Non-Coaxial Event-guided Motion Deblurring with Spatial Alignment

Hoonhee Cho, Yuhwan Jeong, Taewoo Kim, and Kuk-Jin Yoon
Korea Advanced Institute of Science and Technology
{gnsgnsgml,jeongyh98,intelpro,kjyoon}@kaist.ac.kr

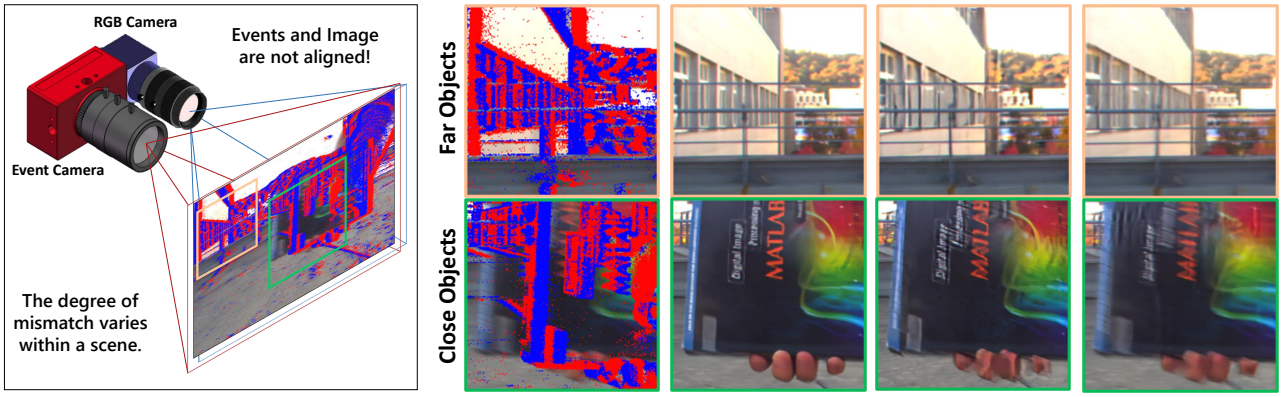

Figure 1: The proposed method aims to restore a sharp RGB image from a blurry image by utilizing events provided by a separate event camera, as shown in (a). Previous event-based motion deblurring methods assume that the image and event are spatially aligned perfectly. However, in real-world images, except for low-quality APS images, there are misaligned pixels with events even when the baseline of two cameras is minimized, as in (b). In general, the misalignment between these pixels depends on the distance to camera. Therefore, as shown in (d) and (e), for distant objects that are partially aligned, the state-of-the-art event-based image deblurring method UEVD [28] performs better than the only image-based method MPRNet [6]. Instead, even when utilizing additional information, the event-based model often fails to preserve details for objects that are close, compared to the image-based model. On the other hand, our approach robustly restores sharp textures for both near and distant objects by matching each pixel between the image and event at the feature-level, as in (c).

Abstract

Motion deblurring from a blurred image is a challenging computer vision problem because frame-based cameras lose information during the blurring process. Several attempts have compensated for the loss of motion information by using event cameras, which are bio-inspired sensors with a high temporal resolution. Even though most studies have assumed that image and event data are pixel-wise aligned, this is only possible with low-quality active-pixel sensor (APS) images and synthetic datasets. In real scenarios, obtaining per-pixel aligned event-RGB data is technically challenging since event and frame cameras have different optical axes. For the application of the event camera, we propose the first Non-coaxial Event-guided Image Deblurring (NEID) approach that utilizes the camera setup composed of a standard frame-based camera with a non-coaxial single event camera. To consider the per-pixel alignment between the image and event without additional devices, we propose the first NEID network that spatially aligns events to images while refining the image features from temporally dense event features. For training and evaluation of our network, we also present the first large-scale dataset, consisting of RGB frames with non-aligned events aimed at a breakthrough in motion deblurring with an event camera. Extensive experiments on various datasets demonstrate that the proposed method achieves significantly better results than the prior works in terms of performance and speed, and it can be applied for practical uses of event cameras.

project: https://sites.google.com/view/neid2023

1. Introduction

Motion blur occurs due to the motions during the exposure time since a frame-based camera records the scene during the exposure time and outputs the averaged signal.

The inverse problem, called deblurring, is restoring a sharp image given a blurry image [41, 35, 5, 13, 67, 71, 1, 22, 30, 39, 65]. Previous learning-based deblurring methods are designed to be sophisticated [6, 69] to improve the deblurred image quality, leading to enormous complexity.

With the advent of event cameras, event-based motion deblurring methods have been proposed to overcome the loss of motion information in the frame-based camera [43, 42, 21, 25, 31, 62, 66, 49, 72, 61]. The event camera asynchronously provides event data with low latency, which has no motion blur even in extreme motion and contains both texture and motion information over continuous time. Despite these potentials of event data, the critical issue is that even the most recent research still exploits a strong assumption that events and images are pixel-wise aligned. However, in real scenarios, it can be challenging to obtain per-pixel aligned event-RGB data unless the event camera provides active-pixel sensor (APS) images [3]; because event and frame cameras have different optical axes. Since APS images are generally of low image quality [16, 2], some works use event simulators (*e.g*. ESIM [46]) on high-frame-rate video datasets (*e.g*. GoPro [37]) to generate synthetic event streams for restoring clean RGB images. However, as mentioned in [19, 52, 44, 4], synthetic event data is far from the output of an actual event camera. A model that has been trained effectively on a synthetic dataset may experience significant performance degradation when evaluated on a real-event dataset.

In the real world, two distinct types of cameras must be used together to acquire high-quality event stream and RGB images simultaneously (see (a) of Fig. 1). Since the two cameras are different devices, they have different optical axes and field-of-views (FoV). It means they are not aligned pixel-wisely. For video interpolation, a task different from motion deblurring, existing works used a feature-based homography warping process [58] or an additional device (*e.g*. *beamsplitter*) [57] for pixel-wise alignment between the two cameras. However, feature-based alignment between events and images [36] is too complicated due to the sequence of image reconstruction, feature matching, and global homography warping. Furthermore, this process only works when the camera is static; the object is front-paralleled in a limited range. In non-coaxial camera setups, the pixel misalignment between two cameras varies with the 3D information of each scene. Therefore, the pixel alignment between two cameras obtained by the homography can not perfectly align all the pixels of both near and faraway objects, even calculated scene by scene as in [58]. In other words, feature-based homography cannot be a general solution to cover most situations. On the other hand, an additional device for pixel alignment could not achieve a compact design and larger FoV, leading to many limitations in real-world applications, such as mobile devices.

We aim to utilize events from non-coaxial event cameras for motion deblurring without homography/additional devices. As shown in (b) of Fig. 1, despite of close placement of the RGB camera and the event camera, the event and image are not aligned. Furthermore, the extent of unalignment is slightly different for each scene and even for each object in the same scene. Therefore, we introduce a Non-coaxial Event-guided Image Deblurring (NEID) task that can unlock the potential of the event camera in a real scenario. NEID aims to spatially align the image and event at the feature level during motion deblurring. To use pixel-wise misaligned events, we need a framework to align events with images, even if the images are blurry.

To this end, we propose a novel framework called Non-coaxial Event-guided Deblurring Network (NED-Net). As a key component for managing alignment between image and event features in the pipeline, we propose the Attention-based Deformable Align (ADA) module. The module leverages both structure-based coarse alignment and patch-based fine alignment. From these two cascade alignment processes, events and images can be aligned well at the feature level despite the distinctive modality differences. After global spatial alignment, the proposed Local Score-based Aggregation (LSA) computes the similarity score with the surrounding image features to remove the unnecessary event features and aggregate confident events to improve the effect of deblurring. Finally, we develop the Cross-Channel Interaction (CCI) module for texture enhancement. In addition, we first present the Non-Coaxial Events and RGB (NCER) dataset consisting of 83 scenes collected with a sophisticated camera setup with temporally synchronized frames and events. This dataset provides multiple scenes with dynamic motion at varying distances suitable for the NEID task. Unlike the existing event-guided motion deblurring dataset [28, 25, 52, 55], the NCER dataset provides high-resolution event data and high-quality RGB images, not APS images.

2. Related Works

Frame-based Motion Deblurring. Some attempts [41, 35, 13, 67, 71, 1, 22, 30] have tried to restore the sharp image given a blurred image without knowing the blur kernel. They have succeeded in modeling simple motions; however, it is still hard to recover under complex motions. With the deep learning-based approach, some methods [65, 39, 40, 37, 69, 70] have tried to resolve complex motions using the convolutional neural network (CNN) and recurrent neural network (RNN) [65, 24, 38, 75]. Recent state-of-the-art methods [69, 6] have adopted a multi-stage scheme composed of smaller sub-tasks. Especially, Chen *et al*. [6] proposed the extended Instance Normalization [59] as an enhancement method of restoration. However, despite these advances, the plausible recovery from a single motion

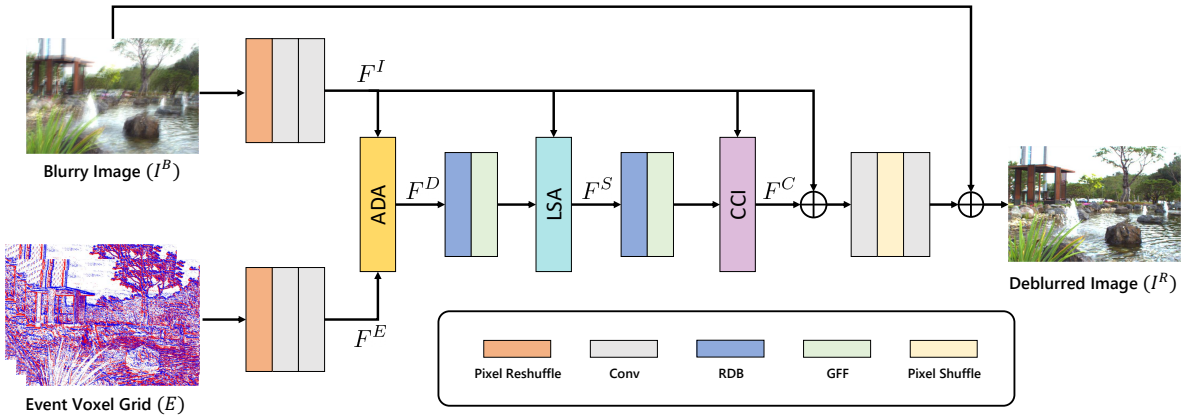

Figure 2: The overview of Non-coaxial Event-based Deblurring Network (NED-Net). NED-Net contains the residual dense blocks (RDB), and global feature fusion (GFF) proposed in [26]. Our main contributions are as follows: Attention-based Deformable Align (ADA), Local Score-based Aggregation (LSA), and Cross-Channel Interaction (CCI) module.

blurred frame is still challenging due to the loss of motion information in the process of degradation.

Event-based Motion Deblurring. To address the lack of information, one alternative is to use additional sensors, such as [76, 29, 57]. Event cameras, a novel bio-inspired sensor, can record the temporally dense brightness change information. Event cameras have been investigated for their potential in diverse environments and tasks [8, 10, 9, 27, 77, 51, 20, 7]. When it comes to deblurring, [42, 43] succeeded in modeling relationships between events, a sharp image, and a blurry image. Haoyu *et al*. [21] also defined the relationship between a blurry image and events, using the modified U-Net [48] with a global residual connection. Other approaches use neural networks to learn the relationships directly between blurry and sharp images with the guide of events. Jiang *et al*. [25] proposed a recurrent architecture with an event filter, which generates sharp boundary prior. Lin *et al*. [31] resolve both problems of deblurring and interpolation using a dynamic filtering method to deal with spatially and temporally variant thresholds that trigger events. Wang *et al*. [62] developed the sparse learning method, which is robust to noise. Recent works have aimed to design more sophisticated architecture [55, 74] and better solve real situations, such as non-consecutive blurry videos [49], unknown exposure time videos [28], and data inconsistency [66]. These works generally assume the event data is per-pixel aligned with the image; however, this assumption is not valid for real cameras because an event camera and an RGB camera are not positioned co-axially. *In this paper, we aim to utilize events for motion deblurring without any additional devices and a hand-crafted alignment process. To the best of our knowledge, ours is the first NEID framework that utilizes a non-coaxial event camera for deblurring a real-world RGB image.*

3. Motion Deblurring with Non-coaxial Events

3.1. Framework Overview

Our goal is to restore a sharp image (I^S) given a blurry image (I^B) and an event (E) corresponding to the exposure time. As shown in Fig. 2, our proposed NED-Net consists of three modules: Attention-based Deformable Align (ADA), Local Score-based Aggregation (LSA), and Cross-Channel Interaction (CCI) module.

First, we extract features from images and events. Then, to account for computation cost while widening the receptive field, we rearrange the image and event through pixel reshuffle [50]. In addition, we adopt the residual dense blocks (RDB) and global feature fusion (GFF) [26] to exploit whole hierarchical features. Then, the extracted image features (F^I) and event features (F^E) are transferred to the ADA module for generating spatially aligned event features (F^D). After the global alignment, LSA module generates the local refined features (F^S) by aggregating the aligned event features (F^D) with the image features (F^I). Finally, the CCI module extracts rich context through channel-wise relation between the two modalities. Then, the output of CCI module (F^C) is used to restore a sharp image (I^R) through pixel shuffle upsampling [50].

3.2. Attention-based Deformable Align

As shown in Fig. 5, events and images are pixel-wise mismatched, varying from scene to scene and object to object. In the case of images, flow estimation [56, 54] is commonly used to estimate per-pixel matching between two adjacent frames. However, in our setup, we need to find the per-pixel matching between the image and the event, which is challenging by solely relying on flow estimation. Without access to ground truths, optical flow networks are generally trained with warping-based photometric loss [34], but it is challenging to apply this due to the sparse nature of events.

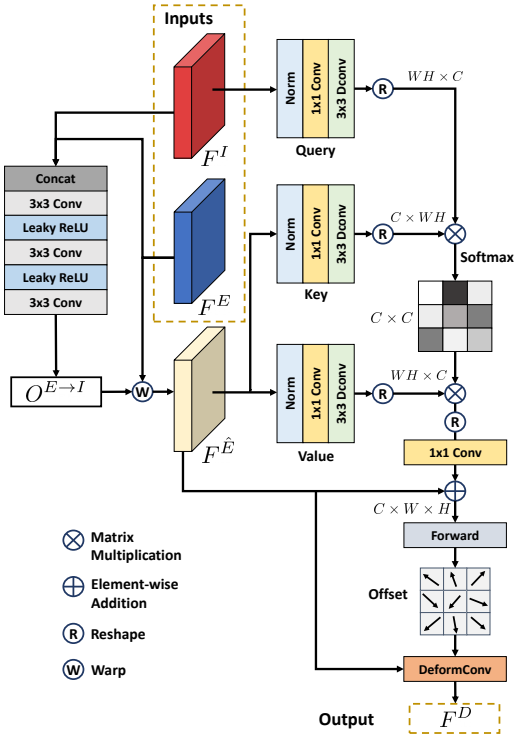

Figure 3: The proposed Attention-based Deformable Align (ADA) module.

Therefore, the correspondences match from events to images is unreliable due to the significant modality difference between events and images.

To solve these issues, we propose the Attention-based Deformable Align (ADA) module that estimates the coarse optical flow only for initial alignment. Then, the ADA module aligns the features precisely through transformer [53, 60] structures and a deformable convolution [12, 64]. The transformer can obtain matching pixels for each pixel through similarity, even between two different modalities. Also, the deformable convolution can handle the various sizes of the pixel-displacement for each pixel. As shown in Fig. 3, we design a shallow structure of motion estimator consisting of a few convolutional and ReLU activation layers. Given image features F^I and event features F^E, the motion estimator generates optical flows $O^{E\to I}$ from events to images. Then, we generate the warped features $F^{\hat{E}}$ by backward warping operation with optical flows $O^{E\to I}$.

After coarse alignment, we perform fine alignment in the local region through deformable convolution. The challenge for such fine alignment is the ambiguity of matching because image features are blurry, not sharp. Therefore, we adopt the more discriminative feature representations, *i.e*. Transformer [53]. To calculate the cross-attention, we generate query $\mathcal{Q}\in\mathbb{R}^{WH\times C}$ from the image feature F^I, and key $\mathcal{K}\in\mathbb{R}^{C\times WH}$ and value $\mathcal{V}\in\mathbb{R}^{WH\times C}$ from the

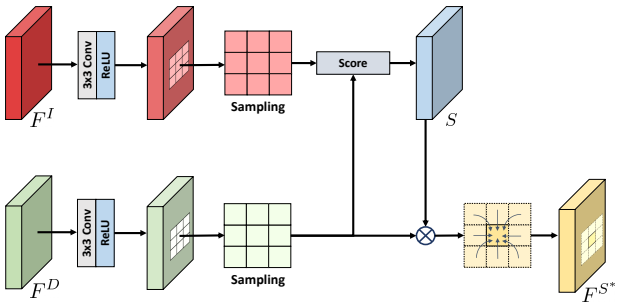

Figure 4: The proposed Local Score-based Aggregation (LSA) module.

warped event feature $F^{\hat{E}}$. The $\mathcal{Q},\mathcal{K},\mathcal{V}$ are projected by efficient depth-wise separate convolutional (Dconv) layer [23], convolution layer and normalization layer. Then, ADA module computes the attention map $\mathcal{A}$, which contains correlation between images and warped events as:

$$\mathcal{A}=\text{softmax}(\mathcal{K}\mathcal{Q}/\alpha)\in\mathbb{R}^{C\times C}, \quad (1)$$

where α is a learnable parameter that adjusts the scale of the attention map. The attention map $\mathcal{A}$ represents the similarity between two features. Then, the offset $\Delta\mathcal{P}$ and modulation weights Δm_k are computed by multiplying the warped event feature with the attention map and going through the feed-forward function [68]. We predefine the deformable convolutional kernels with K^2 sampling points ($K=3$ in our experiment). The spatially aligned event feature F^D at specific position $\mathcal{P}$ can be obtained as:

$$F^D(\mathcal{P})=\sum_{k=1}^{K^2} w_k\cdot F^{\hat{E}}\left(\mathcal{P}+\mathcal{P}_k+\Delta\mathcal{P}_k\right)\cdot\Delta m_k, \quad (2)$$

where w_k is the weight for k-th point, $\mathcal{P}_k$ is the fixed offset to $\mathcal{P}$ in the kernels. Using correlation through an attention mechanism, we can precisely align event features with image features. In general, the previous transformer works [33, 14] correlate the queries and keys spatially so that the computation cost quadratically increases as input resolution increases, *i.e*. $O(W^2H^2)$. To be computationally efficient even for high-resolution, we compute attention channel-wise so that complexity is linear, *i.e*. $O(C^2)$. Nevertheless, our experiments show that it can sufficiently perform spatial alignment.

3.3. Local Score-based Aggregation (LSA)

Event cameras have an inherently natural noise characteristic and have different response functions from the frame-based camera because they are distinct devices. Therefore, the location of the triggered event may not match the area where the blurring occurs in the image. To obtain locations where events exist and blur effects are likely to occur and, simultaneously, to remove unnecessary event regions, we design the Local Score-based Aggregation (LSA)

module. First, as shown in Fig. 4, we feed the image feature F^I and spatially aligned event feature F^D into a 3×3 convolution layer with a ReLU activation layer, and obtain the features of the same size as $\mathbb{R}^{\hat{C} \times \hat{H} \times \hat{W}}$. Next, for each feature, we sample the surrounding pixels by a specified window for each pixel. Let $F^{\hat{I}}, F^{\hat{D}} \in \mathbb{R}^{\sigma \times \hat{C} \times \hat{H} \times \hat{W}}$ denote the features sampled by the number of σ around each spatial location of the image features and spatially aligned event features, respectively. Then, we generate the matching score S by calculating the similarity between the $F^{\hat{I}}$ and $F^{\hat{D}}$ as $S = \langle F^{\hat{I}}, F^{\hat{D}} \rangle \in \mathbb{R}^{\sigma \times \hat{H} \times \hat{W}}$. Here, $\langle \cdot, \cdot \rangle$ is the inner product. The score S contains the similarity score between the sharp representations of the two modalities, thus alleviating unnecessary events for deblurring. Using the score and aligned feature of each surrounding pixel, we can obtain an aggregated feature as:

$$F^{S^*}(\mathbf{p}) = \sum_{i=1}^{\sigma} F^{\hat{D}}(i, \mathbf{p}) \times \text{sigmoid}(S(i, \mathbf{p})). \quad (3)$$

Finally, the output of LSA module is defined by adding the aggregated feature F^{S^*} to the spatially aligned event feature F^D as $F^S = F^{S^*} + F^D$. By considering neighbors together, a more reliable aggregate is possible, and ambiguity can be resolved even if there are repeated patterns.

3.4. Cross-Channel Interaction

Through the ADA and LSA modules, event features are spatially aligned well with the blurry image. To further exploit the rich context in the features, we focus on the relationships between channels. As explored in the previous studies [17, 45, 15, 18], channel-wise attention of features leads to a performance gain of the model. Therefore, we design the Cross-Channel Interaction (CCI) module for context enhancement by adopting the channel-attention mechanism [17]. The Cross-Channel Interaction (CCI) module takes the image feature F^I and the output of the LSA module F^S. To well exploit the spatial information for attention, we feed the features F^I and F^S into the convolutional and activation layers and extract the encoded features $F(I) \in \mathbb{R}^{C \times W \times H}$ and $F(S) \in \mathbb{R}^{C \times W \times H}$, respectively. To compute the relationship between two modalities, we add the two features to get correlated feature $F(A)$. We compute the temporal correlation matrix $\mathbf{A} \in \mathbb{R}^{C \times C}$ by applying a softmax function and multiplication to the reshaped feature $F(A) \in \mathbb{R}^{C \times (W \times H)}$ as follows:

$$\mathbf{A}_{uv} = \frac{\exp((F(A)F(A)^\top)_{uv})}{\sum_{c=1}^{C} \exp((F(A)F(A)^\top)_{uc})}, \quad \sum_{v=1}^{C} \mathbf{A}_{uv} = 1. \quad (4)$$

The feature F^C, which is the output of the CCI module, is obtained through multiplication with encoded feature $F(S)$ and residual addition with the output of the LSA module F^S as $F^C = \mathbf{A}F(S) + F^S$.

Table 1: Comparison of our NCER dataset with publicly available High Quality Frames (HQF) dataset [52], recorded by DAVIS240C.

	NCER (Ours)	HQF [52]
No. sequences	**83**	14
No. frames	**80.6 k**	15.4 k
Event Camera	**640 × 480**	240 × 180
Frame Camera	**640 × 480**	240 × 180
Max Frame Rate	**226 FPS**	40 FPS
Color	✔	✗
Pixel Aligned	✗	✔
Dynamic Range [dB]	**71.43**	55

3.5. Loss Functions

We use the learned perceptual similarity (LPIPS) loss ($\mathcal{L}_{LPIPS}$) [73] for better visual quality and L_1 loss, which are formulated as

$$\mathcal{L}_{restore} = \left\| I^S - I^R \right\|_1 + \lambda_0 \mathcal{L}_{LPIPS}(I^S, I^R), \quad (5)$$

where I^S is ground-truth sharp frame and I^R is the restored sharp frame.

4. Non-Coaxial Events and RGB dataset

Most event datasets for deblurring usually consist of pre-aligned APS images with grayscale. However, since these images have low frame rates and are noisy, it makes sense to anticipate a high-quality RGB image instead, which is the case in [57, 58]. However, additional equipment is used to align accurately, but it is not easy for all users to make such a bulky and elaborate system [57, 47], and calibration and homography calculations are required in the meantime. To demonstrate that our framework can perform feature-level alignment without additional equipment, we propose an Non-Coaxial Events and RGB (NCER) dataset recorded with a high-frame-rate RGB camera in combination with a high-resolution event camera. Each camera is separated from the other, so they are not pixel-wise aligned. We use a 1440×1080 FLIR BlackFly S RGB camera and a 640×480 DVXplorer event camera with approximately the same field of view to generate our dataset. Then, we crop the RGB image so that the resolution of the RGB camera becomes the same as that of the event camera with 640×480. We only use 640×440 resolution for training and evaluation, discarding the last 40 rows that the RGB camera does not capture. We split the NCER dataset into two subsets, namely NCER-F (Far) and NCER-E (Extreme).

NCER-F consists of a total of 39.7k images of 43 scenes. We average varying numbers from 11 to 31 of successive sharp frames to generate blurs of different strengths. Then, we define the sharp ground truth image corresponding to each blurry image as the intermediate image of the sharp sequences used to generate the blurry image. Finally, our NCER-F dataset provides 4,037 pairs of blurry and sharp

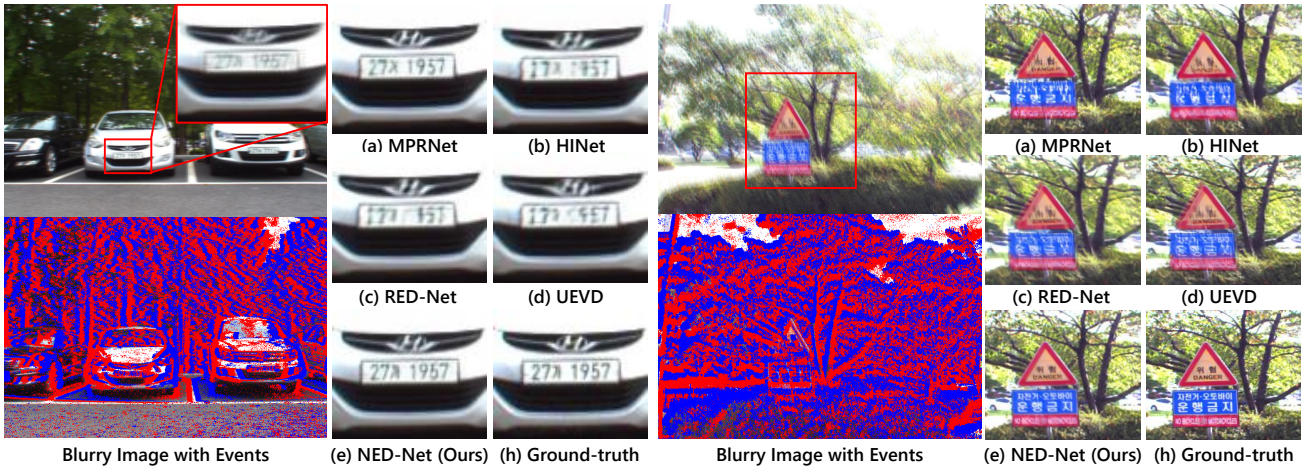

Figure 5: Qualitative results for the proposed method with other methods on NCER-F dataset.

images and corresponding event streams. The dataset is split into 2,583 and 1,454 for training and evaluation, respectively. As shown in Table 1, compared to the existing dataset [52] composed of APS images, the NCER dataset has more scenes and RGB images with higher resolution. In addition, we can generate more realistic motion blur frames due to the use of a high-frame-rate camera while capturing.

NCER-E. The NCER-F focuses on specific situations where the data from the two separate cameras are partially aligned. Therefore, the NCER-F dataset has a minimum distance of 2m between the nearest object and a camera to ensure a certain degree of alignment. On the other hand, the NCER-E dataset targets more general but challenging environments. We establish the NCER-E dataset under the challengeing conditions of three folds: (1) The intensity of the blur can vary from object to object within a scene; (2) As a homography should not correct misalignment, misalignment varies within the scene; (3) There must be a very close object (distance $\ll$ 1m), so the degree of misalignment must be enormous. Therefore, the NCER-E dataset mainly contains significantly close and distant objects in a single scene simultaneously. So, the extent of misalignment varies for each object in the scene. We average the sharp images in the NCER-E dataset with various numbers to generate blurry images, and the train and test set is split into 2,572 and 1,448 images, respectively. More details about the proposed NCER dataset are described in the *supple*.

5. Experiments

5.1. Implementation Details

Following previous studies, we represent an event stream with a voxel grid [78] and set bins of the voxel grid as 24. We set the σ of Sec. 3.3 as 9. The proposed networks are trained by Adam [32] optimizer. The initial learning rate is set to 1×10^{-4} and decreases by the factor of 0.5 at every 100 epochs. All networks are trained up to 300 epochs. We train networks on 192×128 patches with a batch size of 8. We set λ_1 as 0.3 before 50 epochs, then decrease λ_1 to 0.1 after 50 epochs. On the other hand, λ_0 is fixed at 0.01.

Table 2: Comparison of PSNR (dB) and SSIM from other methods and ours without pre-alignment. The best and second-best scores are **highlighted** and underlined.

Method	NCER-F PSNR↑	NCER-F SSIM↑	NCER-E PSNR↑	NCER-E SSIM↑	Parm↓ (M)	Time↓ (ms)
Frame						
MIMO [11]	25.30	0.7810	29.56	0.8443	16.1	133
HINet [6]	27.23	0.8148	30.36	0.8660	88.7	144
MPRNet [69]	27.98	0.8317	31.13	0.8708	20.1	488
Frame + Event						
EVDI [74]	23.58	0.7159	27.88	0.8199	3.9	**40**
LEDVDI [31]	24.35	0.7445	28.11	0.8212	5.5	194
eSL-Net [63]	23.59	0.7112	27.65	0.8176	**1.8**	280
EFNet [55]	27.97	0.8401	30.76	0.8708	8.5	81
RED-Net [66]	27.95	0.8375	32.20	0.8922	9.8	100
UEVD [28]	27.52	0.8246	32.10	0.8913	27.9	197
NED-Net	**28.78**	**0.8476**	**33.20**	**0.9059**	14.5	148

5.2. Experiments Results

We compare our method with state-of-the-art image-only and event-based methods on the NCER dataset. We train each model with the proposed NCER dataset using the official code provided by the authors. We present the result without pre-alignment, as shown in Table 2.

Experiments w/o Pre-alignment. The result of the best frame-based method [69] overcomes the most existing event-based methods. Even though there is no significant misalignment between events and images in the NCER-F dataset, most methods cannot deal with it. Instead, event-based methods benefit at runtime by using events, *i.e.*, EFNet has 0.01dB lower PSNR than MPRNet, but inference time is 6 times faster. Our NED-Net is designed to correct the misalignment between modalities within the network,

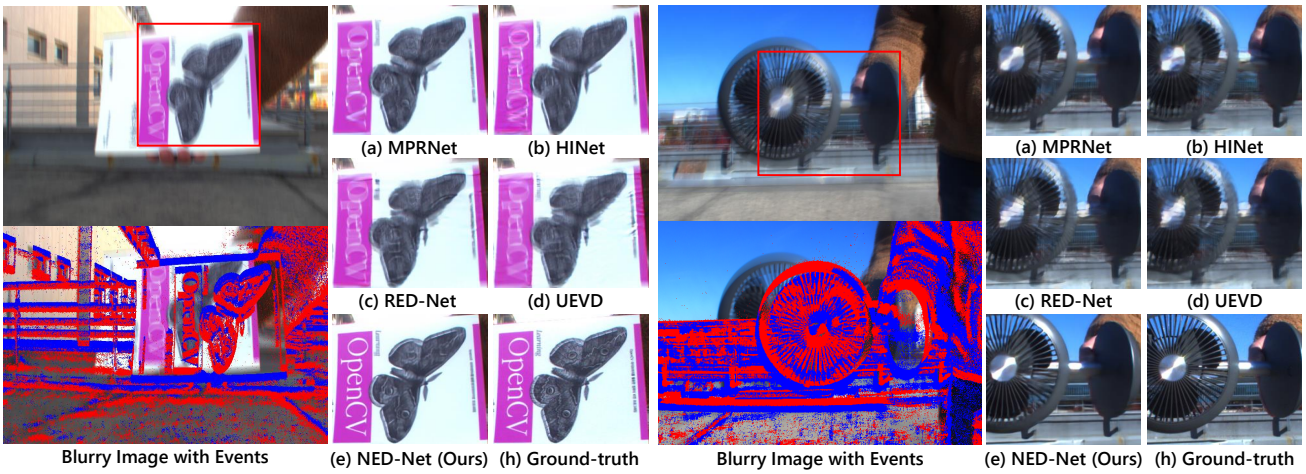

Figure 6: Qualitative results for the proposed method with other methods on NCER-E dataset.

Table 3: Comparison of PSNR (dB) and SSIM from other methods and ours with pre-alignment. The best and second-best scores are **highlighted** and underlined.

Method	Align	NCER-F PSNR↑	NCER-F SSIM↑	NCER-E PSNR↑	NCER-E SSIM↑
Frame					
MIMO-UNet [11]		25.30	0.7810	29.56	0.8443
HINet [6]		27.23	0.8148	30.36	0.8660
MPRNet [69]		27.98	0.8317	31.13	0.8708
Frame + Event					
EVDI [74]	✔	23.64	0.7156	27.62	0.8182
LEDVDI [31]	✔	24.52	0.7481	28.20	0.8245
eSL-Net [63]	✔	23.84	0.7206	27.53	0.8154
EFNet [55]	✔	27.43	0.8249	30.86	0.8779
RED-Net [66]	✔	28.04	0.8350	32.25	0.8904
UEVD [28]	✔	27.98	0.8261	31.53	0.8806
NED-Net		**28.78**	**0.8476**	**33.20**	**0.9059**

Table 4: Ablation of the proposed modules.

Ver.	ADA	LSA	CCI	NCER-F	NCER-E
1				27.73	31.97
2		✔		27.81	32.11
3			✔	27.83	32.17
4		✔	✔	27.98	32.24
5	✔			28.33	32.74
6	✔		✔	28.57	33.04
7	✔	✔		28.64	33.11
8	✔	✔	✔	**28.78**	**33.20**

Figure 7: Qualitative results of the ablation study.

thus achieving the best performance. In addition, NED-Net has better performance than MPRNet and is as fast as 3.3 times, so it can be a step forward to solving the dilemma of the tradeoff between performance and speed. As shown in Fig. 5, our NED-Net robustly restores sharper details than image-only and event-based methods. As shown in Fig. 6, the NCER-E dataset comprises objects at various distances, which are more frequent in the real-world than in the NCER-F dataset. Most of the event-based methods struggle in the NCER-E dataset. The reason is that the NCER-E dataset has a lot of dynamic motion, and the pixel difference between event and image is diverse. Therefore, it is challenging to find the per-pixel alignment between the two modalities with a simple convolution kernel.

Experiments w/ Pre-alignment. For a fair comparison, we conduct additional experiments with pre-alignment, as shown in Table 3. Following [57, 58], we use extrinsic- and feature-based homography to align events to images for the NCER-F and NCER-E, respectively. In NCER-F, where many objects are situated at a distance, a certain degree of alignment is established. For example, UEVD obtains a PSNR of 27.98 dB, which is comparable in some manner to our PSNR score of 28.78 dB. However, in the NCER-E dataset where various distances exist, the effect of pre-alignment is not significant, *i.e.*, performance difference between UEVD and NED-Net widens to 1.67 dB. In the NCER-E, aligning a nearby object results in the misalignment of a distant object and vice versa. Therefore, it is clear that global pre-alignment cannot be a general solution for extensive misalignments and object distance variations. This implies that our feature-level alignment is more effective than explicit approaches. In addition, as shown in Fig. 6, even in the prominent misalignment of the two modalities, our method shows better quality than the existing methods by using event information efficiently. More qualitative results, such as real blurry images, etc., are provided in the *supple*.

Table 5: The results of the ADA module regarding PSNR, SSIM, runtime, model parameters, and GMACs. The unit of time and model parameters are ms and M, respectively.

Method	PSNR	SSIM	Time	Parm.	GMACs
w/o ADA	27.98	0.8316	**130**	**13.4**	**1056**
w/ ADA (Ours)	**28.78**	**0.8476**	148	14.5	1124

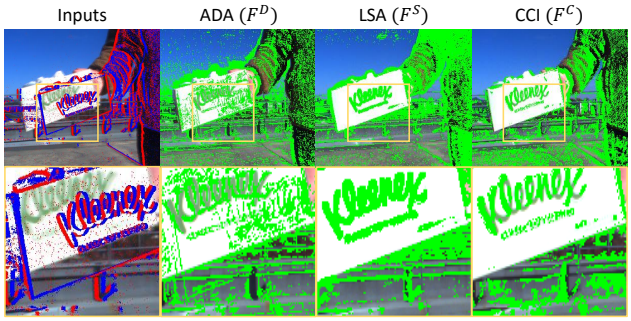

Figure 8: The visualization of the feature maps extracted from each module. Zoom in for a better view.

5.3. Ablation Studies and Discussion

We conduct in-depth examinations of our NED-Net on the NEID task using the NCER dataset.

1. Contribution of each component for the performance. To investigate the contribution of each module, we report the ablation results in Table 4. We set the baseline network by removing all proposed modules, which are flexible structures. Starting from the baseline network, we proceed with the evaluation by adding the proposed components one by one. The noteworthy point is the significant performance improvement when using the ADA module, which performs spatial alignment between two modalities. Even using the ADA module standalone, the PSNR is improved by 0.6 dB and 0.77 dB on the NCER-F and NCER-E, respectively (Ver.1 and 5). On the other hand, if LSA and CCI modules are used without the ADA module, the performance improvement is relatively insignificant because spatially unaligned events are used (Ver.2∼4). Instead, if the ADA module is used together, performance improvement is significant. For example, with the standalone use of the LSA module on NCER-F, improvement rises by only 0.08 dB (Ver.1 and 2), but when used with the ADA, it increases by 0.31 dB (Ver.5 and 7). Finally, the best performance is achieved when all proposed modules are used together. To further validate the proposed modules, we present the qualitative ablation study in Fig. 7.

2. Efficient design of the ADA. We validate how the proposed ADA module, which leads to the most significant performance boost, is designed highly efficiently. We report the results for various metrics in Table 5. The w/o ADA represents a network where only the ADA module is removed from full network (see Ver.4 in Table 4). We confirm that introducing ADA significantly affects deblurring performance (27.98 dB vs. 28.78 dB) compared with the w/o ADA network. Nevertheless, the ADA module is lightweight, which increases the minimal additional time (130 ms vs. 148 ms) and computation cost (1054 GMACs vs. 1124 GMACs) very little. We demonstrate that the ADA module efficiently solves the alignment problem using a few parameters, suggesting our framework can be generally used on various mobile devices.

Table 6: Performance according to voxel grid bin size.

Bin	8	12	16	24	32
PSNR	28.49	28.67	28.75	28.78	**28.80**

Table 7: Results about generalization ability. Networks are trained in NCER-F and tested in NCER-E

	Frame		Frame + Event			
Method	[6]	[69]	[55]	[66]	[28]	NED-Net
PSNR↑	30.36	<u>31.13</u>	28.56	30.48	30.10	**32.60**
SSIM↑	0.8660	<u>0.8708</u>	0.8337	0.8701	0.8621	**0.9004**

Table 8: Comparison of PSNR (dB), runtime (ms), model parameters (M), and GMACs with only flow-based [54] and patch-based attention [53] methods.

Method	NCER-F	NCER-E	Time	Parm.	GMACs
[53]	28.12	31.87	273	17.8	1827
[54]	27.96	32.38	436	23.9	1413
ADA	**28.78**	**33.20**	**148**	**14.5**	**1124**

3. How does each module work for the NEID task? To explain explicitly, we visualize the output features generated from each module in Fig. 8. Looking at F^D, which is the result of the ADA module that aligns globally, the position of event features are mostly moved to the edges of objects in the images, but the mismatched events still remain. After that, the LSA module locally aggregates a few mismatched events in F^s. Finally, the output feature of the CCI module F^C has sharp textures aligned with the blurry regions through cross-channel attention with the image feature. From the results, we demonstrate that our proposed modules are suitable for solving NEID task.

4. Comparison with direct flow- and patch-based methods. We provide the comparisons of the proposed ADA module with existing direct flow [54] and patch-based attention methods [53]. As shown in Table 8, [54] demonstrates a certain level of performance in NCER-F despite lacking an explicit alignment process, but it falls short in NCER-E. On the other hand, [53], which employs a larger model for obtaining optical flow, struggles with precise alignment in NCER-F. Furthermore, [53] imposes a heavy computational burden of $O(C^2WH)$ due to spatial operations, which have significantly higher complexity than ours, $O(C^2)$.

5. Voxel Grid Size. Table 6 reports the performance according to the voxel size on NCER-F dataset. Small sizes perform worse, but it is not significant. The bin size of 16

is the same as in UEVD, but ours still performs better.

6. Generalization Ability. Another issue that can be considered in the NEID task is whether the network can operate for different distributions of misalignment. To validate the generalization ability of our approach, we train networks using the NCER-F dataset and assess their performances on the NCER-E dataset under various misalignment scenarios. As shown in Table 7, event-based networks that do not consider misalignment often struggle to generalize to datasets with different distributions of misalignment. As an example, Table 3 demonstrates that UEVD [28] produces the PSNR that is 0.8 dB lower than our approach in the NCER-F dataset. However, in the case of the NCER-E, the PSNR value is 2.5 dB lower when compared to our approach. In contrast, our method shows to be robust to variations in misalignment and yields the best performance by a significant margin, even trained with little misalignment.

6. Conclusion

In this paper, we first tackle the Non-coaxial Event-guided Image Deblurring (NEID) task with a practical focus on a camera setup consisting of an RGB camera and a non-coaxial event camera. To this end, we propose a novel framework suitable for this task, called NED-Net. For training and evaluation, we propose the first Non-Coaxial Event and RGB (NCER) dataset composed of real-world RGB images with pixel-wise non-aligned events at high-resolution. The experiments on two real-world event datasets demonstrate the effectiveness of our method. Furthermore, our NED-Net achieves high deblurring performance even in the diverse environment, varying mismatches with unaligned events. Although we set the target task as deblurring in this study, our framework and dataset can be flexibly extended to other tasks, such as super-resolution or video frame interpolation that can use temporally dense events efficiently.

Acknowledgements. This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (NRF-2022R1A2B5B03002636).

References

[1] Yuval Bahat, Netalee Efrat, and Michal Irani. Non-uniform blind deblurring by reblurring. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3286–3294, 2017. 2

[2] M Bigas, Enric Cabruja, Josep Forest, and Joaquim Salvi. Review of cmos image sensors. *Microelectronics journal*, 37(5):433–451, 2006. 2

[3] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240× 180 130 db 3 μs latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10):2333–2341, 2014. 2

[4] Marco Cannici, Chiara Plizzari, Mirco Planamente, Marco Ciccone, Andrea Bottino, Barbara Caputo, and Matteo Matteucci. N-rod: A neuromorphic dataset for synthetic-to-real domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1342–1347, 2021. 2

[5] Ayan Chakrabarti. A neural approach to blind motion deblurring. In *European conference on computer vision*, pages 221–235. Springer, 2016. 2

[6] 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. 1, 2, 6, 7, 8

[7] Hoonhee Cho, Jegyeong Cho, and Kuk-Jin Yoon. Learning adaptive dense event stereo from the image domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17797–17807, 2023. 3

[8] Hoonhee Cho, Jaeseok Jeong, and Kuk-Jin Yoon. Eomvs: Event-based omnidirectional multi-view stereo. *IEEE Robotics and Automation Letters*, 6(4):6709–6716, 2021. 3

[9] Hoonhee Cho and Kuk-Jin Yoon. Event-image fusion stereo using cross-modality feature propagation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 454–462, 2022. 3

[10] Hoonhee Cho and Kuk-Jin Yoon. Selection and cross similarity for event-image deep stereo. In *European Conference on Computer Vision*, pages 470–486. Springer, 2022. 3

[11] 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. *arXiv preprint arXiv:2108.05054*, 2021. 6, 7

[12] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 4

[13] Jiangxin Dong, Jinshan Pan, Zhixun Su, and Ming-Hsuan Yang. Blind image deblurring with outlier handling. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2478–2486, 2017. 2

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[15] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019. 5

[16] 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. 2

[17] Yu Gao, Xintong Han, Xun Wang, Weilin Huang, and Matthew Scott. Channel interaction networks for fine-grained image categorization. In *Proceedings of the AAAI*

conference on artificial intelligence, volume 34, pages 10818–10825, 2020. 5

[18] Zilin Gao, Jiangtao Xie, Qilong Wang, and Peihua Li. Global second-order pooling convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2019. 5

[19] Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to events: Recycling video datasets for event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3586–3595, 2020. 2

[20] Mathias Gehrig and Davide Scaramuzza. Recurrent vision transformers for object detection with event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13884–13893, 2023. 3

[21] Chen Haoyu, Teng Minggui, Shi Boxin, Wang YIzhou, and Huang Tiejun. Learning to deblur and generate high frame rate video with an event camera. *arXiv preprint arXiv:2003.00847*, 2020. 2, 3

[22] Michael Hirsch, Christian J Schuler, Stefan Harmeling, and Bernhard Schölkopf. Fast removal of non-uniform camera shake. In *2011 International Conference on Computer Vision*, pages 463–470. IEEE, 2011. 2

[23] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 4

[24] 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. 2

[25] Zhe Jiang, Yu Zhang, Dongqing Zou, Jimmy Ren, Jiancheng Lv, and Yebin Liu. Learning event-based motion deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3320–3329, 2020. 2, 3

[26] Meiguang Jin, Zhe Hu, and Paolo Favaro. Learning to extract flawless slow motion from blurry videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8112–8121, 2019. 3

[27] Taewoo Kim, Yujeong Chae, Hyun-Kurl Jang, and Kuk-Jin Yoon. Event-based video frame interpolation with cross-modal asymmetric bidirectional motion fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18032–18042, 2023. 3

[28] Taewoo Kim, Jeongmin Lee, Lin Wang, and Kuk-Jin Yoon. Event-guided deblurring of unknown exposure time videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 3, 6, 7, 8, 9

[29] Youngrae Kim, Jinsu Lim, Hoonhee Cho, Minji Lee, Dongman Lee, Kuk-Jin Yoon, and Ho-Jin Choi. Efficient reference-based video super-resolution (ervsr): Single reference image is all you need. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1828–1837, 2023. 3

[30] Wei-Sheng Lai, Jia-Bin Huang, Zhe Hu, Narendra Ahuja, and Ming-Hsuan Yang. A comparative study for single image blind deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1709, 2016. 2

[31] Songnan Lin, Jiawei Zhang, Jinshan Pan, Zhe Jiang, Dongqing Zou, Yongtian Wang, Jing Chen, and Jimmy SJ Ren. Learning event-driven video deblurring and interpolation. In *ECCV (8)*, pages 695–710, 2020. 2, 3, 6, 7

[32] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*, 2019. 6

[33] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 4

[34] Simon Meister, Junhwa Hur, and Stefan Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 3

[35] Tomer Michaeli and Michal Irani. Blind deblurring using internal patch recurrence. In *European conference on computer vision*, pages 783–798. Springer, 2014. 2

[36] Manasi Muglikar, Mathias Gehrig, Daniel Gehrig, and Davide Scaramuzza. How to calibrate your event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1403–1409, 2021. 2

[37] 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. 2

[38] Seungjun Nah, Sanghyun Son, and Kyoung Mu Lee. Recurrent neural networks with intra-frame iterations for video deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8102–8111, 2019. 2

[39] 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. 2

[40] Mehdi Noroozi, Paramanand Chandramouli, and Paolo Favaro. Motion deblurring in the wild. In *GCPR*, pages 65–77. Springer, 2017. 2

[41] 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. 2

[42] Liyuan Pan, Richard Hartley, Cedric Scheerlinck, Miaomiao Liu, Xin Yu, and Yuchao Dai. High frame rate video reconstruction based on an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2, 3

[43] 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. 2, 3

[44] Mirco Planamente, Chiara Plizzari, Marco Cannici, Marco Ciccone, Francesco Strada, Andrea Bottino, Matteo Matteucci, and Barbara Caputo. Da4event: towards bridging the sim-to-real gap for event cameras using domain adaptation. *IEEE Robotics and Automation Letters*, 6(4):6616–6623, 2021. 2

[45] Pengfei Zhu Peihua Li Wangmeng Zuo Qilong Wang, Banggu Wu and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 5

[46] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. Esim: an open event camera simulator. In *Conference on Robot Learning*, pages 969–982. PMLR, 2018. 2

[47] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 184–201. Springer, 2020. 5

[48] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3

[49] 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 (ICCV)*, pages 4531–4540, October 2021. 2, 3

[50] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 3

[51] Shintaro Shiba, Yoshimitsu Aoki, and Guillermo Gallego. Secrets of event-based optical flow. In *European Conference on Computer Vision*, pages 628–645. Springer, 2022. 3

[52] Timo Stoffregen, Cedric Scheerlinck, Davide Scaramuzza, Tom Drummond, Nick Barnes, Lindsay Kleeman, and Robert Mahony. Reducing the sim-to-real gap for event cameras. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pages 534–549. Springer, 2020. 2, 5, 6

[53] Maitreya Suin, Kuldeep Purohit, and AN Rajagopalan. Spatially-attentive patch-hierarchical network for adaptive motion deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3606–3615, 2020. 4, 8

[54] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018. 3, 8

[55] 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 (ECCV)*, 2022. 2, 3, 6, 7, 8

[56] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 3

[57] Stepan Tulyakov, Alfredo Bochicchio, Daniel Gehrig, Stamatios Georgoulis, Yuanyou Li, and Davide Scaramuzza. Time lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17755–17764, 2022. 2, 3, 5, 7

[58] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. TimeLens: Event-based video frame interpolation. *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2, 5, 7

[59] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 2

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[61] Patricia Vitoria, Stamatios Georgoulis, Stepan Tulyakov, Alfredo Bochicchio, Julius Erbach, and Yuanyou Li. Event-based image deblurring with dynamic motion awareness. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 95–112. Springer, 2023. 2

[62] Bishan Wang, Jingwei He, Lei Yu, Gui-Song Xia, and Wen Yang. Event enhanced high-quality image recovery. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 155–171. Springer, 2020. 2, 3

[63] Bishan Wang, Jingwei He, Lei Yu, Gui-Song Xia, and Wen Yang. Event enhanced high-quality image recovery. In *European Conference on Computer Vision*, pages 155–171. Springer, 2020. 6, 7

[64] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 4

[65] Patrick Wieschollek, Michael Hirsch, Bernhard Scholkopf, and Hendrik Lensch. Learning blind motion deblurring. In *ICCV*, pages 231–240, 2017. 2

[66] Fang Xu, Lei Yu, Bishan Wang, Wen Yang, Gui-Song Xia, Xu Jia, Zhendong Qiao, and Jianzhuang Liu. Motion deblurring with real events. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2583–2592, October 2021. 2, 3, 6, 7, 8

[67] Yanyang Yan, Wenqi Ren, Yuanfang Guo, Rui Wang, and Xiaochun Cao. Image deblurring via extreme channels prior.

In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4003–4011, 2017. 2

[68] 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. 4

[69] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. *arXiv preprint arXiv:2102.02808*, 2021. 2, 6, 7, 8

[70] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Deep stacked hierarchical multi-patch network for image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5978–5986, 2019. 2

[71] Haichao Zhang and David Wipf. Non-uniform camera shake removal using a spatially-adaptive sparse penalty. In *Advances in Neural Information Processing Systems*, pages 1556–1564. Citeseer, 2013. 2

[72] Limeng Zhang, Hongguang Zhang, Chenyang Zhu, Shasha Guo, Jihua Chen, and Lei Wang. Fine-grained video deblurring with event camera. In *ICMM*, pages 352–364. Springer, 2021. 2

[73] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[74] 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. 3, 6, 7

[75] Zhihang Zhong, Ye Gao, Yinqiang Zheng, and Bo Zheng. Efficient spatio-temporal recurrent neural network for video deblurring. In *European Conference on Computer Vision*, pages 191–207. Springer, 2020. 2

[76] Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, Haozhe Xie, Jinshan Pan, and Jimmy S Ren. Davanet: Stereo deblurring with view aggregation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10996–11005, 2019. 3

[77] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 989–997, 2019. 3

[78] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based optical flow using motion compensation. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 6