Motion Deblur by Learning Residual from Events

Kang Chen, Lei Yu†

***Abstract*—Conventional cameras face challenges when capturing motion information during the exposure due to their physical design, rendering the motion deblurring task ill-posed. To this end, we propose a Two-stage Residual-based Motion Deblurring (TRMD) framework for an event camera, which converts a blurry image into a sequence of sharp images, leveraging the abundant motion features encoded in events. In the first stage, a residual estimation network is trained to estimate the residual sequence, which measures the intensity difference between the intermediate frame and other frames sampled during the exposure. In the subsequent stage, the previously estimated residuals are combined with the blurry image to reconstruct the deblurred sequence based on the physical model of motion blur. To facilitate the efficient integration of image and event modalities for residual estimation, we propose a cross-modal fusion module based on spatial-channel attention, aiming to fuse the complementary spatial-temporal features of two modalities. Extensive experiments demonstrate that our method outperforms current state-of-the-art approaches on the synthetic dataset GOPRO and produces superior visualization with less noise and artifacts on the real blur event dataset REBlur. Our code, data and trained models are available at https://github.com/chenkang455/TRMD.**

***Index Terms*—Event Camera, Motion Deblur, Residual**

I. Introduction

TRADITIONAL cameras necessitate an exposure duration to accumulate sufficient photons for generating high-quality images [1]. During this period, rapid object motion or significant camera shake can cause the resulting image to be blurry [2, 3]. While traditional cameras struggle to record motion characteristics during the exposure interval, there are multiple kinematic scenarios corresponding to the same blurry frame [4, 5], thus making motion deblurring an ill-posed problem. In recent years, Convolutional Neural Networks (CNN) have exhibited outstanding performance in motion-deblurring tasks. Possessing robust fitting capabilities, CNN could directly learn the complex mapping relationship between a blurry image and a sharp image [6–10]. However, the deficiency of motion information during the exposure interval still hinders the effectiveness of motion deblurring.

To tackle this challenge, recent work has leveraged event cameras [11–13], which are bio-inspired sensors capable of recording the log intensity change of each pixel to perform the motion deblurring task. In contrast to conventional cameras that rely on relatively prolonged exposure time to deliver an image, event cameras asynchronously produce dense events with microsecond-level latency, thus capturing abundant motion features during the exposure. Due to this benefit, these event-based deblurring methods not only reconstruct a single blurry image [14–16], but also recover a sequence of sharp images corresponding to the blurry one, which fully exploits the low-latency property of the event camera [4, 17–21].

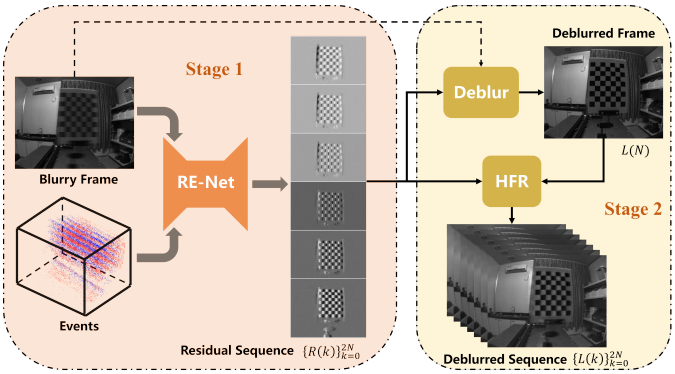


Fig. 1: Illustration of our TRMD framework. In the initial stage, we employ the Residual Estimation Network (RE-Net) with input being the blurry image and events to compute the residual sequence $\{R(k)\}_{k=0}^{2N}$ (the middle residual frame is omitted in the figure as it contains no information), which effectively captures the dynamic motion features during the exposure. In the subsequent stage, the residual sequence and the blurry image are further utilized to obtain the deblurred sequence $\{L(k)\}_{k=0}^{2N}$ based on the physical model of motion blur. Modules "Deblur" and "HFR" (High Frame Rate video generation) correspond to Eq. (11) and Eq. (12) respectively.

Among these methodologies, formulating the event-based motion deblurring task as a residual learning one is a widely adopted strategy [15–18] based on the observation that the residual signifies the difference in intensity between two images, which aligns seamlessly with the mechanism of event generation. Specifically, the approaches detailed in [17, 18] initially compute the residual between the blurry and sharp image with the help of events. Following this, they add the estimated residual to the blurred frame to reconstruct the sharp one. However, the framework that exclusively relies on a single sharp frame as the supervision faces challenges decoupling beneficial features from the temporally rich event stream. Simultaneously, their reconstruction framework does not integrate the physical model of event-based motion deblurring, resulting in their performance entirely dependent on the network architecture, which is hard to improve owing to the limited interpretability of neural networks. Consequently, there is an urgent need to develop a framework that diverges from the previous single-stage end-to-end deblurring approaches.

Moreover, to take advantage of the rich motion informa-

Kang Chen and Lei Yu are with the School of Electronic Information, Wuhan University, Wuhan, China. E-mail: {mrchenkang, ly.wd}@whu.edu.cn.

The research was partially supported by the National Natural Science Foundation of China under Grants 62271354 and 61871297.

† Corresponding author: Lei Yu.

tion encoded in the event stream and rich texture features embedded in the image, which are both indispensable in the motion deblurring task, these residual-based methods [16–18] take both the blurry image and events as the input of the network. Specifically, they concatenate two modalities and feed the combined new vector into the network. However, the asynchronous output of events exhibits spatially sparse and temporally dense features, significantly diverging from the blurry image's spatial and temporal characteristics. The fusion approach of simply concatenating two modalities without considering their statistical differences can significantly decrease the network fitting ability [14]. Therefore, an effective mechanism for fusing the image and event modalities must also be proposed.

In this paper, we propose a Two-Stage Residual-based Motion Deblurring (TRMD) framework, which converts the blurry image into a sequence of sharp images with the assistance of events. The fundamental idea is to leverage the residual sequence as an intermediate variable to build the connection between the two stages. We employ the designed residual estimation network to compute the residual sequence of the blurry image and events in the initial stage. In contrast to the residual proposed in [16–18], which measures the log or fractional intensity difference between images, our proposed residual describes the direct intensity difference between images, addressing the unstable training problem in previous methods. In the subsequent stage, utilizing the previously estimated residual sequence, we recover the middle sharp frame according to the physical mechanism of motion blur. Then, a sequence of sharp images is reconstructed by sequentially adding the residual sequence to the deblurred frame. The whole two-stage process is illustrated in Fig. 1. For the architecture of the residual estimation network, we employ U-Net [22] as the backbone, further cascading a feature extraction module and a feature fusion module, designed to fuse the spatial-temporal features of the event-image modalities.

To evaluate the performance of our approach in motion deblurring tasks, we conduct quantitative evaluations on the synthetic dataset [23]. The results demonstrate that our method outperforms other state-of-the-art algorithms. Furthermore, we visually showcase the performance of our method on the real-world REBlur dataset [14]. The reconstruction results demonstrate that our method effectively restores the blurry image while minimizing noise and artifacts.

To sum up, the main contributions of this paper include:

- We propose a Two-Stage Residual-based Motion Deblurring (TRMD) framework for event cameras. TRMD utilizes the residual sequence as the intermediate variable, which provides a stronger supervision signal for network training than the single-stage approach.
- We propose a novel residual form that describes the direct intensity difference between two images, which addresses the unstable training problem in the previous fraction residual approach.
- We design an event-image cross-modal fusion module for the Residual Estimation Network (RE-Net) based on spatial-channel attention and aims to fuse the spatial-temporal features of two modalities.

II. Related Work

A. Motion deblurring methods based on conventional cameras

In conventional cameras, CNN-based motion deblurring methods [6–9] aim to learn the mapping from a blurry image to a sharp image directly. Due to the limitation that a single blurry image has difficulty capturing motion information during exposure time, its deblurring performance often deteriorates significantly in blurry scenarios beyond the scope of the dataset. To solve this problem, existing approaches [24, 25] use multiple frames to improve image restoration performance. By analyzing the spatial-temporal relationships between these frames, additional motion information can be extracted to aid in restoring sharp images. Nevertheless, the performance enhancement of these methods is limited by the relatively low temporal resolution of video frames.

B. Motion deblurring methods based on event cameras

Event cameras [11–13] can asynchronously record the log-intensity change of individual pixels with low latency, thus capturing abundant motion features during the exposure period. Recent works have incorporated the output of event cameras as the auxiliary input of the deblurring network based on conventional cameras. Zhang et al. [16] initially processed events with a recurrent network to obtain their dense representation, which is further concatenated with the blurry image and fed into the deblurring network. Vitoria et al. [15] proposed to embed modulated deformable convolutions into the deblurring network and estimate their kernel offsets utilizing events.

In addition to reconstructing a single sharp image corresponding to the blurry one, some methods even recover a sequence of sharp images. Pan et al. [20] established the physical model of the event-based motion deblurring by solving a single variable non-convex optimization problem. This model explicitly constructs the relationship between the blurry image, events, and the sharp sequence. However, the assumption of a fixed trigger threshold and the presence of abundant noisy events in the captured scene degrade the restoration effect of the model. To solve these problems, Chen et al. [17] designed an event denoising module and two modified U-Net networks embedded with this module to perform the task of generating motion deblurring and high frame rate video in sequence. Lin et al. [18] incorporated a dynamic filter into their deblurring network to address the problem of spatially variable threshold.

Although the reconstruction of a sharp sequence using these CNN-based methods is limited to fixed moments during exposure time, Song et al. [4] proposed a novel continuous video representation and leveraged U-Net to estimate the critical parameters from a blurry image and events. Nevertheless, capturing a single blurry image and its corresponding sharp sequence, which served as the supervision signal during training, in the real world is challenging. To overcome this limitation, some methods [4, 17, 18] trained networks on synthetic datasets and applied them to other datasets. Still, their performance degraded significantly due to data inconsistency. Furthermore, Xu et al. [26] proposed a self-supervised framework that establishes the constraint between the outputs

of the deblurring network and the optical network. With this framework, their deblurring network can be trained on real-captured datasets without sharp sequence. Similarly, based on the constraint between adjacent blurry frames and events, Zhang and Yu [19] proposed a self-supervised framework that simultaneously performs image deblurring and interpolation tasks.

In summary, the aforementioned learning-based methods adopt an end-to-end one-stage recovery framework. However, this framework does not fully utilize the physical relationships between event streams, blurry images, and sharp images, thus improving a minor improvement.

C. Multi-modal fusion

Multi-modal fusion refers to integrating information from diverse sensors into a single modality, aiming to enhance the comprehensiveness of the data. In recent years, multi-modal fusion has proven to be a highly effective approach in many fields [27–31].

Event cameras are capable of providing event stream and image two modalities. The event modality contains rich motion information, but lacks the intensity description of the image. On the contrary, the image modality records the intensity of each pixel but provides limited kinetic features. Consequently, the appropriate fusion mechanism for integrating these two modalities is crucial in restoring high-quality deblurred images.

To achieve the fusion of the image and events, several methods [4, 17, 18, 21, 32] directly concatenate or multiply two modalities together, resulting in a new vector that serves as the input of the network. However, their fusion mechanism ignores the statistical difference between the two modalities, thus posing a challenge for the network to leverage both modalities' motion and texture features. To better fuse two modalities, Sun et al. [14] designed a cross-modal fusion module based on self-attention. Shang et al. [33] proposed an event fusion module to utilize rich boundaries in events. The weight matrix of this module can guide decoders to pay more attention to features beneficial for deblurring. Yang and Yamac [34] proposed a two-branch network and utilized events to detect the areas with high blur.

However, none of the above methods fully consider the fusion of spatial-temporal features encoded in two modalities.

III. Methods

This paper proposes a Two-Stage Residual-based Motion Deblurring (TRMD) framework for an event camera. We first briefly introduce the physical model of event-based motion deblurring in Section III-A. In Section III-B, we provide a comprehensive exposition of our TRMD framework. Next, a detailed description of the pre-processing operation for events, the structure of Residual Estimation Network (RE-Net), and the form of the loss function are provided in Section III-C. Finally, we highlight the strengths of our method over previous event-based deblurring methods in Section III-D.

A. Physical model of event-based motion deblurring

Motion blur is a visual effect that arises when there is relative movement between the camera and the captured scene during the exposure period. Based on the mathematical formulation of the motion blur as outlined in [35], the blurry image B can be represented as the temporal integration of the latent image $L(t)$ over the exposure duration, *i.e.*,

$$B = \frac{1}{T}\int_{f-T/2}^{f+T/2} L(t)dt \tag{1}$$

where T represents the duration of the exposure period $\mathcal{T}_{[f-T/2,f+T/2]}$ and f is defined as the middle moment during this period.

Event cameras can asynchronously record the log-intensity change of each pixel. For the pixel $\mathbf{x}$ in the image, event $e_{t,\mathbf{x}}$ is triggered at time t when the log-intensity change of this pixel exceeds the given threshold c. The polarity $p_{t,\mathbf{x}}$ of the generated event $e_{t,\mathbf{x}}$ has positive and negative two states, satisfying:

$$p_{t,\mathbf{x}} = \begin{cases} +1, \text{if } \log\frac{L(t,\mathbf{x})}{L(t-\Delta T,\mathbf{x})} \geq c \\ -1, \text{if } \log\frac{L(t,\mathbf{x})}{L(t-\Delta T,\mathbf{x})} \leq -c \end{cases} \tag{2}$$

where $L(t,\mathbf{x})$ represents the intensity of pixel $\mathbf{x}$ at time t, and ΔT represents the time interval between two adjacent events. Since the event generation mechanism is the same for all pixels, we omit the pixel location $\mathbf{x}$ in the subsequent notation for the sake of readability.

According to the event generation model as described in Eq. (2), we can bridge the log-intensity relationship between the middle time f and any time t, *i.e.*,

$$\log L(t) - \log L(f) = c\int_{f}^{t} e(s)ds \tag{3}$$

where $e(s)$ is the continuous expression of the event, satisfying $e(s) = p_{s_0}\cdot\delta(s-s_0)$, where $\delta(\cdot)$ is the Dirac function, s is the time variable in the interval $[f,t]$ and s_0 is the triggered timestamp of the event.

By combing Eq. (1) and Eq. (3) together, we obtain the relationship between the sharp image $L(f)$ and the blurry image B from the perspective of events, *i.e.*,

$$B = \frac{L(f)}{T}\int_{f-T/2}^{f+T/2} E(t)dt \tag{4}$$

where $E(t)$ can be mathematically expressed as follows:

$$E(t) = \exp(c\int_{f}^{t} e(s)ds) \tag{5}$$

We adopt the idea from [17] and define $E(t)$ as the residual, which also represents the ratio of the intensity at time t to time f according to Eq. (3):

$$E(t) = \frac{L(t)}{L(f)} \tag{6}$$

Eq. (4) explicitly constructs the mathematical relationship among the blurry image, events, and the sharp image, which is also known as the EDI model proposed in [20]. However, the residual $E(t)$ can only be accurately calculated at the

triggered instant of the event, which indicates that the intensity within the adjacent events interval is unknown. This precludes the direct computation of the continuous integral in the EDI. To address this issue, an intuitive approach is to employ numerical integration to approximate the continuous integral. Specifically, we divide the exposure period into $2N$ parts with constant duration, resulting in latent $2N+1$ frames and the discrete version of Eq. (4):

$$B = \frac{L(N)}{2N+1}\sum_{k=0}^{2N} E(k) \tag{7}$$

where $L(N)$ represents the intensity of the N-th frame and $E(k)$ denotes the residual between the k-th frame and the N-th frame.

B. Two-stage residual-based motion deblurring framework

The calculation accuracy of the residual E, as defined in Eq. (5), is significantly influenced by the occurrence of events and the threshold c. However, in real-world scenarios, the threshold varies both spatially and temporally [18], and there is a significant amount of noisy events due to the camera's bandwidth limitation [35]. These factors result in significant errors in the calculation of residual E, further degrading the performance of the EDI model. To alleviate this problem, a network $\mathcal{G}$ could be trained to approximate Eq. (5) with the input being events, formulated as:

$$E(k) = \mathcal{G}(\mathcal{E}[N,k];\Theta_1), \quad \forall k \in \{0,...,2N\} \tag{8}$$

where $\mathcal{E}[N,k]$ represents the event stream between the N-th frame and the k-th frame and Θ_1 denotes the parameters of the network.

In the ideal scenario, we can directly train a fully convolutional network such as U-Net to model the mapping between the event stream $\mathcal{E}[N,k]$ and the residual $E(k)$. However, the ground-truth residuals in dark regions of the image exhibit significant variations, which makes the network hard to estimate the accurate residual and the training process unstable. As illustrated in Eq. (6), given the condition that the numerator $L(f)$ is small, the residual $E(t)$ can still be large even if the intensity difference between $L(f)$ and $L(t)$ is small. This indicates that even a subtle change in pixel value can substantially impact the resulting residual.

To alleviate it, we propose a novel residual form $R(k)$. Different from previously proposed residuals, which measure the intensity ratio [18] or log-intensity difference [17], our residual directly describes the intensity difference between sharp latent images, *i.e.*,

$$\begin{aligned} R(k) &= L(k) - L(N) \\ &= L(N)\cdot E(k) - L(N) \end{aligned} \tag{9}$$

while the estimation of residual E solely depends on events according to Eq. (5), our residual R necessitates both the sharp frame $L(N)$ and events for its estimation. In the deblurring process, despite the sharp frame $L(N)$ is unknown, the blurry frame B and the sharp frame $L(N)$ share similar features in slightly blurry regions [17], which paves the way for substituting B for $L(N)$ in estimating the residual $R(k)$. Finally, we utilize the network $\mathcal{G}$ to estimate our proposed residual and the mapping relationship in Eq. (8) can be rewritten as:

$$R(k) = \mathcal{G}(\mathcal{E}[N,k], B; \Theta_1), \quad \forall k \in \{0,...,2N\} \tag{10}$$

Based on the proposed novel residual form, the Eq. (7) can be reformulated by simply substituting $E(k)$ for $R(k)$ according to Eq. (9):

$$L(N) = B - \frac{1}{2N+1}\sum_{k=0}^{2N} R(k) \tag{11}$$

Having obtained the estimated image in the middle frame $L(N)$, a sequence of sharp images can be further recovered by sequentially adding the residual sequence, *i.e.*,

$$L(k) = L(N) + R(k), \quad \forall k \in \{0,...,2N\} \tag{12}$$

To sum up, the pseudocode for our TRMD framework algorithm is listed below:

Algorithm 1: TRMD framework algorithm

Input: the blurry image B, the raw event stream $\mathcal{E}$
Divide the exposure time into $2N$ equal time intervals;
for $k = 0$ **to** $2N$ **do**
  Cut events between the N-th frame and the k-th frame from the raw event stream $\mathcal{E}$
  Estimate the residual according to Eq. (9)
  $R(k) = \mathcal{G}(\mathcal{E}[N,k], B; \Theta_1)$
end
Recover the sharp image at the middle frame according to Eq. (11)
$L(N) = B - \sum_{k=0}^{2N} R(k)/(2N+1)$
for $k = 0$ **to** $2N$ **do**
  Obtain the sharp image according to Eq. (12)
  $L(k) = L(N) + R(k)$
end
Output: Deblurred sharp sequence $\{L(k)\}_{k=0}^{2N}$

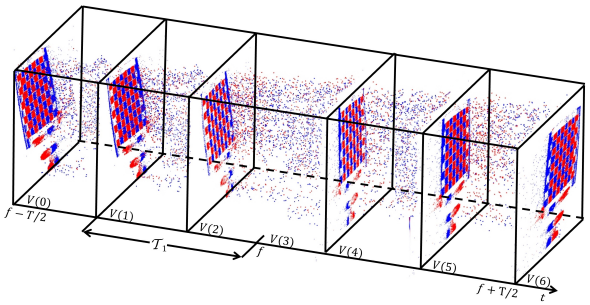


Fig. 2: Voxelization method for the event stream. The exposure period is divided into $2N$ intervals (with $N = 3$ in this example) and the asynchronous events are subsequently processed to derive the voxel sequence $\{V(k)\}_{k=0}^{2N}$.

C. Residual estimation network

1) Pre-processing of the events: Unlike conventional cameras that output 2D images synchronously, event cameras produce sparse events asynchronously, which does not match the input format of the convolutional neural network. To overcome this limitation, previous methods [4, 14, 18, 19, 21]

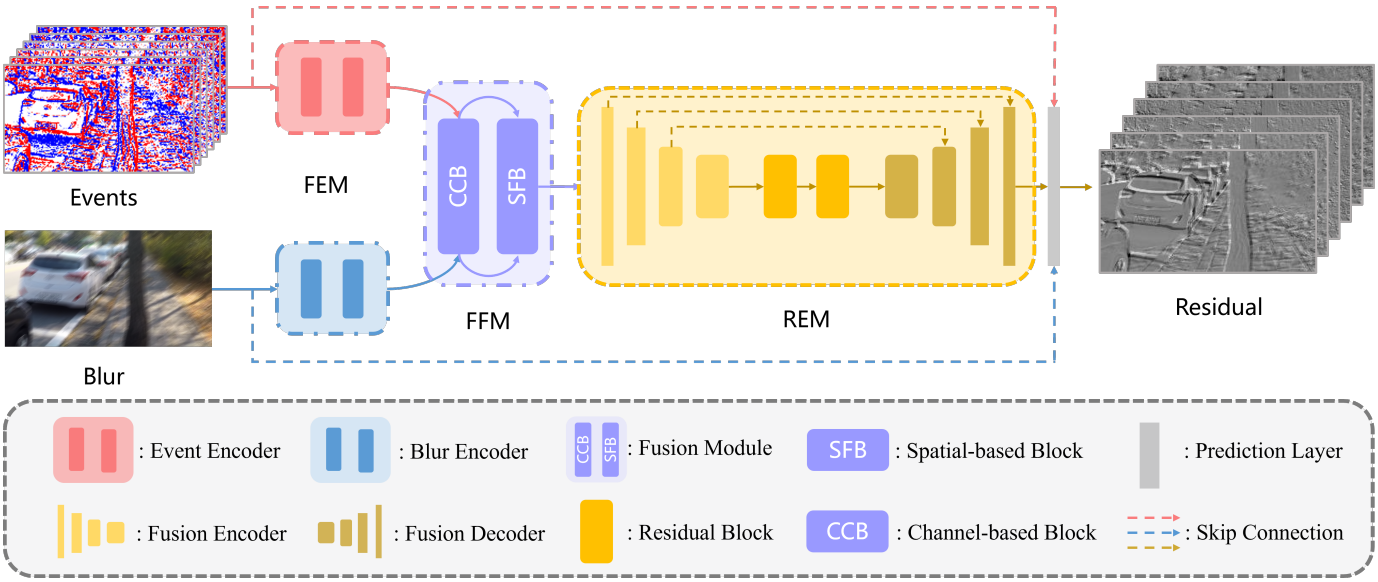


Fig. 3: Architecture of the proposed RE-Net. The event and blurry image modalities are initially processed by two encoders in feature extraction module (FEM) separately to extract shallow features and align channel numbers. Following that, feature fusion module (FFM) is designed to generate fused unimodality based on the events and blurry image. Finally, an encoder-decoder like residual estimation module (REM) is employed to calculate the residual sequence based on fused features.

employ the voxelization technique to convert sparse events into a 3D tensor. Specifically, Sun et al. [14] proposed a novel event representation named "Symmetric Cumulative Event Representation" (SCER). As shown in Fig. 2, the exposure period $\mathcal{T}_{[f-T/2,f+T/2]}$ is divided into $2N$ parts with equal time intervals, resulting in $2N+1$ latent frames. The voxel of the k-th frame $V(k)$ is defined as the accumulation of events between the middle frame N and that frame, which can be mathematically formulated as:

$$V(k) = \text{sgn}(k-N) \int_{s \in \mathcal{T}_k} e(s) ds \tag{13}$$

where $\mathcal{T}_k$ represents the duration period from the N-th frame to the k-th frame and $e(s)$ is the continuous expression of the event. SCER directly calculates the integral of events from the middle frame to any other latent frame. It captures the temporal dynamics of events and facilitates the estimation of residuals through the pre-computation of events integral. Therefore, we choose SCER as the voxelization technique of events.

According to Eq. (10), a single residual can be estimated by utilizing a single voxel and the blurry image. However, sequentially estimating the residuals of multiple frames is computationally demanding, and solely relying on a single voxel may not fully leverage the kinematic features encoded in the event stream. To mitigate these limitations, we take multiple voxels in conjunction with the blurry image as the network input and estimate the residuals of multiple frames simultaneously.

We exclude the voxel of the middle frame that does not convey any useful information and concatenate the remaining voxels to construct the event voxel tensor $E \in \mathbb{R}^{2N \times H \times W}$. Similarly, we discard the residual of the middle frame and concatenate other residuals to obtain the residual tensor $R \in \mathbb{R}^{2N \times H \times W}$ as shown in Fig. 3.

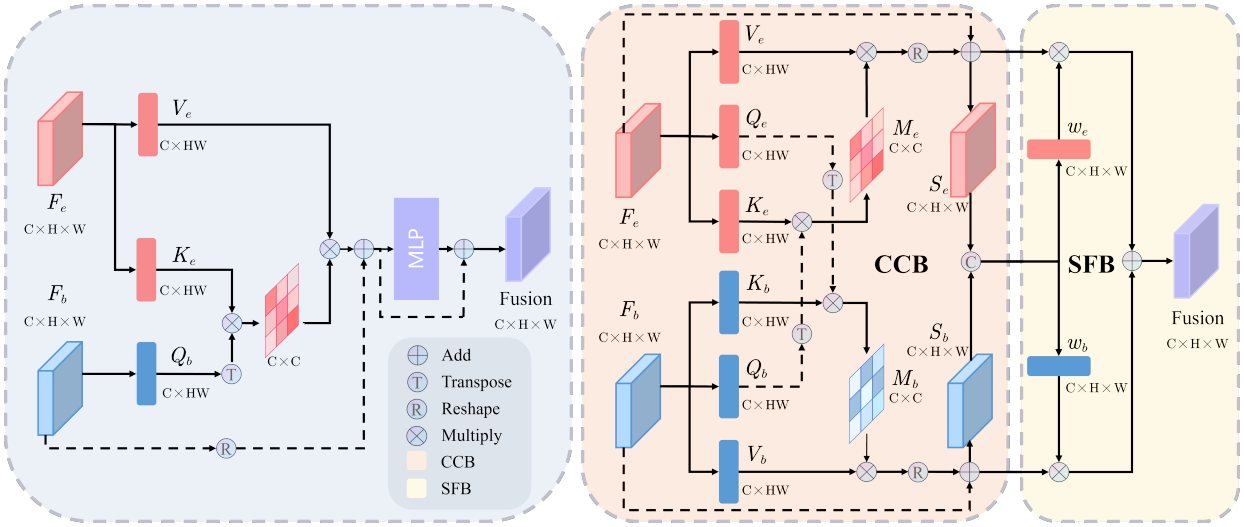
2) Network Architecture: The Residual Estimation Network (RE-Net) is designed to estimate the residual sequence R based on the blurry image B and the event voxel tensor E, as depicted in Fig. 3. The network's architecture consists of three cascaded modules: feature extraction module, feature fusion module, and residual estimation module.

2.a) Feature Extraction Module (FEM)

FEM is designed to extract shallow features from the event stream and the blurry image while aligning their channel numbers, corresponding to $2N$ and 1, respectively. To combine two modalities, methods such as [4, 18] simply concatenate the raw features of events and the blurry image without preprocessing them, taking the resulting tensor as the network input. However, this process omits the fact that two modalities possess distinct statistical distributions and characteristics. The image modality is acquired synchronously and contains abundant spatial features and limited temporal information, whereas the event modality is obtained asynchronously, exhibiting sparse spatial distribution and rich temporal information. The process of directly concatenating two modalities results in difficulties for the network to capture the cross-modal information between them, ultimately leading to a less effective representation.

To address the aforementioned issue, we design an event encoder and an image encoder to separately pre-process two modalities, as shown in Fig. 3. Each encoder consists of two 5×5 convolutional layers with the stride of 1. Given the input $B \in \mathbb{R}^{1 \times H \times W}$ and $E \in \mathbb{R}^{2N \times H \times W}$, the feature extraction process is performed without changing the image size. We obtain the processed event features F_e, and image features F_b of size $C \times H \times W$, where C equals the output channels of the feature extraction module.

2.b)Feature Fusion Module (FFM)



(a) Fusion Module of EFNet [14]. (b) Fusion Module of TRMD (Ours).

Fig. 4: The event-image cross-modal fusion module in EFNet (a) and TRMD (b). The red and blue parts represent the network branch for processing the event and image modality respectively. Our fusion module consists of a channel-based complementation block (CCB) and a spatial-based fusion block (SFB).

On the one hand, events contain abundant motion information that is inadequately represented in the blurry image [35]. On the other hand, the blurry image possesses a wealth of texture features that are relatively less encoded in the event stream [20]. Given the significance of both motion and texture features in motion deblurring, appropriate fusion for these two modalities is crucial in this process.

Inspired by the spatial-channel attention mechanism designed in [27–29], we propose a cross-modal Feature Fusion Module (FFM) consisting of a Channel-based Complementation Block (CCB) and Spatial-based Fusion Block (SFB), as illustrated in Fig. 4. This module takes the image feature F_b and the event features F_e as the input, then outputs the fused unimodal feature $F \in \mathbb{R}^{C\times H\times W}$.

Channel-based Complementation Block. The output of the RE-Net is the residual sequence of $2N$ frames within the exposure period. When the latent frame is far from the middle frame (such as $2N$-th frame), a large amount of event noise will accumulate, and the calculation error caused by the approximation in Eq. (10) increases. To mitigate it, CCB is designed to map the image information into different channels of events adaptively, thus better removing noise and estimating residual.

Attention was initially proposed in natural language processing [36, 37] and recently achieved superior performance in various computer vision tasks [38, 39]. To leverage the powerful cross-modal fusion capability of attention, EFNet [14] proposed a novel fusion module for two modalities, where the queries (Q) come from the image and the keys (K) and values (V) come from events, as shown in Fig. 4. However, their fusion mechanism did not consider the spatial-temporal features of the two modalities, resulting in insufficient cross-modal fusion. In contrast to the single-branch attention mechanism designed in EFNet, we propose a dual-branch attention mechanism to achieve complementarity between two modalities at the channel level. Specifically, the values of Q and V in the single branch are derived from the same modality, while K is obtained from another modality, as shown in Fig. 4.

We take the event branch colored red in Fig. 4 as an example to explain the working flow of the CCB. Given the input feature of events F_e, we initially feed it into three distinct 1×1 convolutional layers to compute $K_e, Q_e, V_e \in \mathbb{R}^{C\times H\times W}$, which are further reshaped into $\mathbb{R}^{C\times HW}$. Meanwhile, the same procedure is performed on the blurry image feature F_b to obtain K_b, Q_b, V_b. Then we multiply K_e and the transpose of Q_b to calculate the similarity between the two modalities in the channel dimension. We further divide the matrix product by $\sqrt{C}$, which is a common technique used to stabilize the training process of the network. After that, the resulting value is normalized using softmax to obtain the channel attention matrix $M_e \in \mathbb{R}^{C\times C}$, which can be mathematically formulated as:

$$M_e = softmax(\frac{K_e Q_b^T}{\sqrt{C}}) \tag{14}$$

The element in the i-th row and j-th column represents the similarity between the i-th channel of the event feature and the j-th channel of the image feature.

Finally, we perform a multiplication operation between M_e and V_e and reshape the result to match the shape of the input feature F_e. To summarize, the process for obtaining the event features S_e incorporated with the image trait can be mathematically formulated as follows:

$$S_e = reshape(M_e V_e) + F_e \tag{15}$$

where the input feature of events F_e is added to the multiplication result, serving as a residual connection.

Spatial-based Fusion Block (SFB). Events are triggered when the log-intensity change of the pixel exceeds a certain threshold, leading to the fact that the density of events in regions with subtle intensity change is relatively low. Thus, we

could utilize the gradients encoded in the image feature S_e, which reflect the spatial variation of pixel intensities, to assess the event density and filter out the event noise. Meanwhile, in areas with high event density, the motion is more severe and the likelihood of causing motion blur increases. Hence, events can serve as valuable indicators for estimating the magnitude of motion in different areas of the image. This enables the network to pay more attention to regions with higher blur levels, thus enhancing the network's ability to reduce motion blur effectively [34]. From the above analysis, we can conclude that the spatial features in the two modalities are complementary to each other, giving the significance of merging them at the spatial level.

Inspired by the spatial attention mechanism proposed in CBAM [40], which employs max and average pooling to process input features along the channel dimension, followed by feeding the concatenation of two processed features into a convolutional layer, we adopt a similar approach in this paper. Specifically, we first concatenate the processed event feature S_e and the blurry image feature S_b. Next, we utilize two convolutional layers to process the fused feature, obtaining two weight vectors w_e and w_b with a size of $C \times H \times W$. These two weight vectors capture the important spatial-temporal regions in the original features. After that, the final fusion feature F is calculated by multiplying the features of the two modalities with their respective weights and summing them together, as formulated by:

$$F = w_e S_e + w_b S_b \tag{16}$$

2.c) Residual Estimation Module (REM):

The framework of the residual estimation module is similar to U-Net [22], as shown in Fig. 3. The fused unimodal from FFM is first processed by four downsampling blocks, each consisting of a 5×5 convolutional layer with a stride of 2 and a ReLU activation function. After each downsampling operation, the size of the input is reduced by half and its channel number doubles, aiming to extract high-level features encoded in the fused modality. Next, the downsampled features are passed through two cascaded residual blocks for a more comprehensive representation of the features. After that, the output from the second residual block is sequentially decoded by four upsampling blocks. Each block comprises a transposed convolutional layer followed by a ReLU activation function. Before each transposed convolutional operation, a skip connection [22] from the corresponding downsampled layer is added to address the problem of information loss during downsampling.

Finally, the output of the last upsampling block, together with the blurry image B and events E, is concatenated and fed into a 1×1 convolutional prediction layer, which ensures that the channel number of the network output matches the length of the residual sequence $2N$.

3) Loss Function: The loss function of RE-Net consists of residual sequence estimation and single image deblurring two parts. The first component is designed to supervise the network in learning the correct mapping from events and the blurry image to the residual sequence. Specifically, we define it as the residual loss $\mathcal{L}_{res}$, which measures the difference between the estimated residual sequence and the ground truth residual sequence, formulated as:

$$\mathcal{L}_{res} = \sum_{k=0}^{2N} ||R(k) - G_R(k)||^2 \tag{17}$$

where $G_R(k)$ represents the ground truth residual of the k-th frame.

The second component single image deblurring loss $\mathcal{L}_{deb}$, intended to supervise the finally deblurred image, is defined as:

$$\mathcal{L}_{deb} = ||L(N) - G_L(N)||^2 \tag{18}$$

where $G_L(N)$ represents the ground truth sharp image corresponding to the blurry image.

The final loss function is mathematically formulated as follows:

$$\mathcal{L} = \mathcal{L}_{res} + \lambda \cdot \mathcal{L}_{deb} \tag{19}$$

where λ is a hyper-parameter set to 1 in this task.

D. Relation to existing models

Our TRMD differs from previous methods in two aspects: the framework for motion deblurring and the fusion mechanism of the events and image. To combine these two modalities, EFNet proposes a multi-modal fusion module based on the self-attention mechanism. Different from the single-branch fusion design in EFNet, our fusion module is based on the dual-branch attention mechanism. Furthermore, while EFNet establishes a direct mapping between the blurry image and sharp image, we describe the image deblurring task as a residual learning one following the techniques proposed in prior research [15–18], which closely mimics the fundamental mechanism of events thus enhancing the utilization of events.

Previous residual-based methods [17, 18] for motion deblurring primarily focus on estimating the residual between the blurry image and sharp image. Our method stands out in that we compute the residual sequence, which captures the intensity difference between the middle frame and other latent frames. The residual sequence encapsulates comprehensive motion information during the exposure and serves as a stronger supervision compared to previous methods during training. Besides, our method leverages the physical mechanism of motion blur more effectively, allowing for better decomposition of the blur process and resulting in superior results, as demonstrated in the subsequent section.

IV. Experiments

A. Data Preparation

1) GOPRO dataset: We utilize the synthetic dataset GOPRO [23], which is widely used in motion deblurring tasks, for the training and evaluation of our TRMD. Similar to the synthetic dataset construction workflow of [4, 19, 21], we first convert the images to the grayscale and downsample them to size 640×360 to keep consistent with the resolution of DVS-Gen2 [41]. Then, we employ RIFE [42] to increase the framerate by interpolating 7 frames between consecutive frames and apply the event generation simulator ESIM [43] to

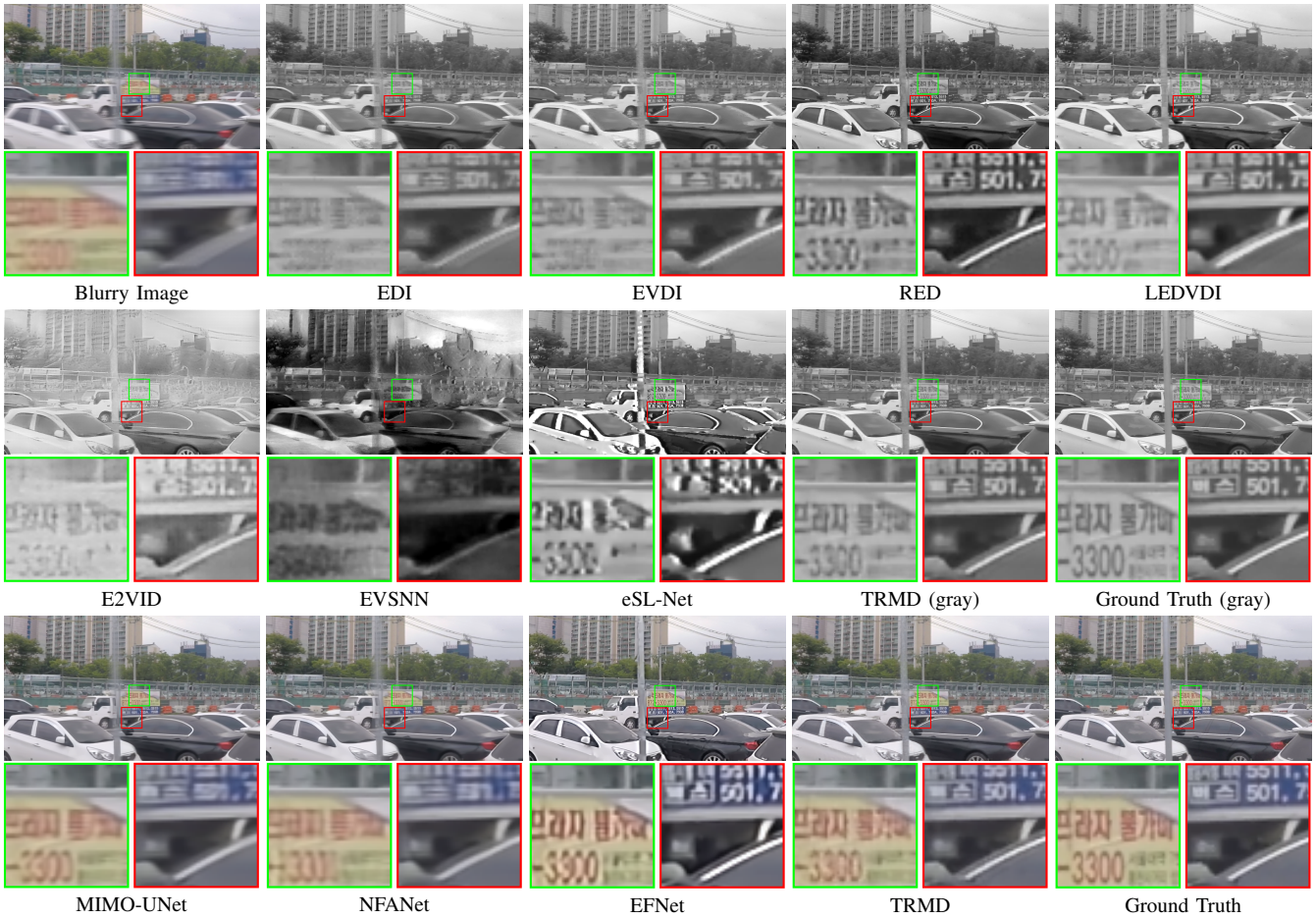


Fig. 5: Qualitative comparison for the single image deblurring task on the GoPro dataset.

generate simulated events with the noise following Gaussian distribution. Additionally, to simulate the motion blur in real-world scenarios, we synthesize one blurry image by averaging a specific number of sharp images on the interpolated high frame-rate sequence. As for the training and testing partitions of the dataset, we follow the official release which contains 22 video sequences for training and 11 video sequences for evaluation.

2) REBlur dataset: We evaluate the effectiveness of our proposed method in the real blur scenario on the Real Event Blur (REBlur) dataset released by [14]. This dataset includes the event stream captured by the DAVIS camera and grayscale images of size 260×360. REBlur includes 12 linear and nonlinear motion scenes under three different motion modes. We follow the official division of the dataset and use 983 pairs of blurry and sharp images from the testing set to evaluate the performance of various methods on the real blur dataset.

B. Training Details

Since the sequence of sharp images corresponding to the blurry image is unavailable on the REBlur dataset, our network can not be trained on this dataset. To overcome this limitation, we adopt the strategy employed by [4, 17, 18], which trains the network on the synthetic dataset and evaluates their performance on real blur datasets. To augment the training data and speed up the training process, we randomly crop images of size 256×200 from the original full-sized images in each epoch. During the testing phase, we feed the entire image into the network to assess performance. We use the PyTorch platform to build and train the network using an NVIDIA GeForce GTX 3090 GPU. The optimizer is Adam with a learning rate 0.001, and the batch size parameter is set to 8 in each epoch. To quantitatively analyze the performance of our model, we utilize the Peak Signal-to-Noise Ratio (PSNR), the Structural Similarity Index (SSIM), and the Learned Perceptual Image Patch Similarity [44] (LPIPS) these three metrics, which are commonly used in motion deblurring tasks.

C. Experimental Results

1) GOPRO dataset: We compare our TRMD in this paper with frame-based methods, *i.e.*, MIMO-UNet [6], NAFNet [45], event-based intensity reconstruction methods without image, *i.e.*, E2VID [46], EVSNN [47], and event-based motion deblurring methods, *i.e.*, EDI [20], EFNet [14], EVDI [19], eSL-Net [21], LEDVDI [18] and RED [26].

This paper does not compare event-based deblurring methods without public source code [15–17]. The quantitative evaluation results for the mentioned methods are presented in Table I. To comprehensively analyze the performance of these methods, we also visualize their reconstruction results. The visual representation for single frame deblurring is presented in Fig. 5 and the restoration outcome of the image sequence is illustrated in Fig. 6. Considering that MIMO-

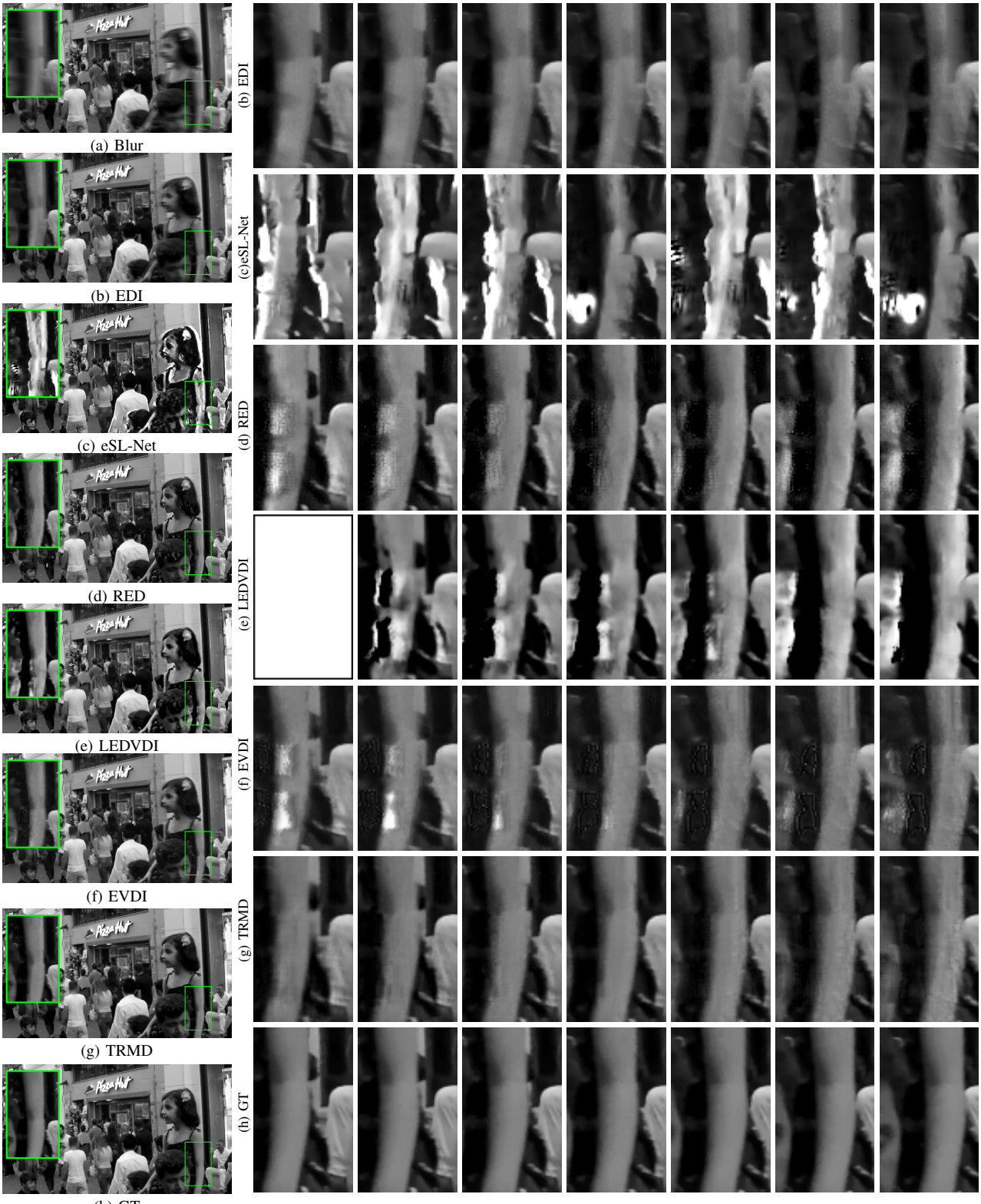


Fig. 6: Qualitative comparison for sequence reconstruction task on the GOPRO dataset. Among these methods, LEDVDI can only restore 6 sharp images from a blurry image, thus we use a blank picture to substitute the position of the first picture for better readability.

UNet, NAFet, and EFNet can restore color blurry images, we retrain our TRMD framework on the color GOPRO dataset for better comparison. The only difference between the color and grayscale models lies in the channel dimensions of the blurry input and the residual sequence. In particular, the channel dimension of B is modified from 1 to 3, while the dimension of the residual sequence is changed from $2N$ to $6N$. The network architecture and training hyperparameters remained unchanged from the original settings throughout the retraining procedure.

TABLE I: Quantitive results of single image deblurring and sequence reconstruction tasks on the GOPRO dataset. All learning-based methods are trained on the same dataset for a fair comparison. Due to the fact that only six sharp images are reconstructed by LEDVDI [18], while other methods and the ground truth clear images have seven images, the corresponding evaluation metrics for LEDVDI are slightly different from other methods. Both PSNR and SSIM are positive metrics denoted as ↑. LPIPS is negetive metric denoted as ↓. The best-performing results are highlighted in **bold**.

Methods	Inputs		Single frame deblurring			Sequence reconstruction			Params	FLOPs
	Image	Events	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓		
MIMO-UNet [6]	✓	×	32.530	0.9461	0.0643	/	/	/	16.1M	235G
NAFNet [45]	✓	×	30.820	0.8730	0.1259	/	/	/	67.9M	**63.3G**
E2VID [46]	×	✓	12.567	0.5584	0.2391	12.562	0.5575	0.2393	10.7M	104G
EVSNN [47]	×	✓	14.842	0.5796	0.2233	14.841	0.5793	0.2241	4.41M	93.3G
eSL-Net [21]	✓	✓	20.471	0.7588	0.1598	22.140	0.8098	0.1262	**0.19M**	552G
EDI [20]	✓	✓	30.608	0.9359	0.0856	28.576	0.9169	0.0959	/	/
RED [26]	✓	✓	30.991	0.9623	0.0728	28.487	0.9427	0.0854	9.76M	562G
EVDI [19]	✓	✓	31.708	0.9572	0.0511	30.798	0.9475	0.0543	0.39M	423G
LEDVDI [18]	✓	✓	34.868	0.9644	0.0425	32.898	0.9583	0.0423	4.99M	509G
EFNet [14]	✓	✓	35.803	0.9725	0.0382	/	/	/	8.47M	894G
TRMD (Ours)	✓	✓	**36.678**	**0.9830**	**0.0200**	**34.370**	**0.9735**	**0.0275**	19.3M	112G

In the blurry scenario depicted in Fig. 5, the deblurred image generated by our method exhibits the highest similarity to the Ground Truth, containing fewer noise and artifacts regardless of whether it is observed in color or grayscale space. Frmae-based methods such as MIMO-UNet and NAFNet fail to reconstruct the letters and numbers in the figure due to the limited information encapsulated in the single blurry frame. Event-based intensity reconstruction methods such as E2VID and EVSNN fall short in capturing the background intensity of the image, as events mainly record the log-intensity change of the pixel. While the deblurred image generated by EFNet exhibits sharper details and more precise edges, it tends to overemphasize visual features, leading to distortion in the color space and ultimately yielding unrealistic results. A comparable issue can also be observed in eSL-Net, where the restored letters are significantly brighter than the non-blurry area. Meanwhile, the non-learning method EDI generates substantial noise in the restored edges, letters, and numbers. This error is attributed to the large amount of noise in the event stream, which produces a discrepancy in the calculation of the double integral. Moreover, while EVDI and RED have designed self-supervised frameworks that allow models to be trained on ground-truth datasets, and LEDVDI has designed modules to denoise events, their lack of robust cross-modal fusion mechanism leads to subpar reconstruction of image details compared to our method.

Furthermore, the visual comparison of sequence reconstruction is illustrated in Fig. 6 and the result demonstrates that our method has the best visual quality.

2) REBlur Dataset: We further visualize performances of EDI [20], LEDVDI [18], RED [26], EVDI [19], EFNet [14], MIMO-UNet [6], NAFNet [45] and eSL-Net [21] on the REBlur dataset for qualitative comparison, as depicted in Figs. 7 to 9. It is worth noting that EFNet has been further fine-tuned on the REBlur, which sets it apart from other learning-based motion deblurring methods exclusively trained on the GOPRO. To ensure a fair comparison, we evaluate the performance of EFNet on the REBlur utilizing the model solely trained on the GOPRO without further fine-tuning.

In the sequence reconstruction task, methods such as EDI, EVDI, EVDI, and LEDVDI exhibit a failure to accurately restore two black holes, resulting in a line-like artifact due to the presence of noise and artifacts as shown in Fig. 7. While eSL-Net successfully restores the two black holes, it encounters a similar issue of over-exaggeration as observed in the GOPRO, where the brightness in the blurry area is inconsistent with that of the sharp area. In contrast, our method successfully restores two black holes while maintaining a high similarity to the Ground Truth.

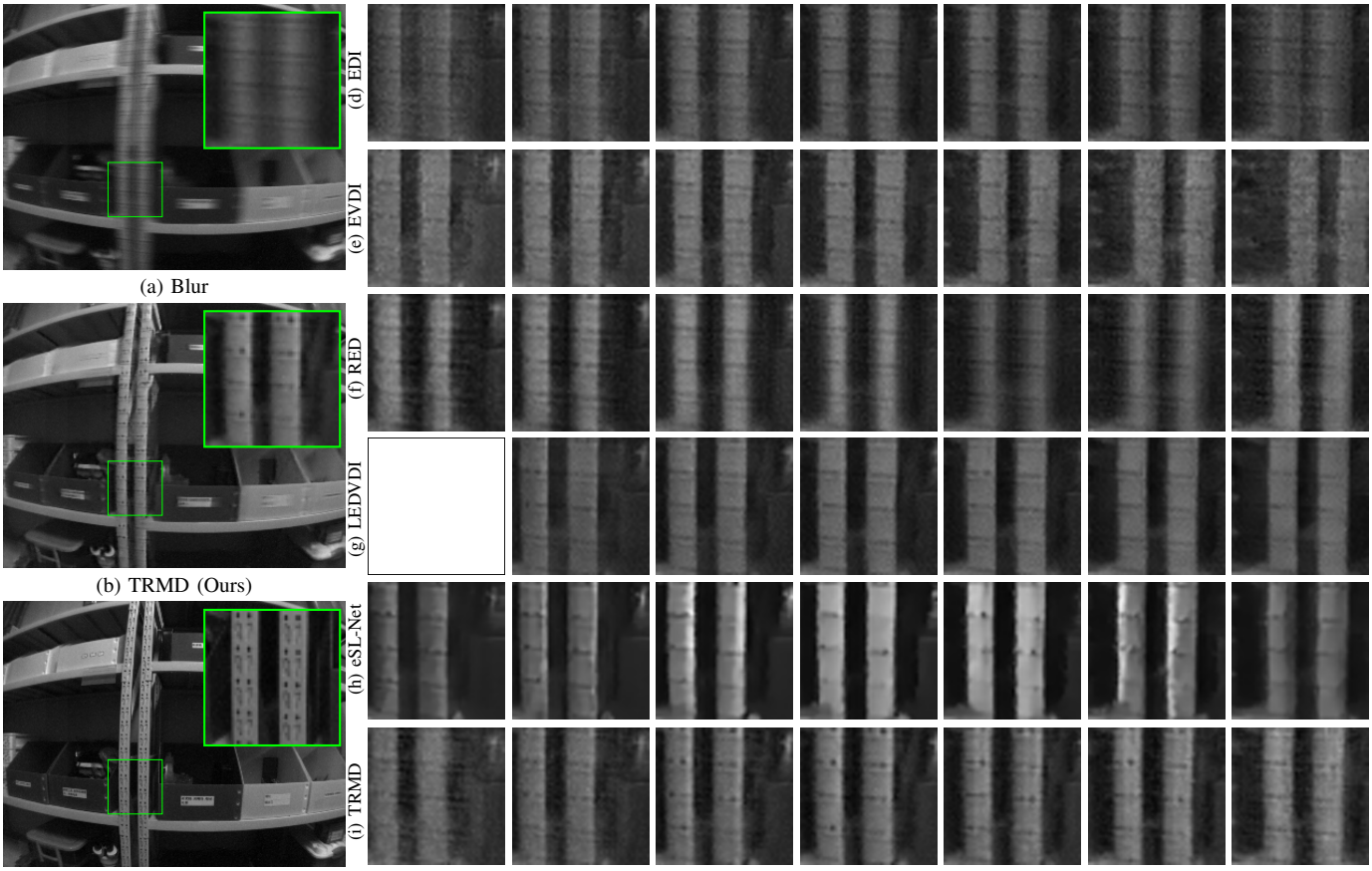
Meanwhile, in the single image deblurring task as depicted in Figs. 8 and 9, our method exhibits an exceptional generalizability, remaining effective even in scenarios whose event format is inconsistent with that encountered during training. Specifically, when restoring the support of the calibration board, EFNet introduces some distortions in the shape, while other methods fail to recover the appearance. In contrast, our method successfully reconstructs the shape and position of the board support and outperforms other methods in terms of restoration quality.

D. Ablation Study

To verify the superiority of our proposed techniques, which include multi-modality input, two-stage and residual strategy, as well as the FEM and FFM in the RE-Net, we conduct comprehensive ablation experiments on the synthetic dataset GOPRO and the quantitative results are listed in Table II.

The ablation experiments are divided into I, II, and III three groups for analyzing the performance of different components in TRMD. Group I serves as the baseline model, aiming to progressively demonstrate the advantages of multi-modality input, residual strategy, and two-stage framework. Experiments of Group II are built upon the previous two strategies, with the goal of demonstrating the roles of FEM and FFM. Group III is carried out based on FEM and FFM, further reinforcing the superiority of the residual strategy and the two-stage framework.

For experiments conducted without the residual strategy, we utilize the network to directly learn the mapping from the input to the sharp sequence. On the other hand, for residual-based experiments without the two-stage strategy, we adopt the framework described in [17] for comparison. This



From top to bottom correspond to EDI(d), EVDI(e), RED(f), LEDVDI(g), eSL-Net(h) and TRMD(i).

Fig. 7: Qualitative comparison for the sequence reconstruction task on the REBlur dataset.

TABLE II: Ablation study on the GoPro dataset. Note that experiments (I-5, II-1) and (II-3, III-3) are repeated for better comparison.

ID	Inputs		Strategy		Module		Single frame deblurring			Sequence reconstruction		
	Blur	Event	Residual	Two-Stage	FEM	FFM	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
I-1	✓	×	×	×	×	×	29.717	0.9250	0.1424	25.978	0.8700	0.1595
I-2	×	✓	×	×	×	×	16.704	0.7464	0.2322	16.670	0.7385	0.2322
I-3	✓	✓	×	×	×	×	34.702	0.9740	0.0352	33.017	0.9628	0.0423
I-4	✓	✓	✓	×	×	×	35.437	0.9772	0.0305	33.152	0.9638	0.0390
I-5	✓	✓	✓	✓	×	×	35.615	0.9790	0.0252	33.300	0.9654	0.0376
II-1	✓	✓	✓	✓	×	×	35.615	0.9790	0.0252	33.300	0.9654	0.0376
II-2	✓	✓	✓	✓	✓	×	36.190	0.9812	0.0238	33.821	0.9702	0.0303
II-3	✓	✓	✓	✓	✓	✓	36.678	0.9830	0.0200	34.370	0.9735	0.0275
III-1	✓	✓	×	×	✓	✓	35.414	0.9769	0.0299	34.212	0.9721	0.0310
III-2	✓	✓	✓	×	✓	✓	36.341	0.9812	0.0217	34.212	0.9719	0.0295
III-3	✓	✓	✓	✓	✓	✓	**36.678**	**0.9830**	**0.0200**	**34.370**	**0.9735**	**0.0275**

framework directly estimates the intensity difference between the blurry image and sharp image to restore the middle sharp frame. Additionally, it incorporates a sequential addition of the residual to the deblurred image for sequence reconstruction. As for the network architecture, our RE-Net consists of FEM, FFM, and REM, where REM serves as the backbone and other modules are integrated in a cascading manner.

For Group I, we take the blurry image, events, and their combination as the input of the RE-Net, respectively. Specifically, we fuse two modalities by concatenating them along the channel dimension as done in [4]. For example, assuming the input size of the blurry image $B \in \mathbb{R}^{1\times H\times W}$ and the event stream $E \in \mathbb{R}^{2N\times H\times W}$, we concatenate two modalities along the channel dimension to obtain a tensor of size $(2N+1)\times H\times W$, which is further fed into the RE-Net to estimate the sharp sequence. The evaluation of their performance is presented in Experiments I-1, I-2, and I-3, respectively. The combination of a single blurry image and events allows for the synergistic utilization of motion and texture features, leading to superior performance compared to using either modality alone. Moreover, the superior performance of Experiment I-5 compared to I-4 and I-3 is mainly attributed to the utilization of residual sequences as intermediate variables, which captures more comprehensive information and provides stronger supervision signals for RE-Net compared to [17]. This decoupling of the complex motion

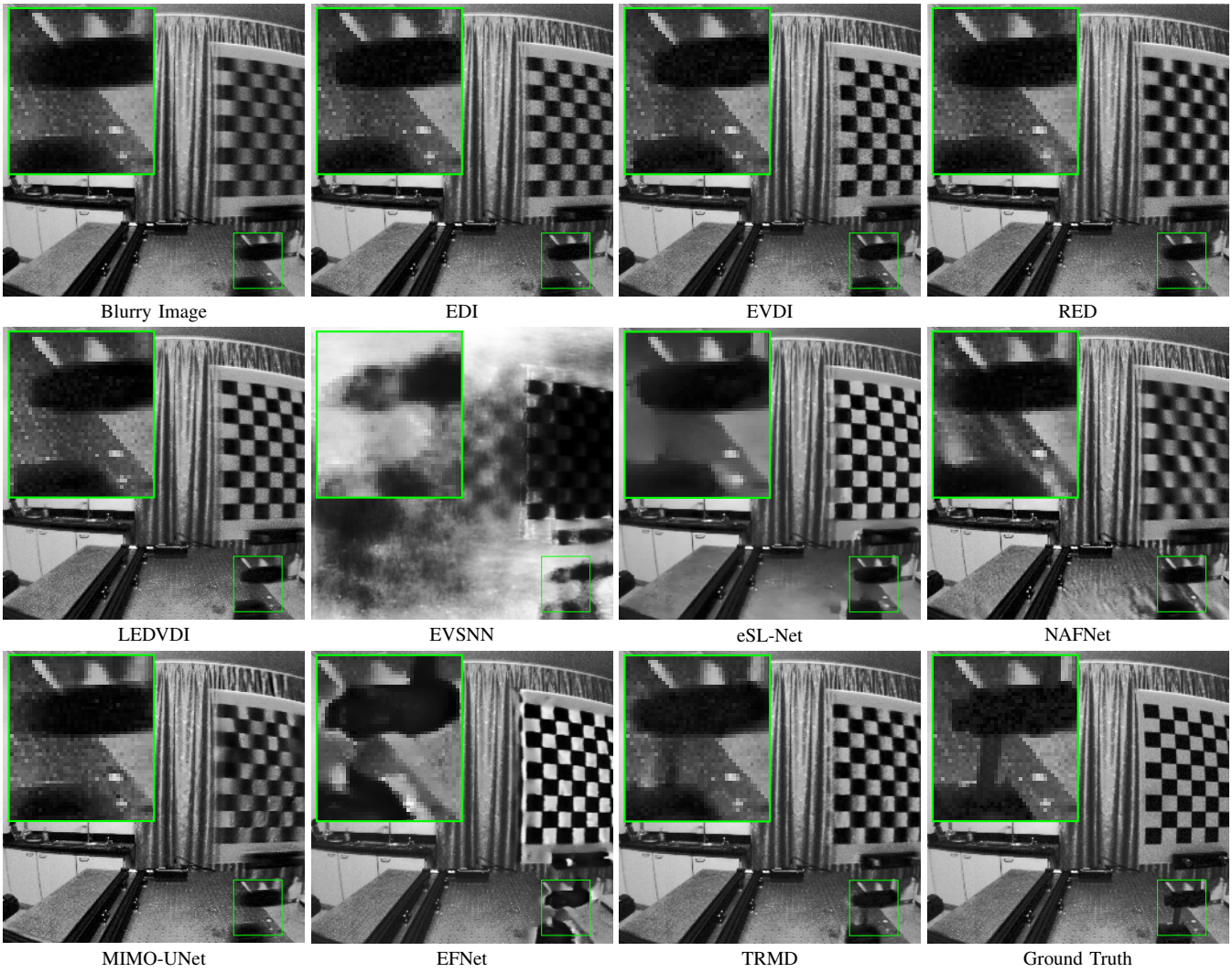


Fig. 8: Qualitative comparison for the single image deblurring task on the REBlur dataset.

blur process contributes to the improved performance achieved by a two-stage residual-based strategy.

For Group II, with the input being the fusion of multi-modality, we sequentially embed the FEM and FFM into the RE-Net as done in Experiments II-1, II-2, and II-3. In Experiment II-2, we initially process the two modalities separately by FEM. Next, we concatenate their outputs along the channel dimension before feeding them into the REM, which is the same as Group I. The superior performance in Experiment II-2 over Experiment II-1 highlights the significance of appropriate feature extraction and alignment before feature fusion. Experiment II-3, which further incorporates the FFM into RE-Net, exhibits better results than Experiment II-4, indicating the benefits of an appropriate modality fusion mechanism for motion deblurring tasks.

To further confirm the proposed two-stage deblurring framework's effectiveness under the RE-Net and multi-modality condition, we conduct Experiments III-1, III-2, and III-3. The improvement in the performance of Experiment III-3 compared to Experiments III-1 and III-2 provides additional evidence for the benefits of residual learning and the two-stage deblurring framework.

E. Discussion on the number N

In Section III, we elaborate on how Eq. (11) serves as a discrete approximation of the continuous integral, with $2N+1$ signifying the number of latent residual frames to be estimated. To further investigate the impact of N on the quality of the restored sharp sequence, we conduct experiments for $N = 1, 2, 3$ respectively. For the blurry image in the GOPRO dataset, we average 7 consecutive sharp images to synthesize one, which implies that there are multiple experimental combinations for $N = 1, 2$ based on positions of the estimated residuals. Taking the $N = 1$ case as an instance, we can restore the residuals of the 0-th frame and the 6-th frame or the residuals of the 1-st frame and the 5-th frame. Altogether, we conduct a comprehensive set of seven experiments as outlined in Table III. The corresponding PSNR values for each image under different experimental combinations are visually represented in Fig. 10.

From Fig. 10, it is evident that the experimental combination with $N = 3$ achieves the best performance across all images, which is consistent with our expectations. Moreover, an intriguing observation can be made from the combination $N = 1$. Combination A-III exhibits inferior performance be-

Fig. 9: Qualitative comparison for the single image deblurring task on the REBlur dataset.

TABLE III: Experimental combinations for different values of N. ID denotes the ID-th frame of the restoration sequence.

N	ID	0	1	2	3	4	5	6
1	A-I	✓			✓			✓
1	A-II		✓		✓		✓	
1	A-III			✓	✓	✓		
2	B-I	✓	✓		✓		✓	✓
2	B-II	✓		✓	✓	✓		✓
2	B-III		✓	✓	✓	✓	✓	
3	C	✓	✓	✓	✓	✓	✓	✓

cause the estimated residuals only capture a minimal portion of the temporal motion features within the exposure time. On the other hand, combination A-I yields poor results primarily because of the high level of event noise, which hinders accurate residual estimation and consequently affects the restoration quality of the clear sequence. In contrast, combination A-II presents the best performance among the $N = 1$ combinations, which can be attributed to its evenly distributed sampling pattern. Simultaneously, the overall restoration of the image sequence exhibits a characteristic pattern of gradually decreasing quality from the center towards both sides. This phenomenon can be attributed to the increasing accumulation of event noise in voxel-level events as the sampling time deviates further from the central moment. Consequently, this accumulation of noise reduces the accuracy of residual estimation, further deteriorating the quality of the reconstructed frame.

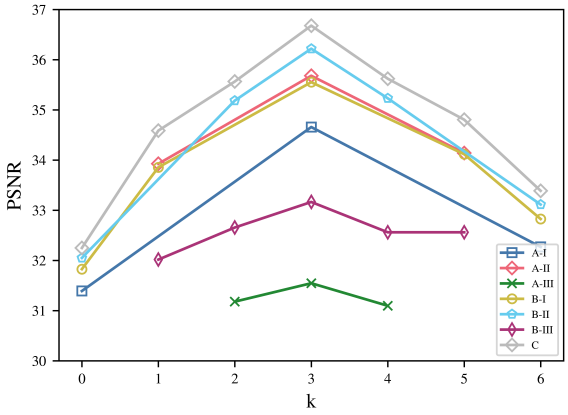


Fig. 10: The visualization showcasing the performance of recovering each frame during the exposure, where k denotes the k-th frame of the reconstructed sequence.

Finally, we qualitatively discuss the impact of N on the REBlur dataset, as shown in Fig. 11. The results show that as N increases, the artifacts and noise generated in the restored image decrease, which meets the quantitative analysis on the GOPRO.

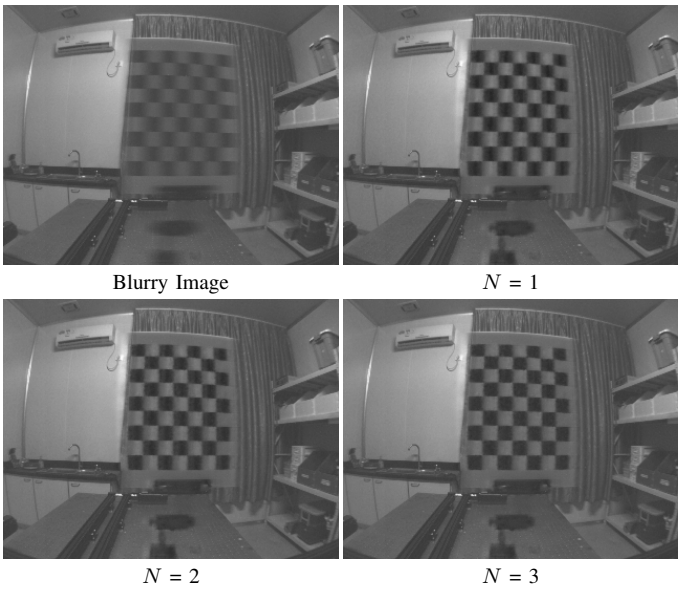


Fig. 11: Qualitative comparison for different values of N on the real-world dataset REBlur. Specifically, $N = 1$ corresponds to the best performance experiment combination A-II as shown in Fig. 10, while $N = 2$ corresponds to B-II.

F. Discussion on the residual

In this subsection, we aim to demonstrate the superiority of the proposed residual form as described in Eq. (9) over the fraction residual employed in [17, 18], which is defined in Eq. (6). Before calculating the residual, we convert the color sharp sequence with the pixel range [0,255] into the grayscale images with the normalized range of [0,1]. Considering the significant issue of the fraction residual in low-light regions as discussed in Section III, we are forced to apply a clip operation to the fraction residual with different thresholds for better visualization, as shown in Fig. 12.

It can be observed that when the threshold is small (2 in this example), although the image's overall brightness appears relatively balanced, the details in certain areas become flattened. On the contrary, when a large threshold (10 in this example) is used, although the overall details of the residual are preserved, its gigantic spatial intensity variation poses significant challenges for RE-Net in estimating the residual accurately. This phenomenon can be inferred from the training and testing curves of RE-Net as depicted in Fig. 13. The training curve indicate that when the threshold for the fraction residual is large, the network's training generates large oscillations. On the other hand, utilizing a small threshold hinders the accurate estimation of the residual by the RE-Net, consequently affecting the overall image restoration performance. Therefore, compared to the fraction residual form, our proposed residual achieves a better balance between these two factors, thus resulting in more stable training and improved performance on the test set.

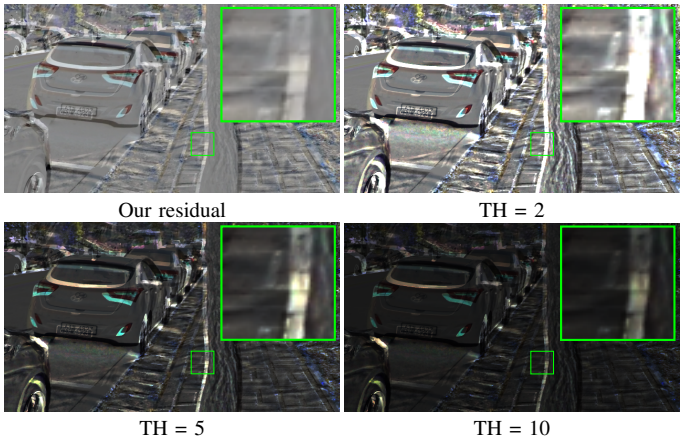


Fig. 12: The visual comparison of the proposed residual $R(k)$ and the fraction residual $E(k)$ with different thresholds, where "TH" denotes the threshold of the fraction residual. Details are zoomed in to illustrate the phenomenon that fraction residual with low threshold tends to smooth out the details in the figure.

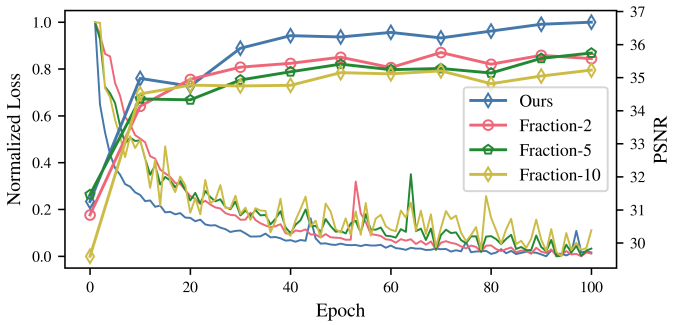


Fig. 13: The comparison of the training (Normalized Loss) and testing curves (PSNR) for the proposed residual $R(k)$ and fraction residual $E(k)$ with different TH (thresholds), which corresponds to the visual comparison in Fig. 12 respectively. Note that the falling curve corresponds to the training loss, while the rising curve corresponds to the PSNR metric. Normalization technique is employed on the training loss to standardize them to the same scale for better comparison.

G. Limitation of the TRMD

Although TRMD has achieved promising results on the synthetic dataset GOPRO, its performance tends to degrade when directly applied to the REBlur dataset. This is mainly due to the inherent differences in scene characteristics and parameter settings between synthetic and real datasets. To address this issue, a straightforward approach, as demonstrated in [14], is to fine-tune TRMD in the REBlur dataset. However, TRMD requires paired data consisting of single frame blurry images and the corresponding sharp video sequence, which is challenging to capture simultaneously in the real world.

Therefore, in future work, we plan to extend the TRMD to be a self-supervised training framework similar to [19, 26]. This would enable TRMD to achieve better generaliza-

tion performance on real-world datasets. By leveraging self-supervision, we can alleviate the reliance on paired training data and enhance the robustness and adaptability of TRMD in real-world scenarios.

V. Conclusion

In this paper, we propose a two-stage residual-based motion deblurring framework, which can be applied to enhance the visual quality of event-based videos, making them more suitable for applications such as autonomous driving, robotics, augmented reality and other low-level tasks[48–51]. We transform a single blurry image into a sequence of sharp images with the help of events. Benefiting from effective integration with the event-based motion deblurring physical model, our two-stage deblurring strategy achieves superior performance compared to previous one-stage reconstruction methods. As for the network architecture, the designed cross-modal feature fusion module effectively fuses the spatial-temporal features of both modalities. Furthermore, our novel proposed residual form exhibits a more stable convergence during training than the previous residual form, resulting in better motion restoration performance in testing. Quantitative and qualitative experiments demonstrate that our method outperforms previous state-of-the-art methods. However, the performance of our method trained on the synthetic dataset deteriorates in real blur scenes due to data inconsistency and the blur generation mechanism. In the future, we plan to design a self-supervised framework to alleviate this problem.

References

[1] B. Peterson, *Understanding exposure: how to shoot great photographs with any camera*. AmPhoto books, 2016. 1

[2] K. Zhang, W. Ren, W. Luo, W.-S. Lai, B. Stenger, M.-H. Yang, and H. Li, "Deep image deblurring: A survey," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2103–2130, 2022. 1

[3] S. Zhao, Z. Zhang, R. Hong, M. Xu, H. Zhang, M. Wang, and S. Yan, "Crnet: Unsupervised color retention network for blind motion deblurring," in *ACMMM*, 2022, pp. 6193–6201. 1

[4] C. Song, Q. Huang, and C. Bajaj, "E-cir: Event-enhanced continuous intensity recovery," in *CVPR*, 2022, pp. 7803–7812. 1, 2, 3, 4, 5, 7, 8, 11

[5] S. Zhao, Z. Zhang, R. Hong, M. Xu, Y. Yang, and M. Wang, "Fcl-gan: A lightweight and real-time baseline for unsupervised blind image deblurring," in *ACMMM*, 2022, pp. 6220–6229. 1

[6] S.-J. Cho, S.-W. Ji, J.-P. Hong, S.-W. Jung, and S.-J. Ko, "Rethinking coarse-to-fine approach in single image deblurring," in *ICCV*, 2021, pp. 4641–4650. 1, 2, 8, 10

[7] K. Kim, S. Lee, and S. Cho, "Mssnet: Multi-scale-stage network for single image deblurring," in *ECCV*. Springer, 2023, pp. 524–539.

[8] Y. Liu, F. Fang, T. Wang, J. Li, Y. Sheng, and G. Zhang, "Multi-scale grid network for image deblurring with high-frequency guidance," *IEEE TMM*, vol. 24, pp. 2890–2901, 2021.

[9] X. Mao, Y. Liu, F. Liu, Q. Li, W. Shen, and Y. Wang, "Intriguing findings of frequency selection for image deblurring," in *AAAI*, vol. 37, no. 2, 2023, pp. 1905–1913. 2

[10] M. Suin, K. Purohit, and A. Rajagopalan, "Spatially-attentive patch-hierarchical network for adaptive motion deblurring," in *CVPR*, 2020, pp. 3606–3615. 1

[11] T. Serrano-Gotarredona and B. Linares-Barranco, "A 128 × 128 1.5% contrast sensitivity 0.9% fpn 3 μs latency 4 mw asynchronous frame-free dynamic vision sensor using transimpedance preamplifiers," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 3, pp. 827–838, 2013. 1, 2

[12] C. Brandli, R. Berner, M. Yang, S.-C. Liu, and T. Delbruck, "A 240× 180 130 db 3 μs latency global shutter spatiotemporal vision sensor," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 10, pp. 2333–2341, 2014.

[13] C. Posch, D. Matolin, and R. Wohlgenannt, "A qvga 143 db dynamic range frame-free pwm image sensor with lossless pixel-level video compression and time-domain cds," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 1, pp. 259–275, 2010. 1, 2

[14] L. Sun, C. Sakaridis, J. Liang, Q. Jiang, K. Yang, P. Sun, Y. Ye, K. Wang, and L. V. Gool, "Event-based fusion for motion deblurring with cross-modal attention," in *ECCV*. Springer, 2022, pp. 412–428. 1, 2, 3, 4, 5, 6, 8, 10, 14

[15] P. Vitoria, S. Georgoulis, S. Tulyakov, A. Bochicchio, J. Erbach, and Y. Li, "Event-based image deblurring with dynamic motion awareness," in *ECCV*. Springer, 2023, pp. 95–112. 1, 2, 7, 8

[16] L. Zhang, H. Zhang, J. Chen, and L. Wang, "Hybrid deblur net: Deep non-uniform deblurring with event camera," *IEEE Access*, vol. 8, pp. 148 075–148 083, 2020. 1, 2

[17] H. Chen, M. Teng, B. Shi, Y. Wang, and T. Huang, "A residual learning approach to deblur and generate high frame rate video with an event camera," *IEEE TMM*, 2022. 1, 2, 3, 4, 7, 8, 10, 11, 14

[18] S. Lin, J. Zhang, J. Pan, Z. Jiang, D. Zou, Y. Wang, J. Chen, and J. Ren, "Learning event-driven video deblurring and interpolation," in *ECCV*. Springer, 2020, pp. 695–710. 1, 2, 3, 4, 5, 7, 8, 10, 14

[19] X. Zhang and L. Yu, "Unifying motion deblurring and frame interpolation with events," in *CVPR*, 2022, pp. 17 765–17 774. 3, 4, 7, 8, 10, 14

[20] L. Pan, C. Scheerlinck, X. Yu, R. Hartley, M. Liu, and Y. Dai, "Bringing a blurry frame alive at high frame-rate with an event camera," in *CVPR*, 2019, pp. 6820–6829. 2, 3, 6, 8, 10

[21] B. Wang, J. He, L. Yu, G.-S. Xia, and W. Yang, "Event enhanced high-quality image recovery," in *ECCV*. Springer, 2020, pp. 155–171. 1, 3, 4, 7, 8, 10

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241. 2, 7

[23] S. Nah, T. H. Kim, and K. M. Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *CVPR*, July 2017. 2, 7

[24] S. Nah, S. Son, and K. M. Lee, "Recurrent neural networks with intra-frame iterations for video deblurring," in *CVPR*, 2019, pp. 8102–8111. 2

[25] X. Zhang, R. Jiang, T. Wang, and J. Wang, "Recursive neural network for video deblurring," *CSVT*, vol. 31, no. 8, pp. 3025–3036, 2020. 2

[26] F. Xu, L. Yu, B. Wang, W. Yang, G.-S. Xia, X. Jia, Z. Qiao, and J. Liu, "Motion deblurring with real events," in *ICCV*, 2021, pp. 2583–2592. 2, 8, 10, 14

[27] L. Liu, J. Chen, H. Wu, G. Li, C. Li, and L. Lin, "Cross-modal collaborative representation learning and a large-scale rgbt benchmark for crowd counting," in *CVPR*, 2021, pp. 4823–4833. 3, 6

[28] Y. Zhang, S. Choi, and S. Hong, "Spatio-channel attention blocks for cross-modal crowd counting," in *Proceedings of the Asian Conference on Computer Vision*, 2022, pp. 90–107.

[29] H. Liu, T. Lu, Y. Xu, J. Liu, and L. Wang, "Learning optical flow and scene flow with bidirectional camera-lidar fusion," *arXiv preprint arXiv:2303.12017*, 2023. 6

[30] G. Zhang, J. Liu, Y. Chen, Y. Zheng, and H. Zhang, "Multi-biometric unified network for cloth-changing person re-

identification," *TIP*, vol. 32, pp. 4555–4566, 2023.
[31] G. Zhang, Y. Ge, Z. Dong, H. Wang, Y. Zheng, and S. Chen, "Deep high-resolution representation learning for cross-resolution person re-identification," *TIP*, vol. 30, pp. 8913–8925, 2021. 3
[32] T. Kim, J. Lee, L. Wang, and K.-J. Yoon, "Event-guided deblurring of unknown exposure time videos," in *ECCV*. Springer, 2022, pp. 519–538. 3
[33] W. Shang, D. Ren, D. Zou, J. S. Ren, P. Luo, and W. Zuo, "Bringing events into video deblurring with non-consecutively blurry frames," in *ICCV*, 2021, pp. 4531–4540. 3
[34] D. Yang and M. Yamac, "Motion aware double attention network for dynamic scene deblurring," in *CVPR*, 2022, pp. 1113–1123. 3, 7
[35] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis *et al.*, "Event-based vision: A survey," *IEEE TPAMI*, vol. 44, no. 1, pp. 154–180, 2020. 3, 4, 6
[36] A. Galassi, M. Lippi, and P. Torroni, "Attention in natural language processing," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 10, pp. 4291–4308, 2020. 6
[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017. 6
[38] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu, "Attention mechanisms in computer vision: A survey," *Computational Visual Media*, vol. 8, no. 3, pp. 331–368, 2022. 6
[39] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *CVPR*, 2018, pp. 7794–7803. 6
[40] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *ECCV*, 2018, pp. 3–19. 7
[41] B. Son, Y. Suh, S. Kim, H. Jung, J.-S. Kim, C. Shin, K. Park, K. Lee, J. Park, J. Woo *et al.*, "4.1 a 640× 480 dynamic vision sensor with a 9μm pixel and 300meps address-event representation," in *2017 IEEE International Solid-State Circuits Conference (ISSCC)*. IEEE, 2017, pp. 66–67. 7
[42] Z. Huang, T. Zhang, W. Heng, B. Shi, and S. Zhou, "Real-time intermediate flow estimation for video frame interpolation," in *ECCV*. Springer, 2022, pp. 624–642. 7
[43] H. Rebecq, D. Gehrig, and D. Scaramuzza, "Esim: an open event camera simulator," in *Conference on robot learning*. PMLR, 2018, pp. 969–982. 7
[44] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018, pp. 586–595. 8
[45] L. Chen, X. Chu, X. Zhang, and J. Sun, "Simple baselines for image restoration," in *ECCV*. Springer, 2022, pp. 17–33. 8, 10
[46] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "High speed and high dynamic range video with an event camera," *IEEE TPAMI*, vol. 43, no. 6, pp. 1964–1980, 2019. 8, 10
[47] L. Zhu, X. Wang, Y. Chang, J. Li, T. Huang, and Y. Tian, "Event-based video reconstruction via potential-assisted spiking neural network," in *CVPR*, 2022, pp. 3594–3604. 8, 10
[48] Z. Zhang, Y. Wei, H. Zhang, Y. Yang, S. Yan, and M. Wang, "Data-driven single image deraining: A comprehensive review and new perspectives," *Pattern Recognition*, p. 109740, 2023. 15
[49] Y. Li, Y. Zhang, R. Timofte, L. Van Gool, Z. Tu, K. Du, H. Wang, H. Chen, W. Li, X. Wang *et al.*, "Ntire 2023 challenge on image denoising: Methods and results," in *CVPR*, 2023, pp. 1904–1920.
[50] F.-A. Vasluianu, T. Seizinger, R. Timofte, S. Cui, J. Huang, S. Tian, M. Fan, J. Zhang, L. Zhu, X. Wei *et al.*, "Ntire 2023 image shadow removal challenge report," in *CVPR*, 2023, pp. 1788–1807.
[51] Y. Dai, C. Li, S. Zhou, R. Feng, Q. Zhu, Q. Sun, W. Sun, C. C. Loy, J. Gu, S. Liu *et al.*, "Mipi 2023 challenge on nighttime flare removal: Methods and results," in *CVPR*, 2023, pp. 2852–2862. 15

Kang Chen is currently working toward the B.S. degree with the Communication Engineering, Wuhan University, Wuhan, China. His research interests include computer vision and neuromorphic computation.

Lei Yu received his B.S. and Ph.D. degrees in signal processing from Wuhan University, Wuhan, China, in 2006 and 2012, respectively. From 2013 to 2014, he has been a Postdoc Researcher with the VisAGeS Group at the Institut National de Recherche en Informatique et en Automatique (INRIA) for one and a half years. He is currently working as a professor at the School of Electronics and Information, Wuhan University, Wuhan, China. From 2016 to 2017, he has also been a Visiting Professor at Duke University for one year. He has been working as a guest professor in the École Nationale Supérieure de l'Électronique et de ses Applications (ENSEA), Cergy, France, for one month in 2018. His research interests include signal processing, neuromorphic vision, and image computation.