A Coarse-To-Fine Fusion Network for Event-Based Image Deblurring

Huan Li[1,2] , **Hailong Shi**[1,*] and **Xingyu Gao**[1,*]
[1]Institute of Microelectronics, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
{lihuan, shihailong, gaoxingyu}@ime.ac.cn

Abstract

Event-driven image deblurring is an innovative approach involving the input of events obtained from the event camera alongside blurred frames to facilitate the deblurring process. Unlike conventional cameras, event cameras in event-driven imaging exhibit low-latency characteristics and are immune to motion blur, resulting in significant advancements in image deblurring. In this paper, we propose a pioneering event-based coarse-to-fine image deblurring network named CFFNet. In contrast to existing deblurring methods, our approach incorporates event data, generating multiple coarse frames from a single frame before further refining them into a sharp image. We introduce an Event Image Fusion Block (EIFB) for the coarse fusion of events and images, producing coarse frames at different time points. Additionally, we propose a Bidirectional Frame Fusion Block (BFFB) for the fine fusion of coarse frames. CFFNet effectively harnesses the spatiotemporal information of event data through a comprehensive fusion process from coarse to fine. Experimental results on the GoPro and REBlur datasets demonstrate that our method achieves state-of-the-art performance for image deblurring task.

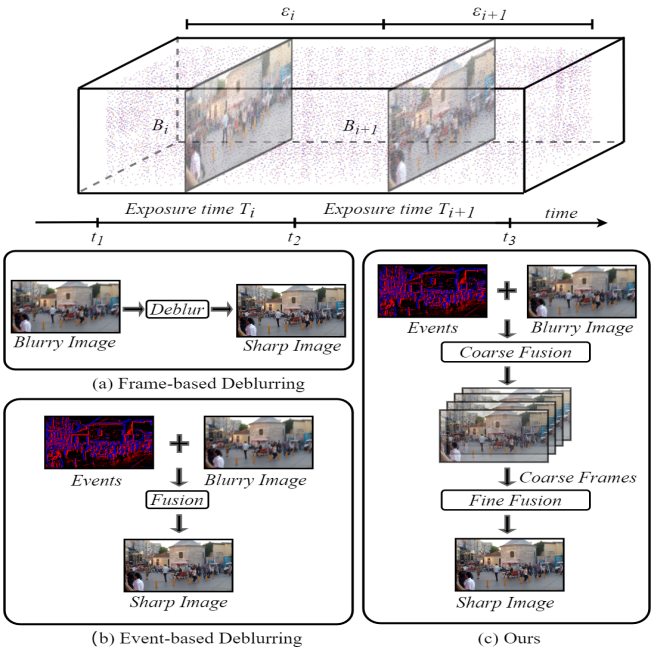

Figure 1: Event data with image illustrations and different deblurring methods. Blurry image B captured during the exposure time T, along with the corresponding event streams ε, where red/blue dots denote positive/negative events, respectively.

1 Introduction

Motion blur stands out as a prominent factor influencing image quality and the pursuit of reconstructing sharp images from blurred counterparts has long been a focal point in computer vision. In the early stage of image deblurring, techniques such as regularization constraints and blur kernel estimation were commonly employed to enhance the quality of image reconstruction [Kundur and Hatzinakos, 1996; Fergus *et al.*, 2006; Dai and Wu, 2008; Krishnan *et al.*, 2011; Wulff and Black, 2014]. However, these traditional approaches may prove ineffective in specific scenarios, particularly when dealing with complex blurs or significant levels of image blurring. In recent years, significant advancements have been achieved in image deblurring, thanks to the emergence of deep learning methods, particularly those grounded in convolutional neural networks [Tao *et al.*, 2018; Park *et al.*, 2020; Cho *et al.*, 2021; Chen *et al.*, 2021; Chen *et al.*, 2022b; Li *et al.*, 2023a; Zhang *et al.*, 2023a]. These methods have showcased impressive performance, excelling in handling complex blurs and large-scale datasets. Despite substantial progress compared to traditional methods, these deep learning approaches still encounter challenges in complex real-world scenarios, especially those involving high-speed movements.

As novel bio-inspired sensor, event cameras diverge in their approach to acquiring visual information compared to traditional cameras [Gallego *et al.*, 2020]. Unlike RGB cameras that sample images with a fixed exposure time, event cameras operate on event-driven principles, detecting and outputting information about changing pixels in real-time. The imaging

*Corresponding author

process of RGB cameras inevitably leads to the loss of some crucial motion information, while the high temporal resolution characteristic of event cameras compensates for potential additional information loss within the exposure time, making them more suitable for high-speed motion scenarios. However, given that event cameras capture asynchronous sparse signals, traditional computer vision methods are no longer suitable for processing such signals. While recent researches have proposed some event-based deblurring algorithm models [Lin *et al.*, 2020; Shang *et al.*, 2021; Cao *et al.*, 2022; Sun *et al.*, 2022; Sun *et al.*, 2023; Yang and Yamac, 2023] and have made great achievement. While they directly fuse events and images, resulting in the loss of a considerable amount of temporal information from event data. To overcome the limitations of previous researches, we introduce a coarse-to-fine deblurring network based on events, named CFFNet. This network takes blurry images and event data as input and outputs sharp images. We introduce a Event Image Fusion Block (EIFB) for the coarse fusion of images with events and a Bidirectional Frame Fusion Block (BFFB) for the fine fusion of coarse frames.

Specifically, the coarse fusion stage utilizes an encoder-decoder structure resembling UNet. This involves feeding blurry images and events into the encoder, extracting features and obtaining multi-scale information through downsampling. Employing cross-attention effectively fuses details from both modalities. The coarse fusion stage further branches into two paths to handle images and events separately. In the image branch, images act as query vectors, events provide key and value vectors, while the roles reverse in the event branch. The upsampled features from the decoder connect with the same-scale encoder output via a Channel Attention Block (CAB) [Zhang *et al.*, 2018]. After passing through a Spatial Attention Module (SAM) [Zamir *et al.*, 2021] and concatenation, the decoder generates multiple coarse frames. Transitioning to fine fusion stage, a stack of Bidirectional Frame Fusion Blocks (BFFB) iteratively facilitate the fusion of multiple frames in forward and backward propagation, enabling effective information integration across multiple frames. Ultimately, multiple UNets are employed to output high-quality sharp images. Rigorous evaluations on the GoPro dataset and the recently introduced REBlur dataset confirm our model's outstanding state-of-the-art performance.

In summary, the main contributions of this paper include:

- We introduce an event-based coarse-to-fine deblurring network, decomposing the image deblurring task into two stages. In the coarse fusion stage, we fuse blurred frames and event streams to generate multiple coarse frames. In the fine fusion stage, we merge these coarse frames to obtain sharp images.
- We propose a novel Event Image Fusion Block (EIFB) for cross-modal information fusion between events and images. Furthermore, we introduce a Bidirectional Frame Fusion Block (BFFB) to learn information from adjacent frames, enabling multi-frame fusion and comprehensive utilization of spatiotemporal information in event data.
- By demonstrating the effectiveness of our deblurring network on mainstream large-scale datasets, our model achieves state-of-the-art results for image deblurring task on the GoPro and REBlur datasets.

2 Related Work

2.1 Frame-Based Deblurring

Reconstructing sharp images from motion-blurred frames has been a long-standing research focus [Hyun Kim *et al.*, 2017; Pan *et al.*, 2016; Nah *et al.*, 2017; Nimisha *et al.*, 2017; Wan *et al.*, 2020; Park *et al.*, 2020]. Traditional deblurring methods often involve the utilization of blur kernels or regularization techniques, employing techniques such as deconvolution [Krishnan *et al.*, 2011], kernel estimation [Xu and Jia, 2010], etc. However, their effect is not ideal.

Recently, deep learning approaches have been employed to achieve improved deblurring results by training convolutional networks to learn the mapping from blurred to sharp images for image reconstruction. For instance, [Zamir *et al.*, 2021] divides the image reconstruction into three stages and utilizes the Supervised Attention Module (SAM) for supervision and efficient feature propagation. Another approach proposed by [Chen *et al.*, 2021] introduces a deep stacked hierarchical network that focuses on different scale features of blurry images and employs the UNet architecture as an encoder-decoder structure. Additionally, [Chen *et al.*, 2022b] introduces linear gating units to replace non-linear activation functions, maintaining the performance of image denoising while bringing nontrivial gains to image deblurring. Furthermore, [Li *et al.*, 2023a] introduces a group spatiotemporal shift module to obtain an extended effective receptive field. However, frame-based image deblurring methods have limitations in accurately capturing motion information in high-speed scenarios.

2.2 Event-Based Deblurring

[Pan *et al.*, 2019] proposed an Event-based Double Integral (EDI) deblurring model, establishing a mapping relationship from blurred image to sharp images with events. However, this model is susceptible to noise from event cameras and its deblurring performance is limited. Another approach [Zhang and Yu, 2022] introduces a self-supervised learning model that estimates event-based residuals to unify motion deblurring and frame interpolation. [Sun *et al.*, 2022] provides voxelized event frames and blurred images to the network by incorporating cross-attention to merge both modalities and introducing a mask to concentrate on the area where the events occurre. Furthermore, [Zhou *et al.*, 2023] utilizes event data to estimate optical flow information and guides image reconstruction through optical flow. While the introduction of optical flow accelerates model convergence through optical flow loss, it is prone to errors in optical flow calculation and increases the complexity of the process. Additionally, [Yang and Yamac, 2023] introduces a deblurring network based on LSTM, allowing a dynamic number of event frames and incorporating deformable convolutions to enhance feature extraction.

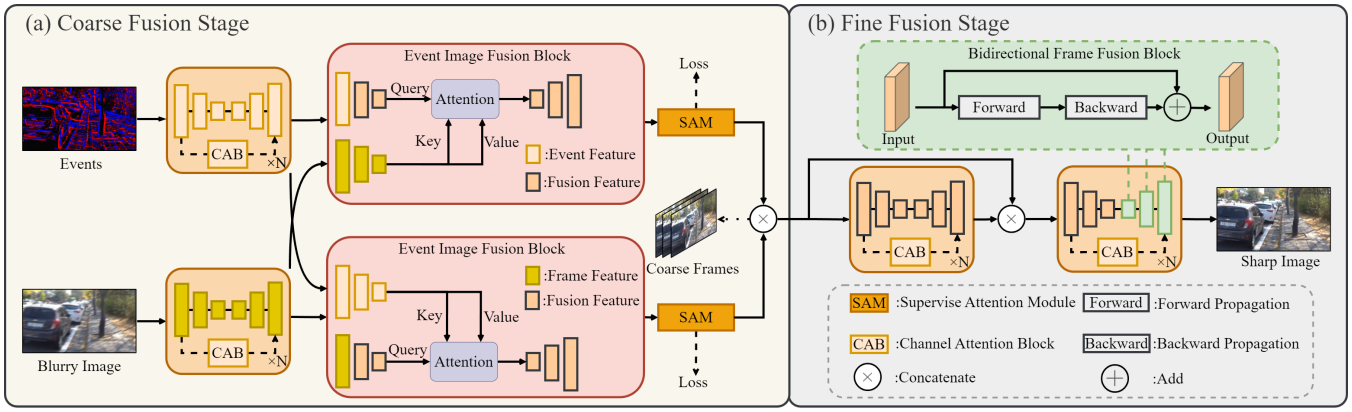

Figure 2: The overview of our method. (a) In coarse fusion stage, we combine the feature of image and events using EIFB to generate coarse frames. (b) In fine fusion stage, better results are achieved by integrating features from both preceding and succeeding frames using BFFB.

Most traditional event-based deblurring methods [Chen *et al.*, 2022a; Yang *et al.*, 2022; Yu *et al.*, 2023; Yang *et al.*, 2023] directly fuse all events and images, losing temporal information in event data, rendering their model performance unsatisfactory. To overcome the limitations of the above models, we designe a coarse-to-fine deblurring network from coarse to fine, mapping single image to multiple frames, then fusing them back into sharp image for improved deblurring results.

3 Problem Formulation and Analysis

Event cameras operate without fixed exposure times and respond asynchronously to changes in pixel logarithmic brightness. When a pixel's log intensity exceeds the contrast threshold c, the pixel generates an output in the form of (x, y, t, δ), where x and y respectively represent the horizontal and vertical coordinates of the pixel on an image. t records the time of the event and $\delta \in (-1, +1)$ indicates polarity, signifying the direction of the intensity change, as expressed below:

$$\delta = \begin{cases} +1, & if \quad \log \frac{I(t)}{I(t-\Delta t)} > c \\ -1, & if \quad \log \frac{I(t)}{I(t-\Delta t)} < c \end{cases} \tag{1}$$

The contrast threshold c is user-defined and may vary depending on different sensors. Here, $I(t)$ represents the image intensity of the point (x, y) at time t and Δt signifies the time interval between adjacent events, determined by the rate and magnitude of changes in pixel intensity. Based on Equation (1), we set $E(t, \hat{t}) = \int_t^{\hat{t}} \delta(s) ds$ to represent the event integral from t to $\hat{t}$ and can derive the relationship between $I(t)$ and $I(\hat{t})$:

$$I(\hat{t}) = I(t) \exp(c \cdot E(t, \hat{t})) \tag{2}$$

[Pan *et al.*, 2019] introduced the Event-based Double Integral (EDI) model to describe the relationship between the blurry image B and the latent sharp image L. If capturing starts at time s and continues for an exposure time T, the resulting blurry image B can be obtained by averaging the integral of the latent sharp image L over the time period T:

$$B = \frac{1}{T} \int_s^{s+T} L(t) dt \tag{3}$$

Handling the direct integration of continuous images can be challenging. Therefore, we assume that the blurry image B is composed of N images taken over a time period of T:

$$B = \frac{1}{N} \sum_{s \le t \le s+T} L(t) \tag{4}$$

If we consider a fixed reference time point r and exposure starting at time s, in conjunction with Equation (2) and Equation (4), we can obtain:

$$B = \frac{L(r)}{N} \sum_{s \le t \le s+T} \exp(c \cdot E(r, t)) \tag{5}$$

This leads to the mapping relationship between the blurry image B and the sharp image L:

$$L(r) = \frac{B \cdot N}{\sum_{s \le t \le s+T} \exp(c \cdot E(r, t))} \tag{6}$$

4 Method

In this section, we present an overview of our model's architecture, as outlined in Section 4.1. Our model is divided into two key phases: coarse fusion and fine fusion. Subsequently, in Section 4.2, we provide a detailed exploration of the coarse fusion stage, focusing on its main module, EIFB. Following that, in Section 4.3, we delve into the specifics of the fine fusion stage and its principal module, BFFB.

4.1 The Overall Architecture of CFFNet

Equation (6) has shown that by accumulating events over time, a mapping from blurred images to sharp images can be established. Based on this relationship, we designed a coarse-to-fine fusion network (CFFNet), which takes events and blurry image as input and outputs sharp images. Given a blurred image $B \in \mathbb{R}^{H \times W \times 3}$ and a set of events $E \in$

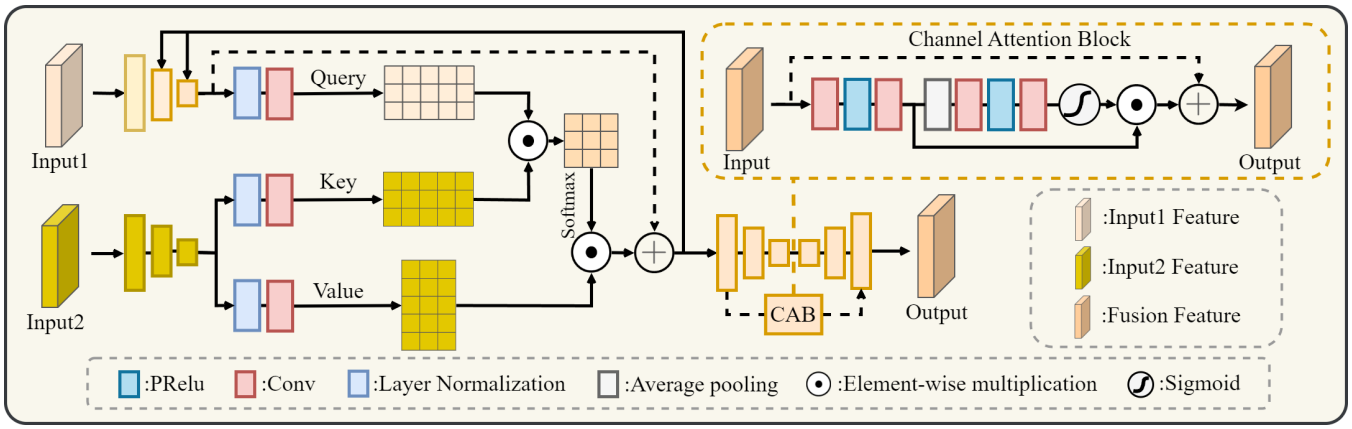

Figure 3: Event Image Fusion Block.

$\mathbb{R}^{H\times W\times C}$, where C symbolizes the quantity of event voxels. The mapping relationship can be described as:

$$\hat{B} = f_2\left(f_1\left(E;B\right), f_1\left(B;E\right)\right), \tag{7}$$

where $\hat{B} \in \mathbb{R}^{N\times H\times W\times 3}$ represents the coarse frames and N represents its count. f_1 represents feature extraction and cross-modal fusion, while f_2 denotes the fusion of features from two branches. The fine fusion stage can be expressed as:

$$I = f_3\left(Backward(Forward(\hat{B}))\right), \tag{8}$$

where $I \in \mathbb{R}^{H\times W\times 3}$ denotes the deblurred image, $Forward$ and $Backward$ refer to the processes of forward and backward propagation. f_3 symbolizes the final image reconstruction.

As illustrated in Figure 2, our model is structured with two stages: (1) Corase fusion. (2) Fine fusion.

4.2 Coarse Fusion Stage

Events and images represent two distinct data types and effectively integrating their features stands as a crucial step in the deblurring process. Events, characterized by asynchronous sparse data, render traditional computer vision methods obsolete. Directly inputting each event into the model would demand a substantial amount of computational resources. Therefore, we adopted the CTER method from [Sun *et al.*, 2022], which averages the events into six parts, accumulating events from the middle time to various time points as voxel blocks.

In dynamic scenes, triggered events suggest positions that are more likely to be blurred, while positions without events tend to remain sharp. This aligns with the voxel input, where positions with higher events accumulation merit increased attention. Hence, we introduce the Event Image Fusion Block (EIFB), as depicted in Figure 3. The process involves two branches for images and events, taking blurred frames and events as input. The module employs an encoder-decoder structure with UNet, extracting multi-scale features through stacked UNets. Both the decoder and encoder utilize Residual Blocks for feature extraction, consisting of two 3×3 convolutions with LeakyReLU activation functions and a residual connection utilizing a 1×1 convolution. The patch size is set as 256×256 and the encoder performs two downsampling steps using convolution with a kernel size of 4×4 and a stride of 2, obtaining feature maps of sizes 128×128 and 64×64, respectively.

The lower scale features are upsampled by the decoder and merged with the same scale features through a skip connection using a CAB [Zhang *et al.*, 2018]. In the image branch, the image features act as Q_{image}, while events serve as V_{event} and K_{event} in the cross-attention module. The event branch in turn, with input event features acting as Q_{event}. EIFB adopts the UNet structure with three scales and utilizes multi-head attention with the number of heads corresponding to the channel numbers of the feature maps. The decoder outputs feature maps at the same scale as the input image. The outputs of both branches are further concatenated after passing through a SAM [Chen *et al.*, 2021]. Considering the different voxels correspond to event integrals from the middle time to different time points, we can change the number of coarse frames generated. Regarding the impact of the number of coarse frames on our model, we will discuss this in the subsequent experiments. SAM propagates useful features from the current stage to the next while simultaneously calculating the loss by combining it with ground truth and expediting the image reconstruction process.

4.3 Fine Fusion Stage

Event data inherently contains a wealth of temporal information which has often been overlooked in previous researches [Han *et al.*, 2021; Kim *et al.*, 2023; Zhou *et al.*, 2023; Zhang *et al.*, 2023b; Sun *et al.*, 2022; Sun *et al.*, 2023; Yang and Yamac, 2023; Li *et al.*, 2023a], where after merging features from both images and events, directly synthesizing a single image disregards the temporal information contained in the events.

In the coarse fusion stage, where we obtain multiple frames at different times, so we designe a Bidirectional Frame Fusion Block (BFFB) to effectively leverage the spatiotemporal information within these frames, as illustrated in Figure 4. This module consists of two sub-modules: the forward prop-

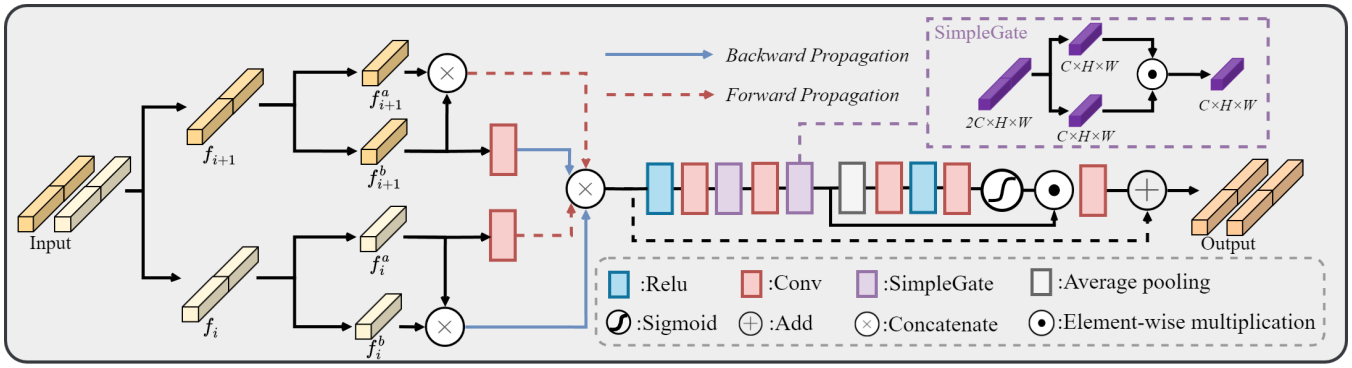

Figure 4: Bidirectional Frame Fusion Block.

agation and backward propagation modules. The red dashed lines represent forward propagation, while the blue solid lines represent backward propagation. During forward propagation, the feature maps of f_i are evenly divided into f_i^a and f_i^b, with one branch concatenating with f_{i+1} in the channel dimension to obtain preliminary fused features. Leveraging half of the channels in the feature maps provides effective information exchange between adjacent frames and reducing the model complexity and computational load. The fused features then undergo convolution and stacking with SimpleGates [Chen *et al.*, 2022b], which replace non-linear activation functions through the product of two feature maps, simplifying the model. The module concludes with an efficient channel attention block, capturing global information and fusing spatiotemporal information between preceding and subsequent frames. Through repeated stacking of forward and backward propagation modules, each frame learns ample information from adjacent frames.

5 Experiments

We assess our approach on two image deblurring datasets: the GoPro dataset [Nah *et al.*, 2017] and the recently introduced REBlur dataset [Sun *et al.*, 2022].

GoPro. The GoPro dataset are commonly employed for motion deblurring task which consist of 3214 pairs of blurred and sharp images with a resolution of 1280×720. Blurred images are generated by averaging 7 to 13 consecutive frames, producing variations in blur intensity. As the dataset solely contains images, we employe the event simulator ESIM [Rebecq *et al.*, 2018] to generate the corresponding events. Among these pairs, 2103 are utilized for training, and the remaining 1111 are reserved for testing.

REBlur. The REBlur dataset combines event data with genuine blurred images, featuring 1469 pairs across various motion scenarios. We split the dataset into training (486 pairs) and testing (983 pairs) sets. To enhance the model's performance, we fine-tune it on the training subset of the REBlur dataset, pretrained with the GoPro dataset and subsequently evaluate its effectiveness on the testing set.

Model	Events	PSNR↑	SSIM↑
D^{2}Nets† [Shang *et al.*, 2021]	✔	31.60	0.940
MPRNet [Zamir *et al.*, 2021]	✘	32.66	0.959
HINet [Chen *et al.*, 2021]	✘	32.71	0.959
Restormer [Zamir *et al.*, 2022]	✘	32.92	0.961
ChaIR [Cui and Knoll, 2023]	✘	33.28	0.963
FSNet [Cui *et al.*, 2023]	✘	33.29	0.963
HINet† [Chen *et al.*, 2021]	✔	33.69	0.961
NAFNet [Chen *et al.*, 2022b]	✘	33.69	0.967
M3SNet [Gao *et al.*, 2023]	✘	33.74	0.967
MADANET+ [Yang and Yamac, 2022]	✔	33.84	0.964
GRL [Li *et al.*, 2023b]	✘	33.93	0.968
Vitoria et al. [Vitoria *et al.*, 2022]	✔	34.33	0.944
NIRE [Zhang *et al.*, 2023b]	✔	35.03	0.973
EFNet [Sun *et al.*, 2022]	✔	35.46	0.972
DLEFNet [Yang and Yamac, 2023]	✔	35.61	0.973
CFFNet (Ours)	✔	**36.26**	**0.976**

Table 1: Deblurring results on GoPro dataset. "Events" indicates whether event data has been introduced. D^{2}Nets† and HINet† indicate the introduction of additional event data.

5.1 Implementation Details

Our model is designed as an end-to-end network, eliminating the need for pre-training. During training, we input cropped images sized at 256×256 and event voxels into the network, incorporating horizontal and vertical flipping for both images and event voxels in heat pixels to augment the dataset. Utilizing the Adam optimizer, we implement a cosine annealing strategy, adjusting the learning rate from 4×10^{-4} to 1×10^{-7}. Training is conducted on 4 NVIDIA 3090 RTX GPUs for 300k iterations. The fine-tuning process on the REBlur dataset involves 2000 iterations on one A100, initializing the learning rate at 4×10^{-5}. All other parameters and configurations remain consistent with the training on the GoPro dataset.

5.2 Evaluation

We evaluate our method on two datasets for image deblurring task and report standard metrics for image restoration, including PSNR and SSIM. To compare the visual effects of different models, we selected several images and compared the results obtained by running them on different pre-trained models.

Figure 5: Visual comparison of the state-of-the-art algorithms on GoPro test set.

Results on GoPro. We present the deblurring results in Table 1. Due to utilizing more data information, event-based methods exhibit a more pronounced deblurring effect. Among these models, our approach maximizes the utilization of spatiotemporal information from event data, yielding the best evaluation indicators. To further underscore the superiority of our model, we introduce event data into two frame-based method: HINet and D^2Net. While there is an improvement compared to the original models, our approach maintains a significant lead. We showcase visualization examples of our model in Figure 5, demonstrating superior detail restoration in complex scenes.

Results on REBlur. Compared with the GoPro dataset, the scenes in REBlur are simpler. As they are grayscale pictures, models performs well on this dataset. The evaluation metrics for our model and others are presented in Table 2, with visual comparisons shown in Figure 6. Our model outperforms other state-of-the-art networks in both evaluation metrics and visual effects. It is evident that most other models struggle in handling blurred scenes, exhibiting noticeable ghosting phenomena. In contrast, our model demonstrates a commendable restoration effect on edges and textures. The quantitative results reveal that our model achieved the best outcomes in both PSNR and SSIM metrics, benefiting from the enhanced exploration of event data.

5.3 Ablation Study

We examine the contributions of each component in our CFFNet on the GoPro dataset and explored the impact of hyperparameters on both the GoPro and REBlur datasets, leading to the following conclusions:

Model	Events	PSNR↑	SSIM↑
SRN [Tao *et al.*, 2018]	✘	35.10	0.961
NAFNet [Chen *et al.*, 2022b]	✘	35.48	0.962
Restormer [Zamir *et al.*, 2022]	✘	35.50	0.959
HINet [Chen *et al.*, 2021]	✘	35.58	0.965
EDI [Pan *et al.*, 2019]	✔	36.52	0.964
SRN† [Tao *et al.*, 2018]	✔	36.68	0.970
HINet† [Chen *et al.*, 2021]	✔	37.68	0.973
EFNet [Sun *et al.*, 2022]	✔	38.12	0.975
REFID [Sun *et al.*, 2023]	✔	38.34	0.975
DLEFNet [Yang and Yamac, 2023]	✔	38.40	-
CFFNet (Ours)	✔	**38.54**	**0.977**

Table 2: Deblurring results on REBlur dataset. Events indicates whether event data has been introduced. SRN† and HINet† indicate the introduction of additional event data.

EIFB. We compare different fusion methods in the coarse fusion stage, as depicted in Table 3, the columns "Frames" and "Events" respectively signify whether image and event data are incorporated. "Cross-attention" denotes the utilization of regular single-scale attention. The model exhibits subpar performance when relying solely on image frames, whereas the introduction of event data markedly enhances its capabilities. Notably, our EIFB module outperforms ordinary attention, showcasing superior integration of information from events and images.

BFFB. In Table 4, we compare the performance of models in the fine fusion stage using different methods to fuse multi-frame features. We employ Add, Concatenate and our proposed Bidirectional Frame Fusion Block (BFFB) for multi-frame fusion. The results demonstrate that introducing additional event data still leads to significant improve-

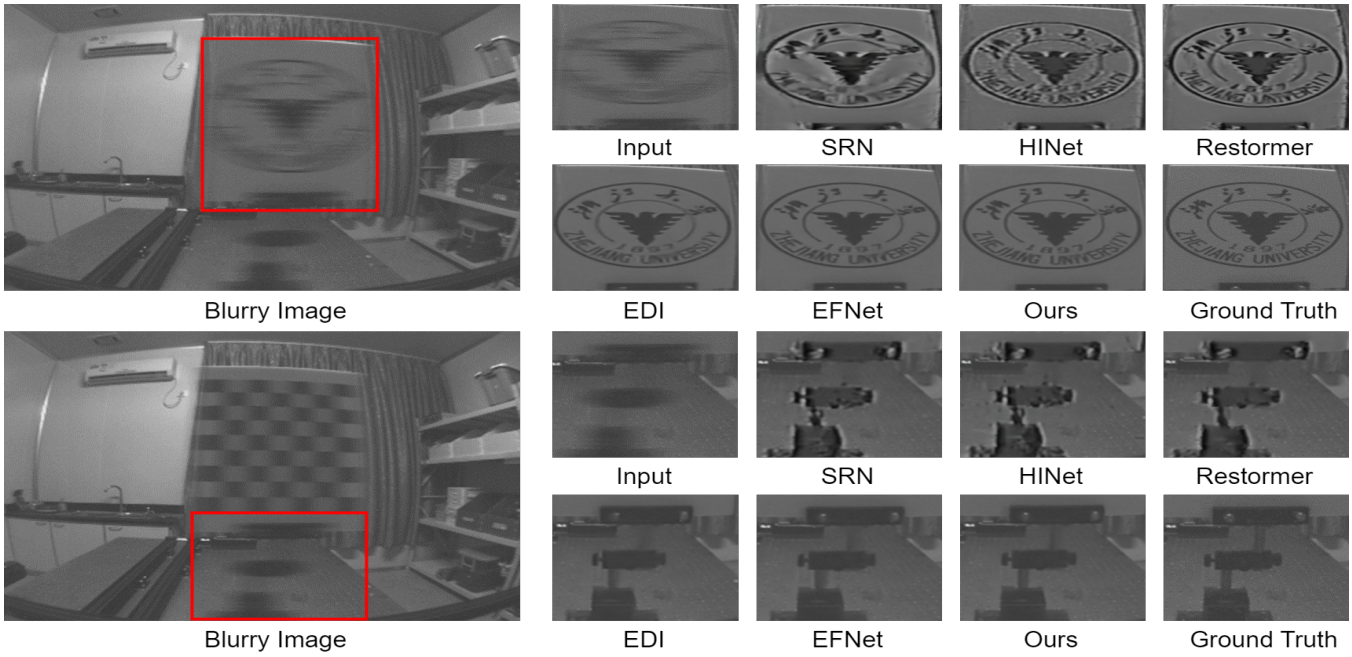

Figure 6: Visual comparison of the state-of-the-art algorithms on REBlur test set.

Frames	Events	Fusion method	PSNR↑	SSIM↑
✔	✘	n/a	29.06	0.936
✔	✔	Cross-attention	33.90	0.966
✔	✔	EIFB	**35.12**	**0.970**

Table 3: Ablation of fusion methods in the coarse fusion stage

Frames	Events	Fusion method	PSNR↑	SSIM↑
✔	✘	BFFB	33.45	0.944
✔	✔	Add	34.92	0.968
✔	✔	Concatenate	34.88	0.967
✔	✔	BFFB	**36.26**	**0.976**

Table 4: Ablation of fusion methods in the fine fusion stage

ments. Among these different fusion methods, the BFFB outperforms others and showcases superior performance, as it is specifically designed to handle coarse frames from the coarse fusion output.

Frames number. We investigate the impact of the quantity of generated frames in the coarse fusion stage. Our exploration delved into the influence of varying frame numbers on model performance, ranging from 2 to 6. The results illustrated in Figure 7 demonstrate that the model achieves optimal performance when generating four coarse frames. This is because a smaller number of generated frames fails to effectively utilize the temporal information from event data. However, exceeding four generated frames results in an extensive temporal span that leads to inadequate information fusion.

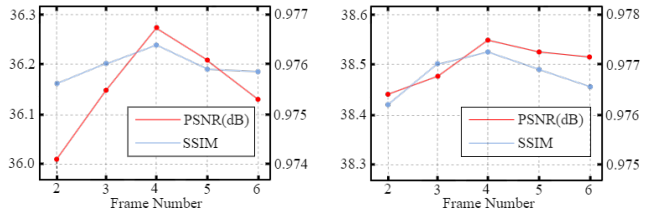

Figure 7: Ablation of frames number in BFFB on different datasets. (a) on GoPro dataset. (b) on REBlur dataset

6 Conclusion

In tackling single-image deblurring task, we introduce a novel event-based Coarse-to-Fine Fusion network (CFFNet), a model grounded in image event fusion principles. The coarse fusion stage employs an Event Image Fusion Block (EIFB), merging image and event information to generate coarse frames. The fine fusion stage incorporates a Bidirectional Frame Fusion Block (BFFB), combining information from coarse frames to effectively learn the mapping relationship from blurry images to sharp images. Evaluation on the GoPro dataset and REBlur dataset demonstrates our method's superiority over state-of-the-art approaches.

Acknowledgments

This work was supported in part by Science and Technology Innovation (STI) 2030—Major Projects under Grant 2022ZD0208700, and National Natural Science Foundation of China under Grant 62376264.

References

[Cao *et al.*, 2022] Chengzhi Cao, Xueyang Fu, Yurui Zhu, Gege Shi, and Zheng-Jun Zha. Event-driven video deblurring via spatio-temporal relation-aware network. In *IJCAI*, pages 799–805, 2022.

[Chen *et al.*, 2021] Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen. Hinet: Half instance normalization network for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 182–192, 2021.

[Chen *et al.*, 2022a] Haoyu Chen, Minggui Teng, Boxin Shi, Yizhou Wang, and Tiejun Huang. A residual learning approach to deblur and generate high frame rate video with an event camera. *IEEE Transactions on Multimedia*, 2022.

[Chen *et al.*, 2022b] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European Conference on Computer Vision*, pages 17–33. Springer, 2022.

[Cho *et al.*, 2021] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4641–4650, 2021.

[Cui and Knoll, 2023] Yuning Cui and Alois Knoll. Exploring the potential of channel interactions for image restoration. *Knowledge-Based Systems*, 282:111156, 2023.

[Cui *et al.*, 2023] Yuning Cui, Wenqi Ren, Xiaochun Cao, and Alois Knoll. Image restoration via frequency selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–16, 2023.

[Dai and Wu, 2008] Shengyang Dai and Ying Wu. Motion from blur. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[Fergus *et al.*, 2006] Rob Fergus, Barun Singh, Aaron Hertzmann, Sam T Roweis, and William T Freeman. Removing camera shake from a single photograph. In *Acm Siggraph 2006 Papers*, pages 787–794. 2006.

[Gallego *et al.*, 2020] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020.

[Gao *et al.*, 2023] Hu Gao, Jing Yang, Ying Zhang, Ning Wang, Jingfan Yang, and Depeng Dang. A mountain-shaped single-stage network for accurate image restoration. *arXiv preprint arXiv:2305.05146*, 2023.

[Han *et al.*, 2021] Jin Han, Yixin Yang, Chu Zhou, Chao Xu, and Boxin Shi. Evintsr-net: Event guided multiple latent frames reconstruction and super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4882–4891, 2021.

[Hyun Kim *et al.*, 2017] Tae Hyun Kim, Kyoung Mu Lee, Bernhard Scholkopf, and Michael Hirsch. Online video deblurring via dynamic temporal blending network. In *Proceedings of the IEEE international conference on computer vision*, pages 4038–4047, 2017.

[Kim *et al.*, 2023] Jeongmin Kim, Dipon Kumar Ghosh, and Yong Ju Jung. Event-based video deblurring based on image and event feature fusion. *Expert Systems with Applications*, 223:119917, 2023.

[Krishnan *et al.*, 2011] Dilip Krishnan, Terence Tay, and Rob Fergus. Blind deconvolution using a normalized sparsity measure. In *CVPR 2011*, pages 233–240. IEEE, 2011.

[Kundur and Hatzinakos, 1996] Deepa Kundur and Dimitrios Hatzinakos. Blind image deconvolution. *IEEE signal processing magazine*, 13(3):43–64, 1996.

[Li *et al.*, 2023a] Dasong Li, Xiaoyu Shi, Yi Zhang, Ka Chun Cheung, Simon See, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. A simple baseline for video restoration with grouped spatial-temporal shift. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9822–9832, 2023.

[Li *et al.*, 2023b] Yawei Li, Yuchen Fan, Xiaoyu Xiang, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Efficient and explicit modelling of image hierarchies for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18278–18289, 2023.

[Lin *et al.*, 2020] Songnan Lin, Jiawei Zhang, Jinshan Pan, Zhe Jiang, Dongqing Zou, Yongtian Wang, Jing Chen, and Jimmy Ren. Learning event-driven video deblurring and interpolation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 695–710. Springer, 2020.

[Nah *et al.*, 2017] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3883–3891, 2017.

[Nimisha *et al.*, 2017] Thekke Madam Nimisha, Akash Kumar Singh, and Ambasamudram N Rajagopalan. Blur-invariant deep learning for blind-deblurring. In *Proceedings of the IEEE international conference on computer vision*, pages 4752–4760, 2017.

[Pan *et al.*, 2016] Jinshan Pan, Deqing Sun, Hanspeter Pfister, and Ming-Hsuan Yang. Blind image deblurring using dark channel prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1628–1636, 2016.

[Pan *et al.*, 2019] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6820–6829, 2019.

[Park *et al.*, 2020] Dongwon Park, Dong Un Kang, Jisoo Kim, and Se Young Chun. Multi-temporal recurrent neural networks for progressive non-uniform single image deblurring with incremental temporal training. In *European*

Conference on Computer Vision, pages 327–343. Springer, 2020.

[Rebecq *et al.*, 2018] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. Esim: an open event camera simulator. In *Conference on robot learning*, pages 969–982. PMLR, 2018.

[Shang *et al.*, 2021] Wei Shang, Dongwei Ren, Dongqing Zou, Jimmy S Ren, Ping Luo, and Wangmeng Zuo. Bringing events into video deblurring with non-consecutively blurry frames. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4531–4540, 2021.

[Sun *et al.*, 2022] 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 *European Conference on Computer Vision*, pages 412–428. Springer, 2022.

[Sun *et al.*, 2023] Lei Sun, Christos Sakaridis, Jingyun Liang, Peng Sun, Jiezhang Cao, Kai Zhang, Qi Jiang, Kaiwei Wang, and Luc Van Gool. Event-based frame interpolation with ad-hoc deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18043–18052, 2023.

[Tao *et al.*, 2018] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8174–8182, 2018.

[Vitoria *et al.*, 2022] Patricia Vitoria, Stamatios Georgoulis, Stepan Tulyakov, Alfredo Bochicchio, Julius Erbach, and Yuanyou Li. Event-based image deblurring with dynamic motion awareness. In *European Conference on Computer Vision*, pages 95–112. Springer, 2022.

[Wan *et al.*, 2020] Shengdao Wan, Shu Tang, Xianzhong Xie, Jia Gu, Rong Huang, Bin Ma, and Lei Luo. Deep convolutional-neural-network-based channel attention for single image dynamic scene blind deblurring. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(8):2994–3009, 2020.

[Wulff and Black, 2014] Jonas Wulff and Michael Julian Black. Modeling blurred video with layers. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 236–252. Springer, 2014.

[Xu and Jia, 2010] Li Xu and Jiaya Jia. Two-phase kernel estimation for robust motion deblurring. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part I 11*, pages 157–170. Springer, 2010.

[Yang and Yamac, 2022] Dan Yang and Mehmet Yamac. Motion aware double attention network for dynamic scene deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1113–1123, 2022.

[Yang and Yamac, 2023] Dan Yang and Mehmet Yamac. Deformable convolutions and lstm-based flexible event frame fusion network for motion deblurring. *arXiv preprint arXiv:2306.00834*, 2023.

[Yang *et al.*, 2022] Wen Yang, Jinjian Wu, Jupo Ma, Leida Li, Weisheng Dong, and Guangming Shi. Learning for motion deblurring with hybrid frames and events. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1396–1404, 2022.

[Yang *et al.*, 2023] Wen Yang, Jinjian Wu, Leida Li, Weisheng Dong, and Guangming Shi. Event-based motion deblurring with modality-aware decomposition and recomposition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8327–8335, 2023.

[Yu *et al.*, 2023] Lei Yu, Bishan Wang, Xiang Zhang, Haijian Zhang, Wen Yang, Jianzhuang Liu, and Gui-Song Xia. Learning to super-resolve blurry images with events. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[Zamir *et al.*, 2021] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14821–14831, 2021.

[Zamir *et al.*, 2022] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022.

[Zhang and Yu, 2022] Xiang Zhang and Lei Yu. Unifying motion deblurring and frame interpolation with events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17765–17774, 2022.

[Zhang *et al.*, 2018] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018.

[Zhang *et al.*, 2023a] Kaihao Zhang, Tao Wang, Wenhan Luo, Wenqi Ren, Björn Stenger, Wei Liu, Hongdong Li, and Ming-Hsuan Yang. Mc-blur: A comprehensive benchmark for image deblurring. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[Zhang *et al.*, 2023b] Xinyu Zhang, Hefei Huang, Xu Jia, Dong Wang, and Huchuan Lu. Neural image re-exposure. *arXiv preprint arXiv:2305.13593*, 2023.

[Zhou *et al.*, 2023] Chu Zhou, Minggui Teng, Jin Han, Jinxiu Liang, Chao Xu, Gang Cao, and Boxin Shi. Deblurring low-light images with events. *International Journal of Computer Vision*, 131(5):1284–1298, 2023.