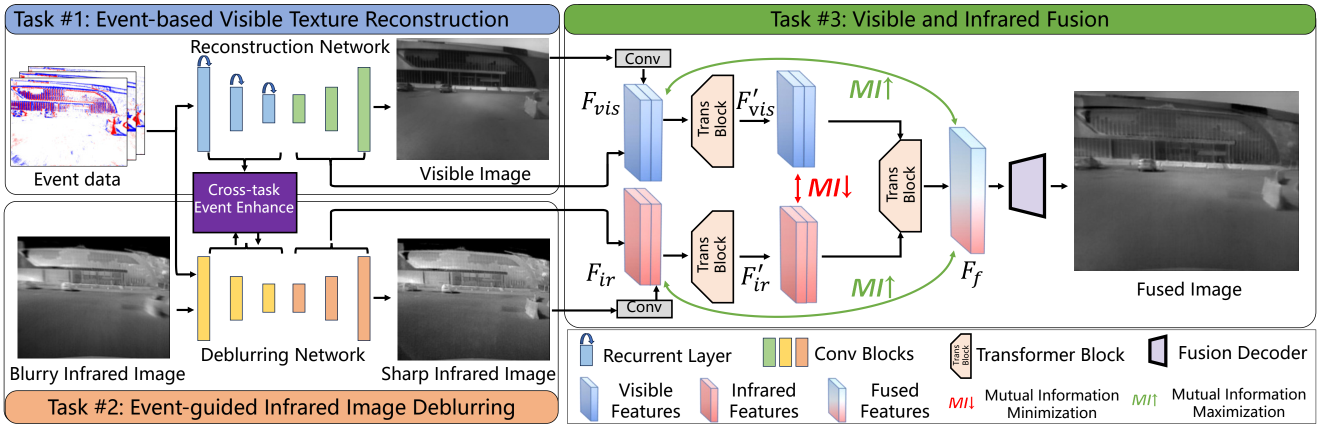

Figure 2. Overall framework of the EVIF system involving three interleaved tasks. To achieve multitask collaboration, EVIF adopts cross-task event enhancement to reinforce useful texture information from the texture reconstruction task into the deblurring task. Meanwhile, the decoded features from both tasks are further involved in the fusion process. Finally, the two modality features are fused with bi-level min-max mutual information optimization to highlight complemented features from both modalities and obtain robust fused images.

aries of VIF, challenges intrinsic to the data quality persist. Due to the hardware limitations of frame-based visible and infrared cameras, VIF systems might capture over/underexposed or blurry images in scenarios with extreme lighting or high dynamic motion, resulting in a notable performance drop.

Event Cameras. Event cameras [15, 24] operate by monitoring changes in per-pixel intensity in an asynchronous manner, as opposed to capturing fixed pixel values. During capturing, event cameras generate an event for a pixel whenever its logarithmic intensity shift surpasses a certain threshold. Due to the unique sampling mechanism and circuit design, event cameras can achieve a significantly higher temporal resolution than traditional frame-based cameras with a broad dynamic range. These advantages make them suitable for many applications, such as high frame rate, high dynamic range video reconstruction [38, 51] and synthetic aperture imaging [60]. Similarly, events are especially useful for VIF systems, as it can recover blurry-free visible textures and provide motion cues for restoring potential sharp infrared images from the captured blurry ones. Therefore, we are motivated to use event cameras to build the EVIF system, aiming to overcome the challenges posted by extreme light conditions and high dynamic motion scenes.

Multi-task Learning. Multi-task Learning (MTL) involves the simultaneous learning of multiple related tasks [48, 62]. By leveraging shared information among tasks, MTL can potentially improve the performance of individual tasks while reducing overall model complexity. MTL has attracted considerable attention due to its versatility, encompassing diverse task combinations and learning paradigms [61, 62]. For EVIF, we dissect the fusion process into three interlinked tasks, thereby addressing the three important questions inherent in MTL [62]:

"When to share": We evaluate the trade-offs between single-task and multi-task models, and confirm the benefits of cross-task collaboration in EVIF.

"What to share": We identify the mode of knowledge transfer among tasks as a hybrid feature and instance-based one, where we transfer useful features across tasks for better learning along with shared data instances (*e.g.*, event signals of the same time interval).

"How to share": We design cross-task event enhancement and bi-level mutual information minimization methods to specify concrete ways to share knowledge among tasks, resulting in improved performance.

3. Methodology

3.1. Problem Definition

Suppose we have a continuous event stream $\mathcal{E}$ within a time interval $[t_0, t_1]$ and a set of M potentially motion-blurred infrared frames $I_{\text{ir}} = \{I_{\text{ir}}^i \mid i = 1, 2, \dots, M\}$. Each I_{ir}^i has an exposure time window $[t_i, t_i+\delta] \in [t_0, t_1]$, where δ is the exposure time length of the infrared camera. The objective of EVIF is to produce sharp infrared-visible fusion images I_f^i for each infrared frame I_{ir}^i. The task involves extracting synchronized visible textures and motion clues from $\mathcal{E}$, and integrating them with I_{ir} to create a clear, comprehensive depiction of the scene. The fusion process aims to overcome the limitations of motion blur and extreme lighting conditions, leveraging the unique strengths of both infrared and visible event data.

3.2. Framework Overview

Fig. 2 illustrates the over all framework of our proposed EVIF system. The framework jointly addresses three interleaved tasks, each achieved by a specific task-related network. Since the first two tasks (*i.e.*, event-based visible tex-

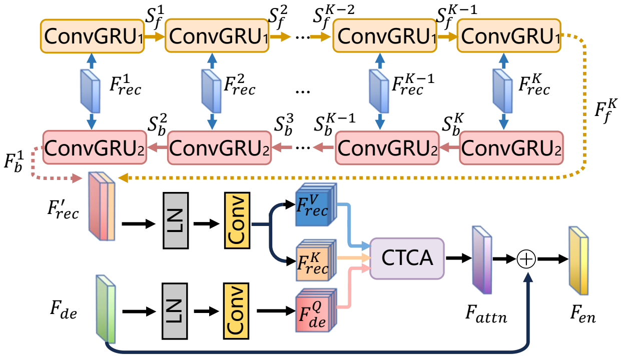

Figure 3. Details of the cross-task event enhancement. LN denotes layer normalization, Conv stands for a 1×1 convolution layer, and CTCA denotes the cross-task channel attention.

ture reconstruction and event-based image deblurring) are well-studied in literature [33, 39, 42, 55], we adopted the state-of-the-art E2VID [39] and EFNet [42] as their task-related networks and put our focus on their synergy[1]. To achieve this, a cross-task event enhancement method is designed to exploit useful texture features from the event reconstruction task. These features are then used to assist infrared image deblurring. Finally, the decoded features from the first two tasks are sent to a fusion network, which employs a bi-level min-max mutual information optimization mechanism to achieve robust fusion.

For each input blurry infrared image I_{ir}^i, its corresponding event segment $\mathcal{E}_{t_i}^{t_i+\delta}$ captured within the exposure time window of I_{ir}^i is treated as input event data. The deblurring network directly takes $\mathcal{E}_{t_i}^{t_i+\delta}$ as input, while the event texture reconstruction network equally divide $\mathcal{E}_{t_i}^{t_i+\delta}$ into K segments over time and process them recurrently, resulting a set of K event features and reconstructed visible images. During fusion, only the middle $\frac{K+1}{2}$-th visible image is used, while all K event features are participated in cross-task event enhancement. Further details on these processes and their integration in our framework are discussed in the subsequent sections.

3.3. Cross-task Event Enhancement

Since the primary aim of the event-based deblurring network is to uncover potential motion clues from events, the texture features inherent in the events might not be fully exploited within it. Considering this, we develop a cross-task event enhancement method. This method is designed to enhance the event features within the event-based infrared image deblurring network, effectively leveraging the texture information learned from the event-based visible texture reconstruction task. As shown in Fig. 3, Given K event features $\{F_{rec}^i \mid i = 1, 2, \dots, K\}$ from the event reconstruction network, we need to summarize the spatial textures within each F_{rec}^i while considering the temporal correlation between them. To achieve this, two ConvGRUs [4] are used to extract spatial-temporal features from F_{rec}^i in a bi-directional recurrent manner:

$$\begin{aligned} S_f^{i+1}, F_f^{i+1} &= \text{ConvGRU}_1(S_f^i, F_{rec}^i), \\ S_b^{i-1}, F_b^{i-1} &= \text{ConvGRU}_2(S_b^i, F_{rec}^i), \end{aligned} \tag{1}$$

where S_f^i and S_b^i are forward and backward hidden states, F_f^i and F_b^i are output features of GRUs. The endpoint features F_f^K and F_b^1 are then stacked along channel dimension to form a single feature F'_{rec}, which contains abundant texture information over time.

After obtaining F'_{rec}, the next key step is to merge the texture information in F'_{rec} into the event feature F_{de} of the deblurring network. A direct solution would be add, multiply, or concatenate F'_{rec} with F_{de}, as adopted in some previous MTL works [9, 47]. However, these approaches overlook the drastically different feature distributions between the two tasks, which may compromise the original motion cues in F_{de}. To avoid this issue, we instead design a Cross-Task Channel Attention (CTCA) to merge F'_{rec} with F_{de}. Different from the vanilla attention mechanism [49] that generates queries, keys, and values from a single input, CTCA calculates the query feature F_{de}^Q from F_{de}, while the key feature F_{rec}^K and value feature F_{rec}^V are obtained from F'_{rec}. All three features are reshaped to $(hw) \times c$. The output of CTCA is then calculated by performing attention along channel dimension:

$$F_{attn} = F_{rec}^V \text{Softmax}\left(\frac{(F_{de}^Q)^T F_{rec}^K}{\sqrt{hw}}\right). \tag{2}$$

Finally, F_{attn} and F_{de} are added to get the enhanced event feature F_{en} for further processing in the deblurring network.

3.4. Fusion by Mutual Information Optimization

The goal of visible and infrared fusion is to acquire images that contain abundant visible texture scene details while highlighting salient targets captured in the infrared modality. Therefore, effectively harnessing the complementary information contained in the two modalities is a key factor in determining the performance of the fusion process.

To encourage complement feature discovery, a proper strategy must be used to balance feature distinctiveness and completeness. Intuitively, the ideal features for fusion should highlight distinct information in each modality while reducing redundancy information that is common across modalities. On the other hand, the fused results must retain the original modality information as much as possible to avoid potential information loss. Based on this analysis, we propose to optimize the mutual information between the two modality features in a bi-level min-max fashion. As

[1] Note that the EVIF system is inherently flexible and also allows the usage of other task-specific networks.

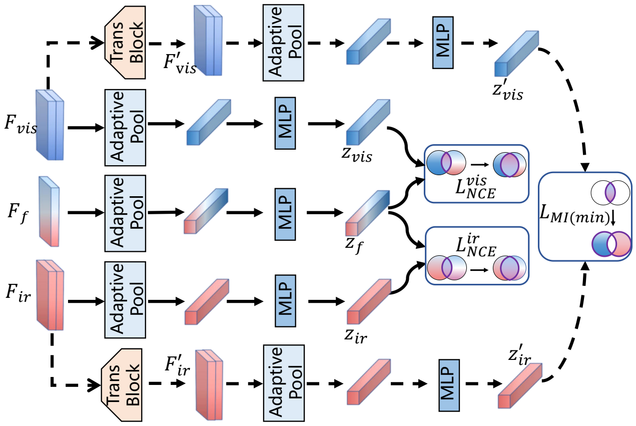

Figure 4. The process of mutual information minimization to highlight distinct modality-specific features, coupling with the process of mutual information maximization to retain original information and avoid potential information loss.

shown in Fig. 2, after obtaining the reconstructed visible image and deblurred sharp infrared image, we pass them into convolution layers and concatenate them with the decoder outputs of the previous task networks. The concatenated results consist of both shallow and deep representations, yielding reliable features F_{vis} and F_{ir} for visible and infrared modality, respectively. We then impose mutual information optimization over the two modality features to encourage complementary information learning.

Specifically, we adopt transformer blocks that apply multi-headed self-attention over spatial locations [12] as our learnable information filter that exploit long-range spatial dependency. After passing the two modality features F_{vis} and F_{ir} into transformer blocks and getting the output features F'_{vis} and F'_{ir}, they are imposed with mutual information minimization to reduce their redundancy and highlight modality-dintinct information. The whole process is illustrated in Fig. 4. As shown in the figure, F'_{vis} and F'_{ir} are adaptively pooled into one-dimensional vectors and passed through MLP layers to obtain two variational latent embeddings z'_{vis} and z'_{ir}. Then, the mutual information between z'_{vis} and z'_{ir} can be expressed as:

$$MI(z'_{vis}, z'_{ir}) = H(z'_{vis}) + H(z'_{ir}) - H(z'_{vis}, z'_{ir}), \quad (3)$$

where $H(z'_{vis})$ and $H(z'_{ir})$ denotes the marginal entropy of z'_{vis} and z'_{ir}, and $H(z'_{vis}, z'_{ir})$ is their joint entropy. To calculate $MI(z'_{vis}, z'_{ir})$, We follow previous works [58, 65] and leverage Kullback-Leibler (KL) divergence to calculate the marginal entropy and obtain the mutual information minimization loss:

$$\begin{aligned}\mathcal{L}_{MI}(z'_{vis}, z'_{ir}) = \hat{H}(z'_{ir}, z'_{vis}) + \hat{H}(z'_{vis}, z'_{ir}) \\ - (KL(z'_{vis}||z'_{ir}) + KL(z'_{ir}||z'_{vis})),\end{aligned} \quad (4)$$

where $\hat{H}(z'_{ir}, z'_{vis})$ is the cross-entropy from z'_{vis} to z'_{ir}.

By minimizing the mutual information between F'_{vis} and F'_{ir}, the modality distinct features are highlighted. We then concatenate F'_{vis} and F'_{ir} and fuse them using another transformer block to get the fused feature F_f. However, solely minimizing mutual information may cause potential information loss, as the network may learn to discard important features to force a lower $\mathcal{L}_{MI}$. To alleviate this issue, we further impose mutual information maximization between F_f and original modality features F_{vis} and F_{ir}. As shown in Fig. 4, three latent embeddings z_f, z_{vis} and z_{ir} are obtained from F_f, F_{vis} and F_{ir}, respectively. To maximize the mutual information, we take inspiration from [41, 46] and optimize the InfoNCE [32] objective by treating each $\{z_{vis}, z_f\}$ as positive and other samples in the same batch as negative:

$$\mathcal{L}^{vis}_{NCE} = -\sum_{i=1}^{N} \log \frac{\exp(z_{f_i}^T z_{vis_i})}{\sum_{j=1}^{N} \exp(z_{f_i}^T z_{vis_j})}, \quad (5)$$

where N is the batch size, z_{vis_i} and z_{f_i} are the i-th corresponded embeddings in the batch. The $\mathcal{L}^{ir}_{NCE}$ for infrared modality can be defined similarly. Note that the loss for MI minimization (Eq. 4) and maximization (Eq. 5) are different since the MI optimization is usually done by adjusting MI bounds and not MI itself [1]. Therefore, $-\mathcal{L}_{MI}$ can not be used to replace $\mathcal{L}_{NCE}$ and vise versa.

By optimizing mutual information in a bi-level min-max manner, the EVIF framework effectively enhances the distinctiveness of features while maintaining a comprehensive representation of each modality. This approach ensures that the fused features capture essential characteristics from both modalities, balancing uniqueness and completeness.

3.5. Progressive Training

Overall, the training of EVIF follows a three-staged manner, in which the three tasks are learned sequentially and progressively. At each stage, we keep training tasks in previous stages, together with the new task at the current stage. For the first two tasks, L2 loss is applied as the training objective. After the event reconstruction network has finished training, the deblurring network is further trained along with cross-task event enhancement. Finally, the fusion network is trained with the following objective:

$$\mathcal{L}_{fuse} = \gamma_1 \mathcal{L}_{SSIM} + \gamma_2 \mathcal{L}_{MI} + \gamma_3 (\mathcal{L}^{vis}_{NCE} + \mathcal{L}^{ir}_{NCE}), \quad (6)$$

where $\mathcal{L}_{SSIM}$ is the SSIM loss between the fused image and the output images of the previous tasks. γ_1, γ_2 and γ_3 are balancing weights of different loss terms.

4. Experiments

4.1. Dataset and Hardware System

Dataset To thoroughly verify the effectiveness of our EVIF system, we conduct experiments on both synthetic data and

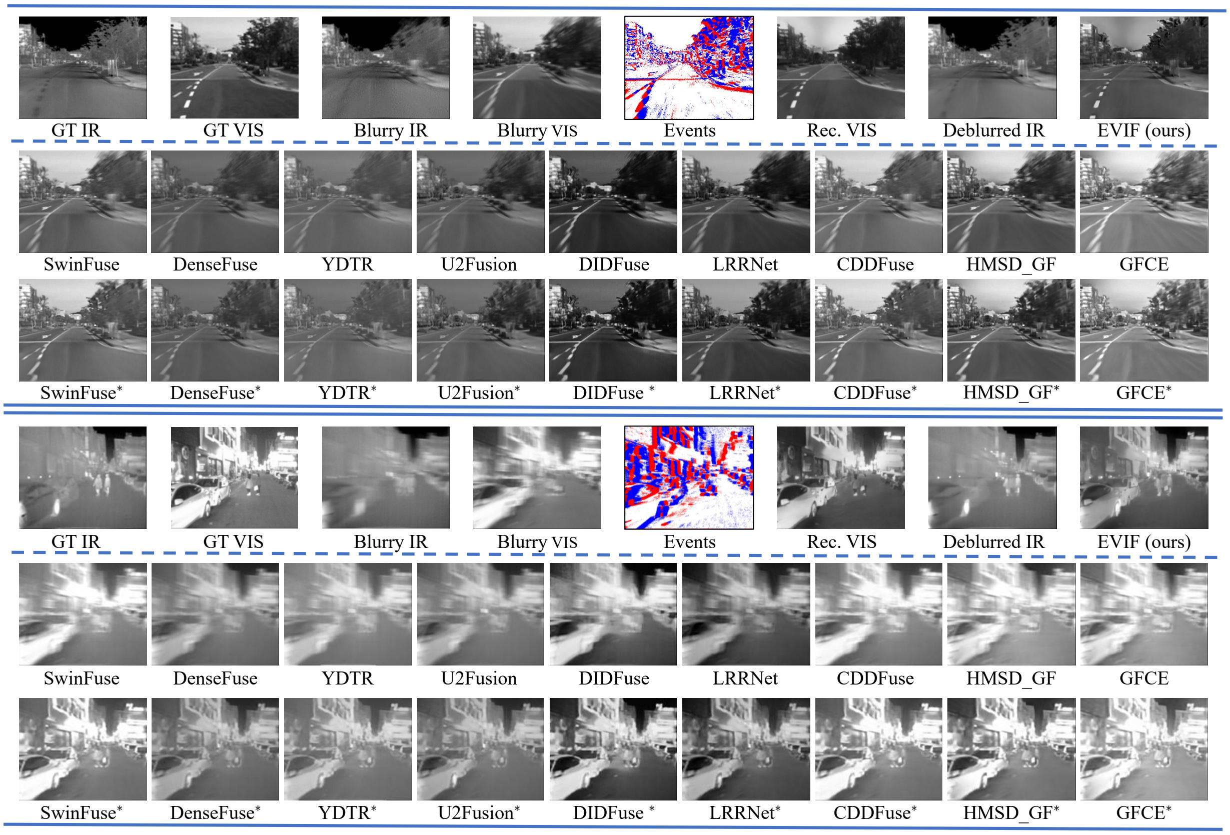

Figure 5. Qualitative comparison of EVIF with nine state-of-the-art frame-based VIF methods on synthetic data. For each test sample, we show in the first row the groundtruth and blurry images, the event data, and the three task outputs of EVIF. We then show fusion results of other methods in the next two rows. Methods with $*$ denotes applying NAFNet deblurring as a preprocessing step before fusion.

real captured data. For the synthetic data, we use the KAIST multi-spectral dataset [10]. This dataset encompasses a wide variety of driving environments and provides consecutive visible and infrared frames. To simulate blurry data to mimic highly dynamic scenes, we average every 7 frames to form one blurry infrared and visible frame. The events are generated using ESIM simulator [37]. For the real data, we build a hybrid camera system as shown in Fig. 6, which consists of a DAVIS346 event camera and an infrared camera. The data are then captured under various challenging scenes (*e.g.*, late night low-light environments and on-board captures on a high-speed vehicle). Since DAVIS346 can also output grayscale APS frames, we use them as visible images for comparison with conventional VIF methods.

4.2. Experimental Settings

Model Training. We use the training split of the KAIST dataset provided in [10] to train our model. For the event data, we use the standard event voxel representation [38] that splits and merges the event sequence into ten channels. Adam [19] optimizer is used with a learning rate of 1×10^{-4} and $[\gamma_1, \gamma_2, \gamma_3]$ in Eq. 6 are set to $[1.0, 0.1, 0.01]$. The training lasts for 100K, 20K and 10K iterations for the three stages on a server equipped with an RTX4090 24GB GPU.

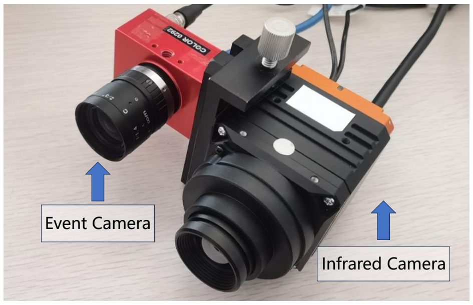

Figure 6. The hybrid camera system used to capture real data. The two cameras are temporally synchronized with a customized circuit and spatially aligned by manual keypoint matching.

Evaluation Protocols. To evaluate and compare methods, we apply six standard VIF metrics, including Cross Entropy (CE)[5], Entropy (EN) [40], Mutual Information (MI) [34], Average Gradient (AG) [11], Structural Similarity Index Measure (SSIM) [53], and Chen-Blum metric (Q_{CB}) [8]. For the synthetic data, we randomly select 1000 test samples from the test split of KAIST. During testing, the simulated blurry visible and infrared frames are directly used as input to conventional VIF methods, while the metrics are

Table 1. Quantitative comparison of EVIF with state-of-the-art conventional VIF methods. For each comparative method, we present two values: one for directly fusing blurry inputs and another for utilizing NAFNet deblurring as a preprocessing step prior to fusion. ↓ indicates smaller is better, and ↑ vice versa. The best results are indicated in **bold**, and the second best are indicated with an underline.

	CE↓	EN↑	MI↑	AG↑	SSIM↑	Q_{CB}↑
DenseFuse	2.321 / 2.302	6.568 / 7.034	1.416 / 1.526	1.423 / 1.574	1.274 / 1.283	0.251 / 0.359
YDTR	2.714 / 2.732	6.751 / 6.562	1.729 / 1.468	1.603 / 1.714	1.253 / 1.258	0.234 / 0.274
U2Fusion	1.678 / 1.696	6.772 / 6.716	1.391 / 1.741	1.498 / 1.688	1.265 / 1.275	0.316 / 0.251
DIDFuse	**1.462** / 1.492	7.103 / 6.775	1.511 / 1.445	2.244 / 2.565	1.041 / 1.064	0.369 / 0.338
LRRNet	2.173 / 2.289	7.245 / 7.122	1.740 / 1.535	2.105 / 2.285	1.202 / 1.216	0.356 / **0.382**
SwinFuse	2.328 / 2.392	7.217 / 7.148	1.489 / 1.767	2.267 / 2.542	1.245 / 1.250	0.285 / 0.349
CDDFuse	2.431 / 2.444	7.211 / 7.192	1.930 / 1.939	2.238 / 2.397	1.245 / 1.265	0.282 / 0.304
HMSD_GF	2.298 / 1.865	7.300 / 7.160	1.416 / 1.881	3.128 / 3.069	1.134 / 1.224	0.330 / 0.300
GFCE	2.650 / 2.586	7.314 / 7.267	1.396 / 1.561	3.325 / **3.645**	1.086 / 1.087	0.327 / 0.366
EVIF (ours)	1.986	**7.326**	**1.978**	2.120	**1.285**	0.377

calculated against the groundtruth sharp images. To further ensure a fair comparison, we additionally test these VIF methods by first using a powerful state-of-the-art frame-based deblurring network NAFNet [7] trained on KAIST data to deblur their input. The real data are directly tested using the model trained on synthetic data without any further finetuning. Due to the lack of groundtruth, we only perform qualitative comparisons on the real data.

4.3. Comparison with State-of-the-art Methods

We first compare our approach on the synthetic KAIST dataset with nine state-of-the-art VIF methods, including SwinFuse [54], DenseFuse [20], YDTR [43], U2Fusion [56], DIDFuse [63], LRRNet [23], CDDFuse [64], HMSD_GF [66] and GFCE [66]. The quantitative results are given in Table 1. Our EVIF system obtains three best values, in which two of them are information theory-based metrics (EN and MI). This suggests that the fused images generated by EVIF contain more information about the input, indicating the effectiveness of our mutual-information optimization mechanism. Moreover, the qualitative results in Fig. 5 also clearly demonstrate the advantage of EVIF. From the first sample in the figure, we see that conventional VIF systems tend to obtain blurry results even with strong input pre-processing (*i.e.*, deblurring with NAFNet). In addition, the unique sampling mechanism of events also makes EVIF suffer less from overexposure issues, as shown by the second sample.

To demonstrate the advantage of our approach under practical scenarios, we further compare different methods on the real-captured data. The qualitative results are given in Fig. 7. We can see that blurry and over/under exposure issue is even more pronounced in the real setting, which severely jeopardizes the performance of conventional VIF methods. Conversely, our EVIF system exhibits superior resilience against these issues, maintaining clarity and detail in the fused imagery. The results in Fig. 7 clearly demonstrate how our approach effectively mitigates the challenges in extreme lighting and high dynamic motion scenarios, offering a significant advancement in VIF technology.

Table 2. Comparison of various infrared image deblurring methods. The cross-task event enhancement notably enhances results.

Method	Blurry	NAFNet	EVIF w/o Enhance	EVIF
PSNR	28.750	31.230	33.320	33.940

4.4. Ablation Study

Effectiveness of Cross-task Event Enhancement. We first perform ablation study to verify the impact of the cross-task event enhancement method to the infrared image deblurring performance. Speficially, we report the deblurring PSNR values of different methods on the synthetic data in Table 2. From the table, we have a few observations. First, events can indeed provide valuable motion cues to facilliate infrared image deblurring, as evidenced by the large performance gap between EVIF and frame-based NAFNet. Second, our cross-task event enhancement method can further boost the performance of the deblurring task. This demonstrates the effectiveness of our cross-task collaborative design, which utilizes the complementary feature across tasks to promote model robustness. To provide more intuitive evidence, we demonstrate some data samples produced by different methods in Fig. 8. As shown in the figure, the sample generated with cross-task event enhancement exhibits richer details, thanks to the additional texture features drawn from the event reconstruction task.

Effectiveness of Bi-level MI Optimization. The bi-level min-max mutual information (MI) optimization in EVIF aims to highlight the complementary information reside in both visible and infrared modalities while reducing information loss. To verify the rational of such design, we provide a qualitative comparison of different MI optimization settings in Fig. 9. As shown in the figure, if no MI optimization is used, the resultant fused image tend to appear like the average of two input images. When only MI minimization is applied, the quality of the fused image become worse due to the potential information loss caused by forcing smaller MI loss. When only MI maximization is applied, the network tend to retain as much information in

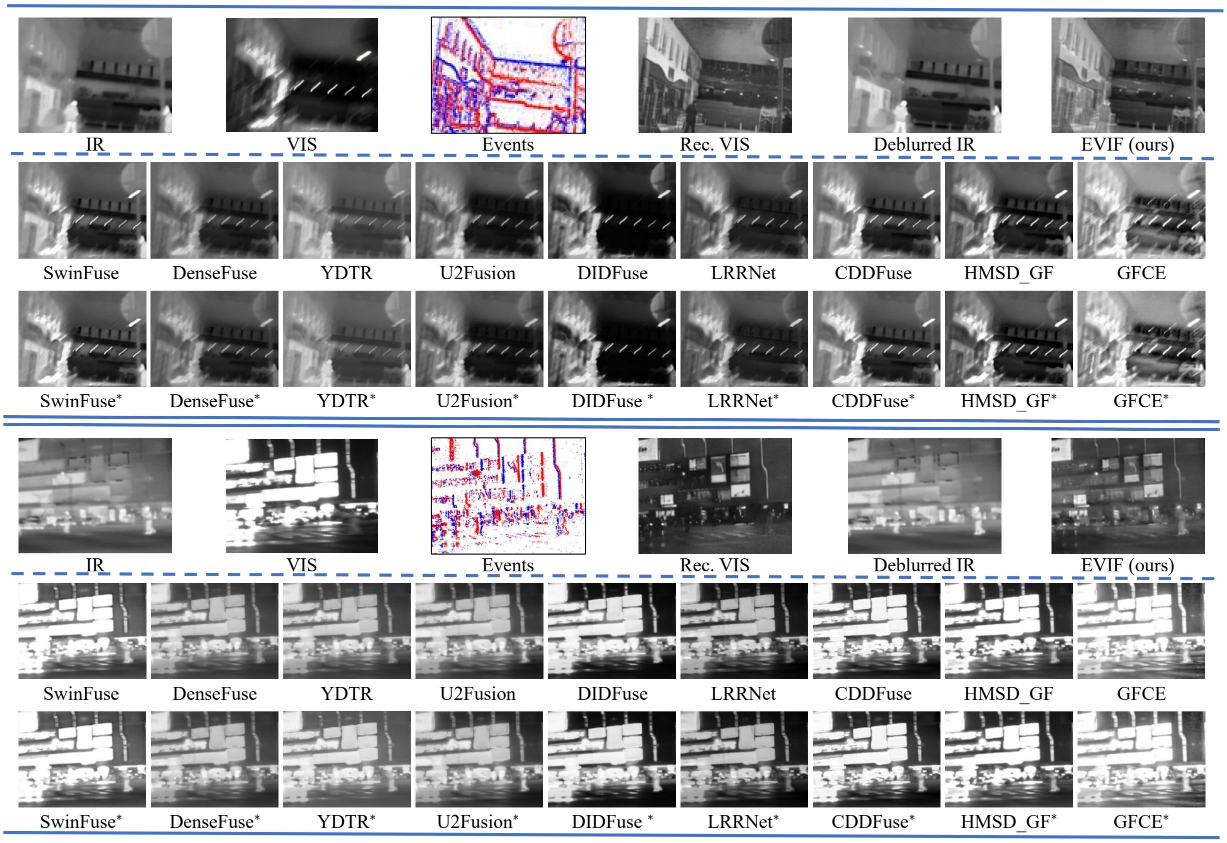

Figure 7. Qualitative comparison on the real-captured data. Please read similarly as in Fig. 5.

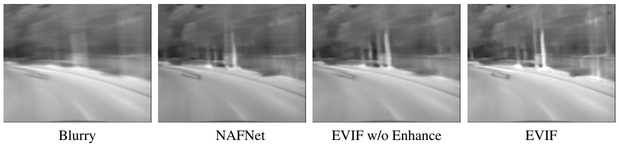

Figure 8. Qualitative assessment of the effectiveness of the cross-task event enhance method.

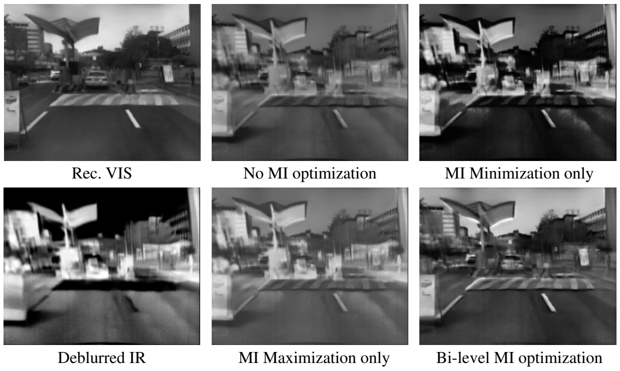

Figure 9. Qualitative assessment of the effectiveness of different MI optimization settings.

the fused image as possible, but this increases redundancy and making salient targets less perceivable. Different from these settings, the proposed bi-level MI optimization clearly enhances the modality-inconsistent regions, while keeps a comprehensive descripsion of the scene.

5. Conclusion

In this paper, we proposed a novel Event-based Visible and Infrared Fusion (EVIF) system. Characterized by the low latency and high dynamic range of event cameras, EVIF is able to handle blurry and over/underexposure issues in extreme lighting and high dynamic motion scenarios. Moreover, we developed a multi-task collaborative framework to obtain robust fusion results from EVIF. Benefiting from the cross-task event enhancement and the bi-level mutual information optimization, our framework can make the most of the event data and provide a comprehensive scene description. Extensive experiments on both synthetic and real data demonstrate that EVIF can effectively handle more extreme conditions than conventional VIF systems, ensuring clearer and more reliable fusion results.

6. Acknowledgment

This work is supported by the National Natural Science Foundation of China under grants No. 62027804, No. 62088102, and No. 62302041.

References

[1] On variational bounds of mutual information, ICML19. 5

[2] TNO image fusion dataset. https://figshare.com/articles/dataset/TNO_Image_Fusion_Dataset/1008029. 2

[3] Videos Analytics Dataset. https://www.ino.ca/en/technologies/video-analytics-dataset/videos/. 2

[4] Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville. Delving deeper into convolutional networks for learning video representations. *Int. Conf. Learn. Represent. (ICLR)*, 2016. 4

[5] D.M. Bulanon, T.F. Burks, and V. Alchanatis. Image fusion of visible and thermal images for fruit detection. *Biosyst. Eng.*, 103(1):12–22, 2009. 2, 6

[6] Xia Chang, Licheng Jiao, Fang Liu, and Fangfang Xin. Multicontourlet-based adaptive fusion of infrared and visible remote sensing images. *IEEE Geosci. Remote Sens. Lett.*, 7 (3):549–553, 2010. 1

[7] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, page 17–33, 2022. 7

[8] Yin Chen and Rick S Blum. A new automated quality assessment algorithm for image fusion. *Image Vis Comput.*, 27 (10):1421–1432, 2009. 6

[9] Jingchun Cheng, Yi-Hsuan Tsai, Shengjin Wang, and Ming-Hsuan Yang. Segflow: Joint learning for video object segmentation and optical flow. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 686–695, 2017. 4

[10] Yukyung Choi, Namil Kim, Soonmin Hwang, Kibaek Park, Jae Shin Yoon, Kyounghwan An, and In So Kweon. Kaist multi-spectral day/night data set for autonomous and assisted driving. *IEEE trans. Intell. Transp. Syst.*, 19(3):934–948, 2018. 6, 1

[11] Guangmang Cui, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition. *Opt. Commun.*, 341(15):199–209, 2015. 6

[12] 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. *Int. Conf. Learn. Represent. (ICLR)*, 2021. 5

[13] Zhizhong Fu, Xue Wang, Jin Xu, Ning Zhou, and Yufei Zhao. Infrared and visible images fusion based on rpca and nsct. *Infrared Phys. Technol.*, 77(1):114–123, 2016. 2

[14] Rikke Gade and Thomas B. Moeslund. Thermal cameras and applications: a survey. *Mach. Vis. Appl.*, 25(1):245–262, 2014. 2

[15] G. Gallego, T. Delbruck, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis, and D. Scaramuzza. Event-based vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(01):154–180, 2022. 3

[16] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1037–1045, 2015. 2

[17] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 3496–3504, 2021. 2

[18] Haiyan Jin, Licheng Jiao, Fang Liu, and Yutao Qi. Fusion of infrared and visual images based on contrast pyramid directional filter banks using clonal selection optimizing. *Opt. Eng.*, 47(2):027002, 2008. 2

[19] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Int. Conf. Learn. Represent. (ICLR)*, 2015. 6

[20] Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE Trans. Image Process.*, 28 (5):2614–2623, 2019. 1, 2, 7

[21] Hongwei Li, Binhai Wang, and Li Li. Research on the infrared and visible power-equipment image fusion for inspection robots. In *Int. Conf. Appl. Robot. Power Ind.*, pages 1–5, 2010. 1

[22] Hui Li, Xiao-Jun Wu, and Josef Kittler. Infrared and visible image fusion using a deep learning framework. In *Int. Conf. Pattern Recogn.*, pages 2705–2710, 2018. 2

[23] Hui Li, Tianyang Xu, Xiao-Jun Wu, Jiwen Lu, and Josef Kittler. Lrrnet: A novel representation learning guided fusion network for infrared and visible images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023. 7

[24] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128 120 db 15 μs latency asynchronous temporal contrast vision sensor. *IEEE J. Solid-State Circuits*, 43(2):566–576, 2008. 2, 3

[25] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 5792–5801, 2022. 2

[26] Yu Liu, Xun Chen, Rabab K. Ward, and Z. Jane Wang. Image fusion with convolutional sparse representation. *IEEE Signal Process. Lett.*, 23(12):1882–1886, 2016. 2

[27] Yu Liu, Xun Chen, Juan Cheng, Hu Peng, and Zengfu Wang. Infrared and visible image fusion with convolutional neural networks. *Int. J. Wavelets Multiresolut. Inf. Process.*, 16(03): 1850018, 2018. 2

[28] Jiayi Ma, Chen Chen, Chang Li, and Jun Huang. Infrared and visible image fusion via gradient transfer and total variation minimization. *Inf. Fusion.*, 31:100–109, 2016. 1

[29] Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: A survey. *Inf. Fusion.*, 45: 153–178, 2019. 1, 2

[30] Jiayi Ma, Wei Yu, Pengwei Liang, Chang Li, and Junjun Jiang. Fusiongan: A generative adversarial network for infrared and visible image fusion. *Inf. Fusion.*, 48:11–26, 2019. 2

[31] Jiayi Ma, Linfeng Tang, Meilong Xu, Hao Zhang, and Guobao Xiao. Stdfusionnet: An infrared and visible image fusion network based on salient target detection. *IEEE Trans. Instrum. Meas.*, 70(1):1–13, 2021. 2

[32] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5

[33] Liyuan Pan, Cedric Scheerlinck, Xin Yu, Richard Hartley, Miaomiao Liu, and Yuchao Dai. Bringing a blurry frame alive at high frame-rate with an event camera. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 6820–6829, 2019. 4

[34] Guihong Qu, Dali Zhang, and Pingfan Yan. Information measure for performance of image fusion. *Electron. Lett.*, 38(7):1, 2002. 6

[35] Manikandasriram Srinivasan Ramanagopal, Zixu Zhang, Ram Vasudevan, and Matthew Johnson Roberson. Pixel-Wise Motion Deblurring of Thermal Videos. In *Proc. Robot. Sci. Syst.*, 2020. 2

[36] Dongyu Rao, Tianyang Xu, and Xiao-Jun Wu. Tgfuse: An infrared and visible image fusion approach based on transformer and generative adversarial network. *IEEE Trans. Image Process.*, 2023. 2

[37] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. Esim: an open event camera simulator. In *Conf. Robot. Learn. (CoRL)*, pages 969–982, 2018. 6

[38] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 3852–3861, 2019. 3, 6

[39] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019. 4, 1

[40] J Wesley Roberts, Jan A Van Aardt, and Fethi Babikker Ahmed. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *J. Appl. Remote Sens.*, 2(1):023522, 2008. 6

[41] Aditya Sanghi. Info3d: Representation learning on 3d objects using mutual information maximization and contrastive learning. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pages 626–642, 2020. 5

[42] Lei Sun, Christos Sakaridis, Jingyun Liang, Qi Jiang, Kailun Yang, Peng Sun, Yaozu Ye, Kaiwei Wang, and Luc Van Gool. Event-based fusion for motion deblurring with cross-modal attention. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pages 412–428. 2022. 4, 1

[43] Wei Tang, Fazhi He, and Yu Liu. Ydtr: Infrared and visible image fusion via y-shape dynamic transformer. *IEEE Trans. Multimedia*, 2022. 7

[44] A. Toet, L. J. van Ruyven, and J. M. Valeton. Merging Thermal And Visual Images By A Contrast Pyramid. *Opt. Eng.*, 28(7):287789, 1989. 1

[45] Atousa Torabi, Guillaume Massé, and Guillaume-Alexandre Bilodeau. An iterative integrated framework for thermal–visible image registration, sensor fusion, and people tracking for video surveillance applications. *Comput. Vis. Image Underst.*, 116(2):210–221, 2012. 1

[46] Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. *Int. Conf. Learn. Represent. (ICLR)*, 2019. 5

[47] Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Mti-net: Multi-scale task interaction networks for multi-task learning. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pages 527–543. Springer, 2020. 4

[48] Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Multi-task learning for dense prediction tasks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(7): 3614–3633, 2022. 3

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. 4

[50] Vibashan Vs, Jeya Maria Jose Valanarasu, Poojan Oza, and Vishal M. Patel. Image fusion transformer. In *Proc. IEEE Int. Conf. Image Process. (ICIP)*, pages 3566–3570, 2022. 2

[51] Lin Wang, I S Mohammad Mostafavi, Yo-Sung Ho, and Kuk-Jin Yoon. Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 10073–10082, 2019. 3

[52] Rui Wang and Linfeng Du. Infrared and visible image fusion based on random projection and sparse representation. *Int. J. Remote Sens.*, 35(5):1640–1652, 2014. 2

[53] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4): 600–612, 2004. 6

[54] Zhishe Wang, Yanlin Chen, Wenyu Shao, Hui Li, and Lei Zhang. Swinfuse: A residual swin transformer fusion network for infrared and visible images. *IEEE Transactions on Instrumentation and Measurement*, 71:1–12, 2022. 7

[55] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based video reconstruction using transformer. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2021. 4

[56] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(1):502–518, 2020. 7

[57] Yong Yang, Jiaxiang Liu, Shuying Huang, Weiguo Wan, Wenying Wen, and Juwei Guan. Infrared and visible image fusion via texture conditional generative adversarial network. *IEEE Trans. Circ. Syst. Video Technol.*, 31(12):4771–4783, 2021. 2

[58] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Xin Yu, Yiran Zhong, Nick Barnes, and Ling Shao. Rgb-d saliency detection via cascaded mutual information minimization. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 4338–4347, 2021. 5

[59] Xingchen Zhang and Yiannis Demiris. Visible and infrared image fusion using deep learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(8):10535–10554, 2023. 1, 2

[60] Xiang Zhang, Wei Liao, Lei Yu, Wen Yang, and Gui-Song Xia. Event-based synthetic aperture imaging with a hybrid network. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 14230–14239, 2021. 3

[61] Yu Zhang and Qiang Yang. An overview of multi-task learning. *Nat. Sci. Rev.*, 5(1):30–43, 2017. 3

[62] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Trans. Knowl. Data Eng.*, 34(12):5586–5609, 2022. 3

[63] Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, Pengfei Li, and Jiangshe Zhang. DIDFuse: Deep image decomposition for infrared and visible image fusion. In *Int. Joint Conf. Artif. Intell. (IJCAI)*, pages 970–976, 2020. 7

[64] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 5906–5916, 2023. 7

[65] Man Zhou, Keyu Yan, Jie Huang, Zihe Yang, Xueyang Fu, and Feng Zhao. Mutual information-driven pan-sharpening. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1788–1798, 2022. 5

[66] Zhiqiang Zhou, Mingjie Dong, Xiaozhu Xie, and Zhifeng Gao. Fusion of infrared and visible images for night-vision context enhancement. *Appl. Opt.*, 55(23):6480–6490, 2016. 7