Asymmetric Hierarchical Difference-aware Interaction Network for Event-guided Motion Deblurring

Wen Yang[1,2], **Jinjian Wu**[1,2 *], **Leida Li**[1], **Weisheng Dong**[1], **Guangming Shi**[1,2]

[1]School of Artificial Intelligence, Xidian University, Xi'an 710071, China
[2]Pazhou Lab, Huangpu, 510555, China
yangwen@xidian.edu.cn, {jinjian.wu, ldli, wsdong, gmshi}@mail.xidian.edu.cn

Abstract

Event cameras are bio-inspired sensors that are capable of capturing motion information with high temporal resolution, which show potential in aiding image motion deblurring recently. Most existing methods indiscriminately handle feature fusion of two modalities with symmetric unidirectional/bidirectional interactions at different-level layers in feature encoder, while ignoring the different dependencies between cross-modal hierarchical features. To tackle these limitations, we propose a novel Asymmetric Hierarchical Difference-aware Interaction Network (AHDINet) for event-based motion deblurring, which explores the complementarity of two modalities with differential dependence modeling of cross-modal hierarchical features. Thereby, an event-assisted edge complement module is designed to leverage event modality to enhance the edge details of the image features in low-level encoder stage, and an image-assisted semantic complement module is developed to transfer contextual semantics of image features to event branch in high-level encoder stage. Benefiting from the proposed differentiated interaction mode, the respective advantages of image and event modalities are fully exploited. Extensive experiments on both synthetic and real-world datasets demonstrate that our method achieves state-of-the-art performance.

Code — https://github.com/wyang-vis/AHDINet

Introduction

Motion blur often occurs due to the fast motion of the object and/or the camera over the period of exposure time. Motion deblurring, a classical yet challenging problem, aims to restore sharp image from blurry input. It is an ill-posed inverse problem, due to the existence of many possible solutions. Traditional methods for accomplishing this task rely heavily on manually crafted image priors and various constraints (Bar et al. 2007; Cho, Wang, and Lee 2012; Bahat, Efrat, and Irani 2017; Kotera, Šroubek, and Milanfar 2013) , which limit the model capacity. With the development of deep neural network (DNN), DNN has also been applied for motion deblurring task to directly learn the relation from blurry images to sharp images under the supervision of a large-scale dataset of blurry-sharp image pairs (Nah, Hyun Kim, and Mu Lee 2017; Zamir et al. 2021; Park et al. 2020; Dong et al. 2023; Kong et al. 2023; Zhang, Xie, and Yao 2024; Liang et al. 2024). Although DNN-based deblurring methods have achieved promising improvements, they may fail to deal with severe blur due to the significant loss of motion information.

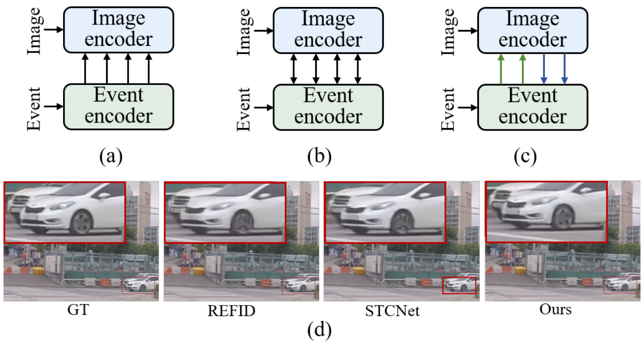


Figure 1: Cross-modal fusion strategies of existing event-based motion deblurring models: (a) Unidirectional interaction; (b) Symmetric bidirectional interaction; (c) Asymmetric bidirectional interaction (Ours). And (d): Some comparison of visualized deblurring results from REFID (Sun et al. 2023), STCNet (Yang et al. 2024), and our AHDINet.

Event cameras are bio-inspired sensors that can record per-pixel intensity changes asynchronously with high temporal resolution and output a stream of *events* encoding time, location and polarity of intensity changes (Gallego et al. 2020) if the intensity changes surpass a threshold. Due to the strong connection they possess with motion information, events have been used for motion deblurring recently. Early methods establish a mathematical event-based model mapping blurry frames to sharp frames (Pan et al. 2019; Scheerlinck, Barnes, and Mahony 2018) ,which are susceptible to sensor noise. Thus, some efforts utilize deep neural network to deal with noise corruption. Mainstream deep learning-based approaches are to model complementarity by elaborating cross-modal fusion schemes in feature encoder, which can be broadly classified into two categories: i) Unidirectional single/multi-level interaction shown

[*]Corresponding authors.

in Figure 1(a) , which uses the event information as auxiliary information to supplement the image modality at single or multiple-level layers of encoder networks (Chen et al. 2022; Sun et al. 2022, 2023; Cao et al. 2022). ii) Symmetric bidirectional single/multi-level interaction shown in Figure 1(b), which treats image and event cues equally to achieve cross-modality interaction at single or multiple-level layers of encoder networks (Yang et al. 2024; Chen and Yu 2024; Yang et al. 2022). However, these methods only adopt an indistinguishable way of integrating multilevel features of two modalities. They overlook a pivotal property that unique characteristics of low-level and high-level features have not been developed and effectively utilized to improve the performance of event-based motion deblurring.

In this paper, we reconsider the interaction patterns of the hierarchical features of the two modalities to fully explore the complementary integration of the them. Because of the differences in imaging principles, image data and event data have their own unique characteristics, even when capturing targets in the same scene. In highly dynamic scenes, event cameras can effectively capture low-level texture details but are insensitive to high-level semantics that are easy for humans to understand, while conventional frame-based cameras are prone to lose low-level texture details by motion blur but can well capture rich high-level semantic information. Hereby, the hierarchical features of the two modalities indicate different complementary directions between them. We argue that the interaction of the hierarchical features of the two modalities should take place in an independent and differential manner, i.e., low-level event features can help image features obtain clear edge texture, while high-level image features can further enrich semantics of event and refine background information. Our expectation is to construct asymmetric bi-directional hierarchical difference interactions (shown in Figure 1(c)) to model the complementary features of the different levels between the two modalities, taking full advantage of the respective strengths of the different modalities.

Based on the above analysis, we propose an asymmetric hierarchical difference-aware interaction network (AHDINet) for event-based motion deblurring, which considers the discrepant dependence in multi-level features of two modalities to conduct cross-modal complementary fusion. First, an event-assisted edge complement (EEC) module is designed to supplement the edge features of event to image in shallow feature encoding. EEC could adaptively control the reliably edge message passing based on event-driven confidence mask. Second, an image-assisted semantic complement (ISC) module is designed to supplement semantic information of image branch to the event features in deep feature encoding. ISC uses semantic enhancement and semantic injection based on channel-spatial attention for selective feature fusion. Thanks to this differentiated interaction mode, image and event can give full play to their respective advantages for motion deblurring. And our framework achieves state-of-the-art performance of event-based motion deblurring (some visual comparisons are shown in Figure 1(d)). The main contributions of our work are as follows.

- We propose a novel asymmetric hierarchical difference-aware interaction network (AHDINet) for event-based motion deblurring, which explores the complementarity of two modalities by modeling the different dependence in cross-modal hierarchical features. Our method outperforms previous state-of-the-art works.
- We design an event-assisted edge complement (EEC) module, containing an event-driven detail message passing controller supplemented by cross-modal channel attention, to adaptively transfer finer edges from events to image in low-level encoder.
- We develop an image-assisted semantic complement (ISC) module to assist event modality in capturing global context semantic attributes from image, by the semantic enhancement followed semantic injection with channel-spatial attention in high-level encoder.

Related Work

Image Deblurring

Image deblurring is a challenge task because it requires extracting clear latent images from blurred ones. Conventional methods for accomplishing this task rely heavily on manually crafted a priori and hypotheses (Cho, Wang, and Lee 2012; Hyun Kim and Mu Lee 2015; Bahat, Efrat, and Irani 2017; Kotera, Šroubek, and Milanfar 2013; Levin et al. 2009), which, although insightful, tend to limit their generalizability and representativeness. However, the emergence of deep neural networks (DNNs) has revolutionized image deblurring by implicitly discovering the intricate relationship between blurred and clear images. 1) Single-Stage Approaches. These methods (Zhang et al. 2020; Kupyn et al. 2018, 2019) strive to produce hyper-realistic images by leveraging sophisticated network architectures tailored for high-level vision tasks. 2) Multi-Stage Approaches. These techniques (Nah, Hyun Kim, and Mu Lee 2017; Tao et al. 2018; Zamir et al. 2021; Chen et al. 2021) break down the complexity problem into simpler, more manageable subtasks. By progressively restoring an image at multiple scales, these techniques can gradually reveal clearer, sharper versions. 3) Coarse-to-Fine Strategies. These methods (Park et al. 2020; Cho et al. 2021; Dong et al. 2023) can gradually restore a sharp image with multiple input images on different resolutions. 4) Attention Modules. To further improve performance, spatial and/or channel attention modules are also integrated (Suin, Purohit, and Rajagopalan 2020; Tsai et al. 2022; Purohit and Rajagopalan 2020; Liang et al. 2021; Kong et al. 2023; Zhang, Xie, and Yao 2024; Liang et al. 2024). These mechanisms allow the network to selectively attend to relevant information and filter noise and interference, thus perfecting the deblurring process.

Despite good performance, these learning-based deblurring methods fail to deal with severe blur, as deblurring is a highly ill-posed problem with infinite feasible solutions that cannot be trivially addressed from only the blur set of input.

Event-based Motion Deblurring

Event cameras provide visual information with low latency and with strong robustness against motion blur, which offers great potential for motion deblurring. Early works (Pan et al. 2019; Scheerlinck, Barnes, and Mahony 2018) succeeded in modeling relationships between a sharp image and a blurry image using the physical model-based formulation. Regrettably, there is inevitable noise in events due to the non-ideality of physical sensors (Zhang and Yu 2022), resulting in degraded performance.

In recent works, data-driven methods tackle above limitations by learning-based approaches (Lin et al. 2020). Some efforts have been directed towards designing more advanced architecture for better integration strategies. On the one hand, some methods take event as auxiliary information for image branch, forming the unidirectional interaction mode. Several approaches fuse single-level event features into image branches with simple integration strategies (Chen et al. 2022; Shang et al. 2021; Jiang et al. 2020; Wang et al. 2020). And to improve the effectiveness of cross-modal fusion, cross-modal attention modules applied at multiple levels of event features are designed to complement the image information (Sun et al. 2022, 2023; Cao et al. 2022). On the other hand, some researchers are accustomed to treat the two modalities as equal, forming undifferentiated bidirectional interaction mode. Several works design the symmetric single-level bidirectional interactions (Yang et al. 2023, 2024; Chen and Yu 2024) and multi-level bidirectional interactions (Yang et al. 2022) to model cross-modal complementarities. Moreover, some endeavors have been focused on addressing real-world scenarios (Cho et al. 2023; Zhang and Yu 2022; Xu et al. 2021; Zhang et al. 2023; Sun et al. 2023; Kim, Cho, and Yoon 2024), including challenges such as videos with unknown exposure time (Kim et al. 2022).

Crucially, there are different dependencies between the hierarchical features of the event modality and the image modality. Existing methods take unidirectional or bidirectional undifferentiated hierarchical feature interactions, failing to adequately model cross-modal complementarities. Unlike these works, we propose a method that combines complementary information between the two modalities in an asymmetric bidirectional hierarchical difference interaction manner.

Method

Problem Statement

Given a blurry image B and the corresponding event stream $E_T \triangleq \{(x_i, y_i, p_i, t_i)\}_{t_i \in T}$ containing all asynchronous events triggered during exposure time T, where $p = \pm 1$ is polarity, which denotes the direction (increase or decrease) of the intensity changes at that pixel (x, y) and time t, the proposed method is to recover a sharp image I by exploiting both blurry image B and event stream E_T, which can be modeled as $I = g_{\theta^*}(B, E_T)$, where g_{θ^*} is deep learning model.

Overall Framework

Figure 2 illustrates the overall framework of the proposed AHDINet, in which differential dependence in hierarchical features of image and event is considered for cross-modal complementary modeling. Specifically, our method first extracts target features f_b and f_e via two parallel backbones from blurry image and its corresponding events, separately. Next, we design an event-assisted edge complement (EEC) module to transfer detail supplement information from image modality to event modality in the low-level feature encoding stage (i.e., the first two ConvBlock of backbone), thereby enhancing the edge information of image features. Besides, we develop an image-assisted semantic complement (ISC) module to utilize the rich color appearance and global scene context of image to assist event branch in capturing fine-grained semantic attributes. Finally, the attention-based averaging module aims to optimally combine the event branch-based and image branch-based results. Below we detail the main parts: EEC and ISC module. More details about our AHDINet can be found in the supplement.

Event-assisted Edge Complement (EEC)

Compared with the blurry image, event modality contains complex texture information and can intuitively describe the shape and position of the moving objects. In this way, for the low-level encoder features that contain more detailed information (such as boundaries and edges), event features can provide more straightforward and instructive details than blurry image features. Thus, we design the EEC module hoping to transfer event features to image features at the low-level encoder to replenish them with missing edge information. To reliably complement information from event features to image features, an event-driven detail message passing controller (DMPC) is first proposed to control the edges passing between two modalities in the feature map, and then channel-wise attention is employed to further boost discriminative channels of controlled event features. Finally, the boosted event features are injected into the image features. The structure of the EEC is shown in top of Figure 2.

Detail Message Passing Controller (DMPC). Event data can accurately capture objects in degraded scenarios like high-speed motion or large dynamic range scenes, providing sharper edges. However, event data often contains various types of noise. To maximize the advantages of event data, we design an event-driven detail message passing controller to highlight reliable event features. The detail message passing controller is realized by an appropriate Gaussian blur to the event data, which helps eliminate noise while retaining essential information. Specifically, the detail message passing controller M_D can be represented as:

$$M_D = \mathcal{G} * (\mathcal{G} * E_T)\,, \tag{1}$$

where $\mathcal{G}$ is a Gaussian kernel with variance 3 and support 7×7; $*$ denotes the convolution operator, which is used to propagate the influence of the sparsely distributed events to their neighboring regions.

To guide the transfer of event edge features with a smaller spatial size (in the first two ConvBlock of backbone), the values of M_D is propagated according to the receptive field of each ConvBlock. Specifically,

$$M_D^{(0)} = M_D, \quad M_D^{(i)} = G_{K_i} * M_D^{(i-1)}, \tag{2}$$

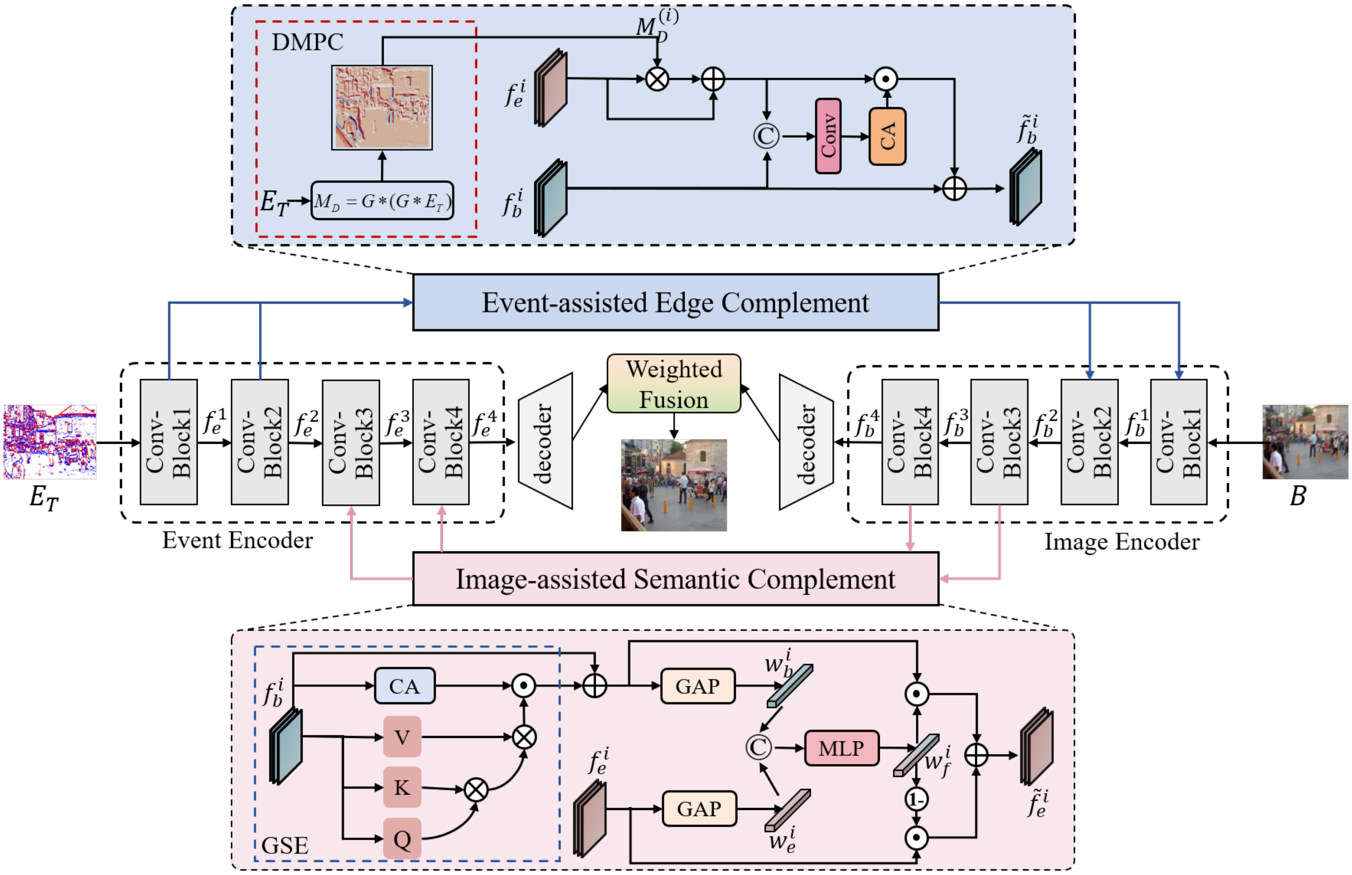


Figure 2: Framework of our AHDINet, containing two parts: event-assisted edge complement module and image-assisted semantic complement module.

where G_{K_i} denotes a Gaussian kernel with variance 3 and support $K_i \times K_i$ whose size is the same as the i-th ConvBlock.

Edge Injection. Based on the detail message passing controller M_D, as well as channel attention mechanism, the edge features extracted from the event branch are injected into the image branch, which can adaptively boost discriminative channels of event features, while suppressing those non-discriminative ones. More specifically, given the event features f_e^i and image features f_b^i from i-th ConvBlock, the $M_D^{(i)}$ is employed as the spatial edge prior to weigh the event features, thus obtaining the initial spatially-enhanced event features $\hat{f}_e^i$:

$$\hat{f}_e^i = f_e^i \otimes M_D^{(i)} \oplus f_e^i, \tag{3}$$

where $\otimes$ and $\oplus$ refer to the element-wise multiplication and element-wise addition.

On after of that, the channel weights of controlled event features $\hat{f}_e^i$ are calculated from the interaction of the two features $\hat{f}_e^i$ and f_b^i, and guides their fusion:

$$\widetilde{f}_b^i = CA(Conv(Cat(\hat{f}_e^i, f_b^i))) \odot \hat{f}_e^i \oplus f_b^i, \tag{4}$$

where $\widetilde{f}_b^i$ denotes the image features that incorporates event edge features, $CA(\cdot)$ refers to channel attention, $Conv(\cdot)$ is convolution operation, and $\odot$ is the channel-wise multiplication.

Image-assisted Semantic Complement (ISC)

For blurred image, despite the loss of local edge details, it still contains rich color appearance and global scene content, i.e., its global semantic information is also more comprehensive than the event modality. In the high-level layers of encoder stage, the learned features of the network contain more semantic information. Therefore, we design the ISC module in the high-level encoder stage to enrich the event semantic features with the help of the image modality. ISC requires highlighting meaningful high-level image semantics and suppressing distracting event high-level semantics. Inspired by the self-attention mechanism, we propose a global semantic enhancement (GSE) module with channel-spatial attention to enhance image semantic features and introduce a semantic injection module to transfer image features into event branch. The detailed ISC architecture is shown in the bottom of Figure 2.

Global Semantic Enhancement (GSE). To enhance image semantic features, we use channel-spatial attention to model the spatial information of the global context to extract rich semantic information, and weight the channel information of the global context to decrease information loss in the channels.

Specifcally, channel-spatial attention has two branches. The first branch is the channel-attention branch that weights the channel information in the global context. Given the image features f_b^i from i-th ConvBlock (i.e., the last two ConvBlock of backbone), the channel attention module is used to generate the channel weight w_c^i of f_b^i. This process can be expressed as:

$$w_c^i = Sig(Avg(Conv(f_b^i))), \tag{5}$$

where $Avg(\cdot)$ is the global average pooling, $Sig(\cdot)$ denotes sigmoid function, $Conv(\cdot)$ refers to the convolution layer.

The second branch is spatial-attention branch, which uses the powerful long-range dependencies modelling capability of self-attention to model the global context to enhance global semantic information. Specifically, the image features f_b^i are first transformed into Query Q, Key K, and Value V, then spatial attention maps w_s^i of f_b^i can be calculated as:

$$w_s^i = Softmax\left(QK^T\right). \tag{6}$$

Finally, we use channel-space attention to weight and globally model image features to enhance semantic information:

$$\hat{f_b^i} = V \otimes w_s^i \odot w_c^i \oplus f_b^i, \tag{7}$$

where $\hat{f_b^i}$ denotes semantic-enhanced image features.

Semantic Injection. Different channels in features maps usually represent different semantics. We expect to highlighting meaningful high-level image semantics and suppressing distracting event high-level semantics. And channel attention could select and reweight the channels of each modality features to improve feature representation. In this step, we design cross-modal interaction module based on channel attention to adaptively enhance or suppress bimodal channel semantic features and remix them into new event high-level semantic features.

Specifcally, we first obtain the channel weights of each modality from the concatenated multi-modal features, then adjust the channelwise relationships of each modality features, and finally aggregate these features into cross-modal features by weighted summation. This process can be expressed as:

$$\begin{aligned} w_f^i &= MLP(Cat(GAP(\hat{f_b^i}), GAP(f_e^i))), \\ \widetilde{f_e^i} &= \hat{f_b^i} \odot w_f^i \oplus f_e^i \odot (1 - w_f^i) \end{aligned} \tag{8}$$

where $\widetilde{f_e^i}$ refers to semantic-enhanced event features, $MLP(\cdot)$ denotes two linear units with Sigmoid activation function, $GAP(\cdot)$ denotes the global average pooling operation.

Loss Function

In this paper, we use the Charbonnier loss (Charbonnier et al. 1994) to train our network in an end-to-end fashion:

$$L_{\text{char}} = \frac{1}{CHW}\sqrt{\|I - G\|^2 + \varepsilon^2}, \tag{9}$$

where I and G is deblurred out and ground truth, respectively, C, H, W are dimensions of frame, and constant ε is empirically set to 10^{-3} as in (Zamir et al. 2021).

Experiments

Experimental Settings

Datasets. Our AHDINet is evaluated on 1) Synthetic dataset. *GoPro* (Nah, Hyun Kim, and Mu Lee 2017) and *DVD* (Su et al. 2017) datasets are widely adopted for image-only and event-based deblurring such as (Sun et al. 2022), which contains synthetic blurring images and sharp clear ground-truth images, as well as synthetic events generated by simulation algorithm ESIM (Rebecq, Gehrig, and Scaramuzza 2018). 2) Authentic dataset. *REBlur* (Sun et al. 2022) is a genuine event deblurring dataset collected by DAVIS, with an image resolution of 360 × 260. This dataset comprises 1,389 sample pairs encompassing diverse 12 distinct types of linear and nonlinear motions, for three different moving patterns and the camera itself. Among these samples, there are 486 training sets and 903 test sets. The scenes depicted in this dataset accurately capture regular indoor movements.

Implementation Details. Our method is implemented using Pytorch on NVIDIA RTX 3090 GPU. The size of training patch is 300 × 300 with minibatch size of 8. The optimizer is ADAM (Kingma and Ba 2015), and the learning rate is initialized at 2×10^{-4} and decreased by the cosine learning rate strategy with a minimum learning rate of 10^{-6}. For data augmentation, each patch is horizontally flipped with the probability of 0.5. The training ends after 200k iterations for *GoPro* dataset and 100k iterations for *REBlur* dataset. The Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index (SSIM) are adopted as the evaluation metrics.

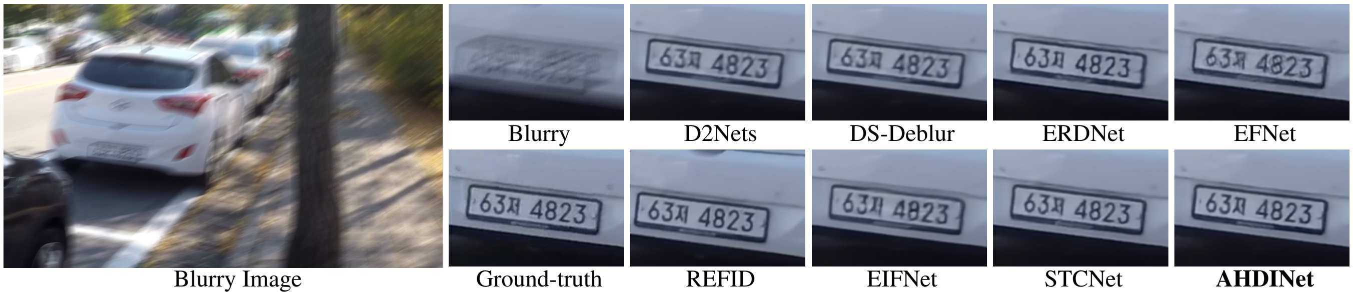
We compare the proposed AHDINet to state-of-the-art event-based deblurring methods, including RED (Xu et al. 2021), eSL-Net (Wang et al. 2020), D2Nets (Shang et al. 2021), LEBMD (Jiang et al. 2020), EVDI (Zhang and Yu 2022), DS-Deblur (Yang et al. 2022), MADANET+ (Yang and Yamac 2022), ERDNet (Chen et al. 2022), EFNet (Sun et al. 2022), REFID (Sun et al. 2023), EIFNet (Yang et al. 2023), STCNet (Yang et al. 2024) and FAEVD (Kim, Cho, and Yoon 2024).

GoPro: Table 1 reports the quantitative results on synthetic GoPro dataset. Compared to the best existing event-based deblurring methods, our method achieves outstanding performance improvements (0.39 dB improvement in terms of PSNR), showing the superiority of our symmetric bidirectional hierarchical difference-aware fusion strategies. We show in Figure 3 a visual comparison between our method and several state-of-the-art methods. Overall, visual quality comparisons demonstrate that our method can recover sharper texture details that are closer to the ground-truth, while the results restored by other methods still suffer from motion blur, losing sharp edge information. More visual comparisons are available in the supplement.

***DVD*:** To prove the generalization ability of the proposed AHDINet, cross-database experiments are conducted. The AHDINet is trained on *GoPro* dataset and tested on *DVD* dataset. Table 2 reports the quantitative results on the *DVD* dataset. Our method significantly outperforms other state-of-the-art competitors (0.4dB improvement in terms of

Method	RED	LEBMD	eSL-Net	EVDI	D2Nets	DS-Deblur	MADANET+
PSNR	28.98	29.67	30.23	30.40	31.76	33.13	33.84
SSIM	0.8499	0.9270	0.8703	0.9058	0.9430	0.9465	9640
Method	**ERDNet**	**EFNet**	**REFID**	**EIFNet**	**STCNet**	**FAEVD**	**AHDINet**
PSNR	34.25	35.46	35.91	35.99	36.45	36.70	**37.09**
SSIM	0.9534	0.9720	0.9730	0.9785	0.9809	0.9780	**0.9820**

Table 1: Comparison of event-based motion deblurring methods on *GoPro* dataset.


Blurry Image

Blurry

D2Nets

DS-Deblur

ERDNet

EFNet

Ground-truth

REFID

EIFNet

STCNet

AHDINet

Figure 3: Visual comparisons on *GoPro* datatset. Best viewed on a screen and zoomed in.

Method	D2Nets	eSL-Net	DS-Deblur	ERDNet	EFNet	REFID	STCNet	**AHDINet**
PSNR	26.64	27.50	31.63	32.29	32.85	33.15	33.94	**34.34**
SSIM	0.8819	0.8914	0.9436	0.9506	0.9571	0.9611	0.9692	**0.9712**

Table 2: Comparison of event-based motion deblurring methods on *DVD* dataset.

Method	D2Nets	eSL-Net	ERDNet	EFNet	REFID	STCNet	**AHDINet**
PSNR	35.10	35.50	37.98	38.12	38.34	38.98	**40.85**
SSIM	0.9621	0.9563	0.9506	0.975	0.9752	0.9820	**0.9900**

Table 3: Comparison of event-based motion deblurring methods on *REBlur* dataset.

PSNR over best event-based methods), demonstrating the superior generalization ability of the proposed framework.

REBlur: Real events dataset is more challenging than synthetic events dataset because the former has more noise than the latter due to the non-ideality of physical sensors. To further verify the superiority of our model, we compare the performances of the proposed AHDINet on real events dataset with other state-of-the-art competitors. Table 3 shows the quantitative comparison results on *REBlur* dataset. Note that for a fair comparison, we retrain several event-guided methods using the publicly available code provided by the authors. Compared with other methods, AHDINet achieves the best performance, outperforming the second best method by a remarkable improvement of 1.87dB in terms of PSNR. Moreover, we show the qualitative visual quality comparisons on *REBlur* dataset in Figure 4. We can see that the proposed method can achieve higher reconstruction quality and recover more details of the textures and edges than other methods.

EEC	ISC	Gropo	
		PSNR	SSIM
✗	✗	33.40	0.9615
✓	✗	36.50	0.9808
✗	✓	36.08	0.9788
✓	✓	**37.09**	**0.9820**

Table 4: Ablation study on EEC and ISC in AHDINet.

Ablation Study

To evaluate the effectiveness of the key components (EEC and ISC) in our model, we conduct ablation studies on *GoPro* dataset. A baseline is first experimented with, which simply concatenates image features f_b^4 and event features f_e^4. First row of Table 4 shows the performance of baseline.

Effectiveness of EEC and ISC module. First, we verify the effectiveness of EEC. We append it to *Baseline* to complement the event's edge texture features to the image. There

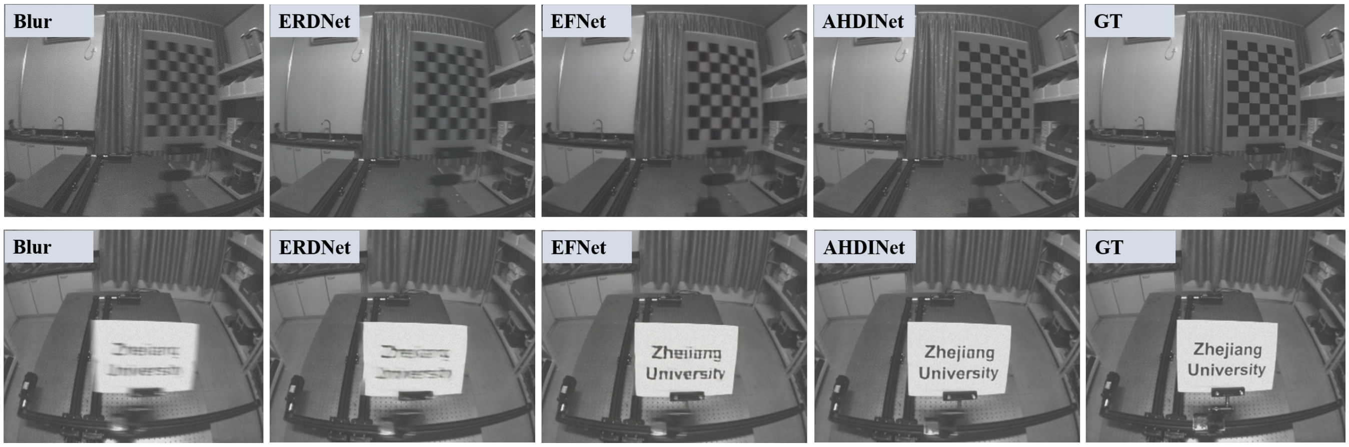


Figure 4: Visual comparisons on *REBlur* datatset. Best viewed on a screen and zoomed in.

DMPC	Gropo	
	PSNR	SSIM
✗	36.14	0.9791
✓	**36.50**	**0.9808**

Table 5: Ablation study on DMPC in EEC module.

GSE	Gropo	
	PSNR	SSIM
✗	35.76	0.9772
✓	**36.08**	**0.9788**

Table 6: Ablation study on GSE in ISC module.

Fusion Type	Gropo	
	PSNR	SSIM
UMI	35.95	0.9775
SBMI	36.61	0.9797
ABMI	**37.09**	**0.9820**

Table 7: Ablation study on asymmetric structure in AHDINet.

is a great performance gap in the first two rows of Table 4, which shows the effectiveness of the event modality to supplement the image modality with its edge features through the EEC module. Then we verify the effectiveness of ISC. We also directly delete the ISC module in our complete AHDINet. The results in the third row of Table 4 demonstrate the positive effect of the ISC module. Further, from this we can see the importance of low-level edge features for motion deblurring.

Ablation study on detail message passing controller (DMPC) in EEC module. Event data often contains various types of noise. The DMPC is designed to control the edge message passing from event branch to image branch. Thus, we validate the importance of DMPC strategy in EEC. We delete the DMPC module in EEC to allow uncontrolled transfer of spatial event features to the image branch. Table 5 shows that DMPC can effectively attenuate the interference of noise in the event.

Ablation study on global semantic enhancement (GSE) in ISC module. The GSE is designed to use channel-spatial attention to weight the semantic information of the global context for information enhancement. Thus, we validate the importance of GSE strategy in ISC. We delete the GSE module in ISC and do not enhance the image semantic information. Table 6 shows the effectiveness of GSE.

Ablation study on asymmetric structure in AHDINet. In our work, we consider the different dependence in cross-modal hierarchical features of two modalities to model the complementary fusion, and propose an asymmetric hierarchical difference-aware fusion method. In order to validate the basic idea of the proposed method, the different fusion strategies, i.e., unidirectional multi-level interaction (UMI), symmetric bidirectional multi-level interaction (SBMI), our asymmetric bidirectional multi-level interaction (ABMI), are verified. In this experiment, the fusion between cross-modal hierarchical features is accomplished with EEC or ISC. Table 7 summarizes the experimental results on different fusion type, demonstrating that our qualitative analysis is correct.

Conclusion

In this work, we propose an asymmetric hierarchical difference-aware interaction network (AHDINet) for event-based motion deblurring, which models the different dependence in hierarchical features of two modalities for complementary fusion. Specifically, we first employ an event-assisted edge complement (EEC) module in the early layers of the encoding network, which can supplement more detailed information of event to enhance the image features. Then, we design an image-assisted semantic complement (ISC) module in later layers of feature encoding to enhance the event features with image semantics. Both subjective and objective experiments on synthetic datasets and real-world datasets have demonstrated the effectiveness of our method.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.62402366), Fundamental Research Funds for the Central Universities (No.ZYTS24006, QTZX23038), China Postdoctoral Science Foundation (No.2024M752535) and Postdoctoral Fellowship Program of China Postdoctoral Science Foundation (No.GZC20241320).

References

Bahat, Y.; Efrat, N.; and Irani, M. 2017. Non-uniform blind deblurring by reblurring. In *ICCV*.

Bar, L.; Berkels, B.; Rumpf, M.; and Sapiro, G. 2007. A variational framework for simultaneous motion estimation and restoration of motion-blurred video. In *ICCV*.

Cao, C.; Fu, X.; Zhu, Y.; Shi, G.; and Zha, Z.-J. 2022. Event-driven Video Deblurring via Spatio-Temporal Relation-Aware Network. In *IJCAI*.

Charbonnier, P.; Blanc-Feraud, L.; Aubert, G.; and Barlaud, M. 1994. Two deterministic half-quadratic regularization algorithms for computed imaging. In *ICIP*.

Chen, H.; Teng, M.; Shi, B.; Wang, Y.; and Huang, T. 2022. A Residual Learning Approach to Deblur and Generate High Frame Rate Video With an Event Camera. *IEEE TMM*.

Chen, K.; and Yu, L. 2024. Motion Deblur by learning residual from events. *IEEE TMM*.

Chen, L.; Lu, X.; Zhang, J.; Chu, X.; and Chen, C. 2021. HINet: Half instance normalization network for image restoration. In *CVPR*.

Cho, H.; Jeong, Y.; Kim, T.; and Yoon, K.-J. 2023. Non-coaxial event-guided motion deblurring with spatial alignment. In *ICCV*.

Cho, S.; Wang, J.; and Lee, S. 2012. Video deblurring for hand-held cameras using patch-based synthesis. *ACM TOG*.

Cho, S.-J.; Ji, S.-W.; Hong, J.-P.; Jung, S.-W.; and Ko, S.-J. 2021. Rethinking Coarse-to-Fine Approach in Single Image Deblurring. In *ICCV*.

Dong, J.; Pan, J.; Yang, Z.; and Tang, J. 2023. Multi-scale residual low-pass filter network for image deblurring. In *ICCV*.

Gallego, G.; Delbrück, T.; Orchard, G.; Bartolozzi, C.; Taba, B.; Censi, A.; Leutenegger, S.; Davison, A. J.; Conradt, J.; Daniilidis, K.; et al. 2020. Event-based vision: A survey. *IEEE TPAMI*.

Hyun Kim, T.; and Mu Lee, K. 2015. Generalized video deblurring for dynamic scenes. In *CVPR*.

Jiang, Z.; Zhang, Y.; Zou, D.; Ren, J.; Lv, J.; and Liu, Y. 2020. Learning event-based motion deblurring. In *CVPR*.

Kim, T.; Cho, H.; and Yoon, K.-J. 2024. Frequency-aware Event-based Video Deblurring for Real-World Motion Blur. In *CVPR*.

Kim, T.; Lee, J.; Wang, L.; and Yoon, K.-J. 2022. Event-guided deblurring of unknown exposure time videos. In *ECCV*.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Kong, L.; Dong, J.; Ge, J.; Li, M.; and Pan, J. 2023. Efficient Frequency Domain-based Transformers for High-Quality Image Deblurring. In *CVPR*.

Kotera, J.; Šroubek, F.; and Milanfar, P. 2013. Blind deconvolution using alternating maximum a posteriori estimation with heavy-tailed priors. In *CAIP*.

Kupyn, O.; Budzan, V.; Mykhailych, M.; Mishkin, D.; and Matas, J. 2018. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *CVPR*.

Kupyn, O.; Martyniuk, T.; Wu, J.; and Wang, Z. 2019. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *ICCV*.

Levin, A.; Weiss, Y.; Durand, F.; and Freeman, W. T. 2009. Understanding and evaluating blind deconvolution algorithms. In *CVPR*.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. Swinir: Image restoration using swin transformer. In *ICCV*.

Liang, P.; Jiang, J.; Liu, X.; and Ma, J. 2024. Image deblurring by exploring in-depth properties of transformer. *IEEE TNNLS*.

Lin, S.; Zhang, J.; Pan, J.; Jiang, Z.; Zou, D.; Wang, Y.; Chen, J.; and Ren, J. 2020. Learning event-driven video deblurring and interpolation. In *ECCV*.

Nah, S.; Hyun Kim, T.; and Mu Lee, K. 2017. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*.

Pan, L.; Scheerlinck, C.; Yu, X.; Hartley, R.; Liu, M.; and Dai, Y. 2019. Bringing a blurry frame alive at high frame-rate with an event camera. In *CVPR*.

Park, D.; Kang, D. U.; Kim, J.; and Chun, S. Y. 2020. Multi-temporal recurrent neural networks for progressive non-uniform single image deblurring with incremental temporal training. In *ECCV*.

Purohit, K.; and Rajagopalan, A. 2020. Region-adaptive dense network for efficient motion deblurring. In *AAAI*.

Rebecq, H.; Gehrig, D.; and Scaramuzza, D. 2018. ESIM: an open event camera simulator. In *CoRL*.

Scheerlinck, C.; Barnes, N.; and Mahony, R. 2018. Continuous-time intensity estimation using event cameras. In *ACCV*.

Shang, W.; Ren, D.; Zou, D.; Ren, J. S.; Luo, P.; and Zuo, W. 2021. Bringing Events Into Video Deblurring With Non-Consecutively Blurry Frames. In *ICCV*.

Su, S.; Delbracio, M.; Wang, J.; Sapiro, G.; Heidrich, W.; and Wang, O. 2017. Deep video deblurring for hand-held cameras. In *CVPR*.

Suin, M.; Purohit, K.; and Rajagopalan, A. 2020. Spatially-attentive patch-hierarchical network for adaptive motion deblurring. In *CVPR*.

Sun, L.; Sakaridis, C.; Liang, J.; Jiang, Q.; Yang, K.; Sun, P.; Ye, Y.; Wang, K.; and Van Gool, L. 2022. Event-Based Fusion for Motion Deblurring with Cross-modal Attention. In *ECCV*.

Sun, L.; Sakaridis, C.; Liang, J.; Sun, P.; Cao, J.; Zhang, K.; Jiang, Q.; Wang, K.; and Van Gool, L. 2023. Event-Based Frame Interpolation with Ad-hoc Deblurring. In *CVPR*.

Tao, X.; Gao, H.; Shen, X.; Wang, J.; and Jia, J. 2018. Scale-recurrent network for deep image deblurring. In *CVPR*.

Tsai, F.-J.; Peng, Y.-T.; Tsai, C.-C.; Lin, Y.-Y.; and Lin, C.-W. 2022. BANet: A Blur-aware Attention Network for Dynamic Scene Deblurring. *IEEE TIP*.

Wang, B.; He, J.; Yu, L.; Xia, G.-S.; and Yang, W. 2020. Event enhanced high-quality image recovery. In *ECCV*.

Xu, F.; Yu, L.; Wang, B.; Yang, W.; Xia, G.-S.; Jia, X.; Qiao, Z.; and Liu, J. 2021. Motion Deblurring with Real Events. In *ICCV*.

Yang, D.; and Yamac, M. 2022. Motion Aware Double Attention Network for Dynamic Scene Deblurring. In *CVPRW*.

Yang, W.; Wu, J.; Li, L.; Dong, W.; and Shi, G. 2023. Event-based Motion Deblurring with Modality-Aware Decomposition and Recomposition. In *ACM MM*.

Yang, W.; Wu, J.; Ma, J.; Li, L.; Dong, W.; and Shi, G. 2022. Learning for Motion Deblurring with Hybrid Frames and Events. In *ACM MM*.

Yang, W.; Wu, J.; Ma, J.; Li, L.; and Shi, G. 2024. Motion Deblurring via Spatial-Temporal Collaboration of Frames and Events. In *AAAI*.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; Yang, M.-H.; and Shao, L. 2021. Multi-stage progressive image restoration. In *CVPR*.

Zhang, H.; Xie, H.; and Yao, H. 2024. Blur-aware Spatio-temporal Sparse Transformer for Video Deblurring. In *CVPR*.

Zhang, X.; and Yu, L. 2022. Unifying Motion Deblurring and Frame Interpolation with Events. In *CVPR*.

Zhang, X.; Yu, L.; Yang, W.; Liu, J.; and Xia, G.-S. 2023. Generalizing event-based motion deblurring in real-world scenarios. In *ICCV*.

Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; and Fu, Y. 2020. Residual dense network for image restoration. *IEEE TPAMI*.