Motion Aware Double Attention Network for Dynamic Scene Deblurring

Dan Yang Mehmet Yamac
Huawei Technologies Oy (Finland) Co. Ltd
{dan.yang1, mehmet.yamac}@huawei.com

Abstract

Motion deblurring in dynamic scenes is a challenging task when the blurring is caused by one or a combination of various reasons such as moving objects, camera movement, etc. Since event cameras can detect changes in intensity with a low latency, necessary motion information is inherently captured in event data, which could be quite useful for deblurring standard camera images. The degradation intensity does not show homogeneity across an image due to factors like object depth, speed, etc. We propose a two-branch network structure, Motion Aware Double Attention Network (MADANet), that pays special attention to areas with high blur. As part of the network, event data is first used by the high blur region segmentation module that creates a probability-like score for areas exhibiting high relative motion to the camera. Then, the event data is also injected to feature maps in the main body, where there is a second attention mechanism available for each branch. The effective usage of event data and two-level attention mechanisms makes the network very compact. During the experiment, it was shown that the proposed network could achieve state-of-the-art performance not only on the benchmark dataset from GoPro, but also on two newly collected datasets, one of which contains real event data.

1. Introduction

In daily taken photographs, blur is a common distortion generally caused by a variety of causes, including object motion, camera shake, depth variations in the scene, and more. Earlier approaches modeled degradation using a global blurring kernel applied to sharp images or using kernels that are applied locally. However, estimating both blur kernels and then the sharp image is a challenging ill-posed problem, which may require a computationally demanding iterative solution. Meanwhile, the most recent Convolutional Neural Network (CNN)-based approaches [7, 31, 47, 48, 58] have shown significant improvements in reconstruction accuracy as well as computation efficiency. Most of them omit the kernel estimation step and are trained directly to map blurry images to sharp images using more realistic datasets, such as GoPro dataset [31]. Despite the significant advances [7, 48] in deblurring networks for single images, their performance still falls short in challenging situations such as those involving fast-moving objects.

Figure 1. The proposed MADANet and the state-of-the-art deblurring methods, BANet [48], MIMO+ [7] and HINet [6].

Furthermore, multiple input frames have also been tried in an attempt to increase recovery performance, rather than using only single frames, such as burst frames [1, 52] or differently exposed frames [5, 29, 56, 60]. The latter approach, which combines long and short exposure frames and produces a sharper, more pleasing output, has been shown to perform better in recent studies [29, 60]. However, using a long/short exposure frames strategy may still be inadequate in some cases, for example when the relative motion between the camera and the object is fast and the motion trajectory is nonlinear. Throughout this paper, the terms ”long-exposure frame” and ”blurry image” will be used interchangeably.

The objective of the present study is to use an event-based camera as an aid for improving the quality of the blurry image captured by a standard camera. Bio-inspired sensors power event cameras [26, 41, 43], which detect changes in pixel brightness at the microsecond level. When

the intensity changes over a predefined threshold, a pixel in the event camera is triggered. As a result, spatiotemporal information of rapidly changing scenes is encoded intrinsically. This information can be extremely useful for deblurring tasks, especially for images of dynamic scenes with moving objects. However, there are only a few efforts to deblur intensity images using event frames [18, 34, 49, 53]. Though [18, 49] are deep unfolding approach, [18, 34, 49] techniques work iteratively and may be computationally demanding for portable devices like smartphones.

In comparison with traditional cameras, event cameras possess many advantages, such as high dynamic range (more than 120dB, compared to 60dB for traditional cameras), high temporal resolution, and low power consumption. However, their limited resolution (about 1 Megapixel maximum array size [43]) makes them unsuitable for practical image restoration (for example, mobile phone cameras have a resolution of over 10 Megapixels). Even though deep neural networks are being used to super resolve event streams [24], using a super-resolution (SR) algorithm for event stream before combining it with the standard camera image increases computational burden, especially for compact, portable devices. Moreover, due to their asynchronous structure, event streams cannot be directly used for standard computer vision algorithms.

In this study, an event camera aided blind deblurring network family, Motion Aware Double Attention Deblurring Networks (MADANet), is introduced. MADANet is a two branch neural network, and one branch give special attention to high blur regions caused by high relative motions. This high blur regions are localized with a event-aided segmentation module since Event cameras are capable of detecting such motions. Each branch contains another attention mechanism (e.g., blur aware (BA) modules from [48]) that further differentiates varying blurring level regions as channel attention maps. In this way, a two-level of attention mechanism (i.e., double attention mechanism) is presented. MADANet, as opposed to complicated event-based algorithms, directly feeds a series of event frames to network modules, each of which is obtained from accumulated events over a shorter period of time than the exposure time of a blurry standard camera image. In that way, instead of super resolving the event frames in order to fuse them with the blurry input image, MADANet directly injects these frames in a lower resolution feature space. The MADANet variations surpass state-of-the-art deblurring algorithms such as single-frame algorithms, short/long frame fusing methods, and previous event-aided methods with a significant margin despite its lightweight structure on benchmark GoPro dataset. The results were validated using TSlowMotion data, which was collected similarly to GoPro, but with a wider range of blurring cases and better quality ground truth images. Furthermore, the TReal dataset is collected using a real event camera coupled with a standard camera. In addition, even if the MADANet was designed as event aided deblurring solution, it still shows comparable performance to the state-of-the art one when it is used as either single image or long/short frame fusion deblurring solution. The evaluation on TReal dataset dataset also demonstrates the benefits of the proposed method visually. Contributions to the proposed design, which make it an effective tool for dynamic scene deblurring, are as follows:

- The blur degradation level does not exhibit homogeneity in a dynamic scene. In this work, for the first time in literature, we introduce an event-aided high blur region localization sub-network. Thanks to the resulting attention map, one of the network branches is able to give special attention to the high blur level regions.
- Additionally, we explore for the first time in literature how low spatial resolution event data may be used to deblur HR RGB frames without the need for SR event data.

2. Related Works

2.1. Single Image Deblurring

A severely challenging problem is finding the sharp image from the blurred one under an unknown degradation operation, a problem called blind deblurring. In the earliest approaches, blur is modeled as a spatially invariant linear system known as uniform blurring [3, 9, 22], and this fails to accommodate blur variations due to moving objects, different depth levels, etc, in dynamic scenes. Variations such as these are partly included in non-uniform [17, 51] or depth-aware [33, 54] models which use overlapping patches or different depth levels in order to estimate the pixel level varying blurring kernels. In order to have a unique solution, such severe ill-posed problems need some prior assumptions about the latent image space, which leads to iterative and therefore computationally costly solutions [55].

Recently, Convolutional Neural Network (CNN) based works have significantly improved the estimation of uniform [4, 25] and non-uniform [14, 46] blurring kernels with less computational costs and less estimation error, but still, any kernel estimation error may lead to undesired ringing artifacts. Furthermore, the kernel-based blur approximation for a blurry image in a dynamic scene does not accurately reflect real-world blur. In the more recent CNN-based deblurring solutions [31, 47, 58], the kernel estimation is bypassed and the blurry images are directly mapped to the sharp estimations. In most of these solutions, the training is carried out on benchmark datasets such as GoPro [31], REDS [30], and RealBlur [40] whose degradation of blurry images more closely mimics a real-world situation. While the recent improvements in single image deblurring networks [7, 48] are significant, their performance is still lim-

ited by challenging blurring types such as the ones caused by fast-moving objects [21].

2.2. Deblurring by Aid of Short Exposure Frame

Additionally, as a complement to single image deblurring techniques, a number of recent techniques use a burst of blurry images that are each captured sequentially, with equal exposure times, to produce a sharp image estimation [1,52]. The exposure time of an image, on the other hand, is a very important factor that determines the type of distortion in it. Images that are underexposed, or sometimes called short-exposure frames, appear sharper, but noisy, while long-exposure frames are blurred, but carry much accurate color information. Several studies [5, 29, 56, 60] have attempted to solve the sharp image recovery problem by leveraging differently exposed images; earlier studies [51,56] attempted to estimate the blur kernel of long-exposure frames from long/short exposure frames. The most recent learning-based methods [29,60] are trained to have a direct map from short/long exposure image pair to restored sharp image. The fusion strategy still fails, however, in a variety of situations, including when fast-moving objects are present, or shutter delays between capturing long and short exposure images are quite long.

2.3. Event Processing, Challenges and Event-Based Image Deblurring

A typical event camera detect the intensity change, then records pixel location, change time and polarity of change. Considering the standard camera coupled to event one, the blurry image intensity can be regarded as integral over sharp latent images changing with time, and each of these small changes between any so-called sharp images can be calculated over the sum of the event stream (in ideal case). This assumption led [34] to propose a model for deblurring intensity images using a double integral technique. A recent approach [18], models deblurring as a Maximum-a-Posteriori (MAP) problem and solves it using a neural network in a deep unfolding manner.

Event cameras obtain sparse and asynchronous information with a high temporal resolution, which makes them very different than RGB cameras. Therefore, the event data can not be directly plug in conventional computer vision algorithms. One direction is to use model based approaches as it is the case in above mentioned deblurring solutions [18, 34], or new categories of neural network designs such as bio-inspired neural networks Spiking Neural Networks [28]. On the other hand, in a few recent works event streams are converted to event frames by simple accumulation way for different computer vision tasks, including depth estimation [20], optical flow estimation [27], and feature tracking [13]. We will be developing an akin accumulation strategy to have a series of event frames for a corresponding blurry image.

One of the main drawbacks of the current event camera solutions is the large pixel sizes of the available products [11]. Their low resolution renders them unusable for practical image restoration. About 1 Megapixel resolution is the largest array size available [43]. There are some recent attempts to use deep neural networks for event stream superresolution [24]. Nevertheless, using a super-resolution for event frames before combining them with the standard camera image for image restoration adds to the computational load, especially for compact, portable devices.

3. Methodologies

Let us assume that a blurry image, $\mathbf{B}$ is the average of the corresponding latent images that can be hypothetically captured during exposure time interval, $[0, T]$, i.e., $\mathbf{B} = \frac{1}{T}\int_0^T \mathbf{I}(t)\,dt$, where $\mathbf{I}(t)$ is the corresponding sharp latent image at time t. The blurring effect occurs when objects move or handshaking occurs during the exposure time. The blurring distortion on an object's appearance depends on a combination of various factors, such as the object's distance from the camera, parameters of the camera, the speed of the object, etc. Even if only camera movement takes place, the blurring distortion amount, we call it blurring level, changes locally in the observed image, e.g., if the object is close to the camera, it will be much higher [54]. Exposure time influences the overall blurring level most significantly. As exposure time decreases, the blurring level will be reduced at the expense of true colors and signal-to-noise ratio. This is why there are some attempts to use multiple exposure frames, or a long/exposure frame pair to have a better estimation of the sharp image. All these strategies will, however, be constrained by the trade-off between exposure time to gather sufficient information and the speed at which the scene changes for the standard camera.

Unlike traditional cameras, new emerging event cameras do not have a predetermined exposure time. The bio-inspired sensors detect changes in brightness in a scene asynchronously at a microsecond level. The event camera does not capture the intensity at pixel (x, y) at time t; instead, it outputs the sequence of events, each denoted by (x, y, t, δ). In this event representation, t is the time of event occur and the polarity, δ, is determined as follows,

$$\delta = \begin{cases} +1, & g \geq thr \\ -1 & g \leq -thr \\ 0 & \text{else} \end{cases}$$

where $g = \log\left(\frac{I_{x,y}(t)}{I_{x,y}(t_{ref})}\right)$, t_{ref} is the timestamp of previous event for same pixel, and thr is the pre-determined threshold. We first subdivide the event stream during the standard exposure time, T, into a number of chunks. Each

Figure 2. The blurry image and the corresponding 6 event frames

chunk is accumulated, then it is quantized to a ternary 2-D data. Mathematically speaking, each event frame, $\mathbf{e^i}$ is obtained as follows,

$$e^i_{x,y} = Q\left(\int_{\frac{T(i-1)}{n}}^{\frac{Ti}{n}} \varepsilon_{x,y}(t)\,dt\right), \quad \text{for } i = 1, ..., n, \quad (1)$$

where $Q(h) = sgn(h) : \mathbb{R} \rightarrow \{+1, 0, 1\}$ and $\varepsilon_{x,y}(t)$ is individual event occurs at time t, and pixel location (x, y). Different from akin event representation schemes [13, 20, 27], our event frames have ternary data, $e^i_{x,y} \in \{1, 0, -1\}$. In this work, we stack $n = 6$ number of event frames for each blurry image. The corresponding event stack, $\mathbf{E} = \{\mathbf{e^1}, ..., \mathbf{e^n}\}$, will then be used in both high blur region localization module and main body of the deblurring network. An example blurry image and corresponding event frames are illustrated in Figure 2.

3.1. High Blur Region Localization Module

In photographs, fast-moving objects and those in close proximity to the camera may appear more blurry than those far away from the camera. It might be more efficient to devote more resources to areas with more blur distortion. We propose segmenting high blur level regions using a light High Blur Region Segmentation (HBRS) sub-network to locate pixels in the image plane that correspond to the points in 3D space having fast relative motion to the camera. The existing event cameras are lower resolution than the standard cameras. In this study, the resolution ratio between RGB images and event frames is assumed to be 4:1. Therefore, as inputs, the proposed HBRS sub-network takes the event stack, $\mathbf{E}$, and down-scaled blurry image, $\mathbf{B}\downarrow$, where $\downarrow$ is the standard bicubic down-sampling operation with scale factor 2. Then the network outputs an attention map, $\mathbf{A}$, that gives the information about the probability of each pixel being in high blur region, i.e., $\mathbf{A} \leftarrow \mathcal{F}_{HBRS}(\mathbf{B}\downarrow, \mathbf{E})$ such that $A_{i,j} \in [0, 1]$.

As illustrated in Figure 3, the proposed module is a lightweight encoder-decoder network. The network consists of 3 main components, encoder, residual groups, and decoder. In the encoder part, we use two convolutional layers with 32 hidden neurons. The filter size is 3 × 3 for both layers and the second one has stride 2. The decoder part has three convolutional layers; the first one consists of transposed convolutions with stride 2x2 and kernel size 4x4, while the second and third are convolutional layers with

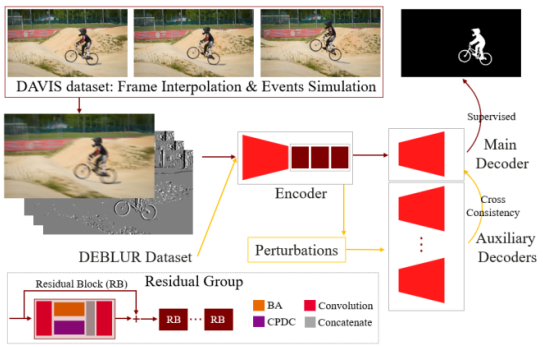

Figure 3. High Blur Region Segmentation (HBRS) network

kernel size 3x3 and 1x1, respectively. Both hidden layers have 24 neurons. ReLU is used as the activation function at end of each convolutional layer except the last output layer where Sigmoid is used. The residual group contains 3 successive residual blocks (RBs), each with 32 channels input and output feature maps. Blur Aware Module from [48] is borrowed as the residual block in this study. Each RB consists of a Blur-aware Attention (BA) block and a Cascaded Paralled Dilated Convolution (CPDC) with multiple dilation rates. The residual group is illustrated in Figure 3, and the details about the architecture of BA, and CPDC blocks can be seen in [48].

There is no labeled dataset for this task in the literature. In order to overcome this limitation, a semi-supervised method Cross-Consistency Training (CCT) that was proposed in [32] is employed. As labeled data, the video segmentation dataset, DAVIS [37], is used to generate blurry image and moving object mask. The idea of CCT is to push the encoder and residual blocks of the network to be consistent under different types of small perturbations. In order to achieve this, in each epoch, the same number of labeled data and unlabeled data is used during the training. The labeled data is used to update the encoder, residual blocks and the main decoder, while the unlabeled data is for updating the network except the main decoder. During this later update process, the feature maps from residual blocks are perturbated as suggested in [32] before fed into each auxiliary decoder. The outputs of the auxiliary decoders are forced to be close to the output of the main decoder. The overall pipeline of applied CCT is shown in Figure 3.

3.2. MADANet: Double-branch Mechanism

The proposed MADANet consists of two sub-networks; one is HBSR, $\mathcal{F}_{HBRS}(.)$, and the other one is MADANet deblurring network, $\mathcal{M}$. An encoder, $\mathcal{M}_{\mathcal{E}}$, a feature processing part (or main body), $\mathcal{M}_{\mathcal{F}}$, and a decoder, $\mathcal{M}_{\mathcal{D}}$, compose the deblurring network. Let $\mathbf{B} \in \mathbb{R}^{S_1 \times S_2 \times 3}$ be our blurry RGB image. The encoder part maps the blurry

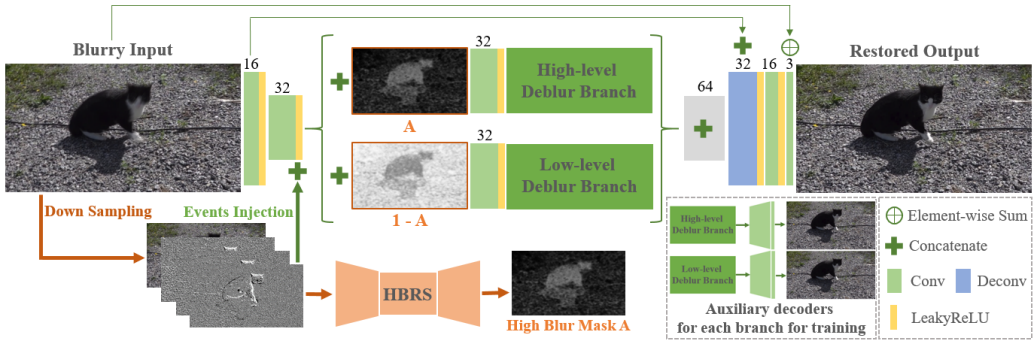

Figure 4. The pipeline of the proposed MADANet: 1) Encoding the blurry standard image to a low-resolution multi-channel feature space. 2) Injecting a series of event frames onto this feature space. 3) Dividing the network into two branches. 4) Injecting the high blur region segmentation mask, into the high blur region deblurring branch while injecting a complement of the mask into the other branch of the network. 5) Integrating the feature maps from these two branches with a decoder to produce the final image.

image to a new representation space, i.e.,

$$\mathbf{F} \leftarrow \mathcal{M}_{\mathcal{E}}(\mathbf{B}) \tag{2}$$

where $\mathbf{F} \in \mathbb{R}^{s_1 \times s_2 \times c}$ is c channel representation of size $s_1 \times s_2$. In this work, we set $c = 32$, $s_1 = S_1/2$ and $s_2 = S_2/2$.

The feature processor is composed of 2 branches, $\mathcal{M}_{\mathcal{H}}$ and $\mathcal{M}_{\mathcal{L}}$ for high blur and low blur regions, respectively. The high blur region feature processor, takes c-channel feature map, $\mathbf{F}$, then it fuses them with n number of event frames as well as with attention map, $\mathbf{A}$. Hereafter, using this $c + n + 1$ channel inputs, it outputs a an c-channel processed feature map, $\mathbf{F_H} \in \mathbb{R}^{s_1 \times s_2 \times c}$, as output, i.e.,

$$\mathbf{F_H} \leftarrow \mathcal{M}_{\mathcal{H}}(\mathbf{F}, \mathbf{A}, \mathbf{E}). \tag{3}$$

In the module $\mathcal{M}_{\mathcal{H}}$, the first layer is a convolution layer with c-neurons, e.g., in our setup it reduces the number of channels from 39 to 32. Then, multiple residual blocks follow this layer in order to produce $\mathbf{F_H}$. Low blur region feature processor branch has a similar architecture with $\mathcal{M}_{\mathcal{H}}$, except it takes the complement of the attention map, $\mathbf{1} - \mathbf{A}$ as input, i.e.,

$$\mathbf{F_L} \leftarrow \mathcal{M}_{\mathcal{L}}(\mathbf{F}, \mathbf{1} - \mathbf{A}, \mathbf{E}), \tag{4}$$

where $\mathbf{F_L} \in \mathbb{R}^{s_1 \times s_2 \times c}$ is the processed feature map. The residual blocks are the same with the ones described in Section 3.1. The number of residual blocks for high and low deblurring branches are empirically set to 6 and 5, respectively. Having feature maps from both branches, decoder fuse them by simple concatenation and produce the final output, i.e.,

$$\mathbf{I_r} \leftarrow \mathcal{M}_{\mathcal{D}}(\mathbf{F_H}, \mathbf{F_L}), \tag{5}$$

where $\mathbf{I_r}$ is the final output of the network. The first layer of $\mathcal{M}_{\mathcal{D}}$ is a $c = 32$ neuron transposed convolution layer which takes the concatenation of feature maps, $\mathbf{F_H}$ and $\mathbf{F_H}$. Then, the 32 channels output of this layer is also concatenated with the $c/2 = 16$ channels feature map from the first layer of the encoder. Finally, $c/2 = 16$-neuron and 3-neuron convolution layers complete the decoder part. The filter size of all the layers in MADANet is 3×3. The overall structure of the network is given in Figure 4.

3.3. Loss

Our version of semi-supervised training is based on the proposed method in [32], in which a set of perturbations are added to the unlabeled data, including prediction-based, feature-based, and random perturbations. Binary Cross-Entropy (BCE) loss is used as supervised training loss and Mean Square Error (MSE) between the output of the main decoder and auxiliary decoders is used for unsupervised learning. The combined loss of supervised $\mathcal{L}_s$ and unsupervised $\mathcal{L}_u$ is as following

$$\mathcal{L}_{HBRS} = \mathcal{L}_s + w_u * \mathcal{L}_u \tag{6}$$

where the unsupervised loss weight w_u follows a Gaussian distribution from zero up to a fixed weight.

Having HBRS trained, the deblurring network is trained using a pixel reconstruction loss $\mathcal{L}_r$ and a SSIM-based loss $\mathcal{L}_s$ to measure how close the reconstructe image $\mathbf{I_r}$ to the ground-truth sharp image $\mathbf{I_{gt}}$. The reconstruction loss is given by the l_1 norm:

$$\mathcal{L}_r = \|\mathbf{I_r} - \mathbf{I_{gt}}\|_1 \tag{7}$$

$\mathcal{L}_s$ is a differentiable version of the well-known full-reference image quality metric, Structural Similarity Index Measure (SSIM) [50]. Inspired by [61] we use the following format of loss where the SSIM produces 1.0 as the best score:

$$\mathcal{L}_s = \frac{1 - \text{SSIM}(\mathbf{I_r}, \mathbf{I_{gt}})}{2} \tag{8}$$

The final deblurring loss $\mathcal{L}_{deblur}$ can be writen as:

$$\mathcal{L}_{deblur} = \mathcal{L}_r + \lambda \mathcal{L}_s \tag{9}$$

where the λ is the weight for SSIM loss and we use 0.5 to match the the scale of both loss types.

An auxilary decoders after each branch is employed during training as shown in Figure 4 and both auxiliary decoders generate restored images $\mathbf{I_{high}}$ and $\mathbf{I_{low}}$. To force the low and high blur branches to focus on handling with different blur area according to the estimated mask $\mathbf{A}$, the losses $\mathcal{L}_{high}$ and $\mathcal{L}_{low}$ measuring the regional differences between restored image and ground-truth: $\mathcal{L}_{deblur}(\mathbf{I_{high}} \cdot \mathbf{A}, \mathbf{I_{gt}} \cdot \mathbf{A})$ and $\mathcal{L}_{deblur}(\mathbf{I_{low}} \cdot (\mathbf{1} - \mathbf{A}), \mathbf{I_{gt}} \cdot (\mathbf{1} - \mathbf{A}))$.

4. Experiments

4.1. Data for motion segmentation

DAVIS [37] dataset contains sequences of video frames and the annotated moving object segmentation mask for each frame. The training of HBRS module requires blurry images, events frame stack, and the moving object segmentation map as ground-truth. Inspired by the motion blur synthesis method proposed in [2], the adjacent triplet frames in DAVIS were recursively interpolated 4 times and, then the 33 interpolated frames were averaged to produce a motion-blurred image. The segmentation mask of the middle frame was selected as the ground-truth mask. A video interpolation network proposed in [8] was utilized to generate the middle frame from two adjacent frames. The public event simulator [39] is used to simulate event signals from the interpolated frames. There are 1086 training instances generated in total among them 67 are for validation. The same amount of unlabeled instances from the deblurring training dataset are selected for cross-consistency training.

4.2. Deblur dataset

GoPro dataset – We follow the official suggestion of training and testing split and by averaging nearby (the number varies from 7 to 13) frames to produce the blurry image. The corresponding events are synthesized by ESIM simulator [39]. Based on the number of frames that are averaged for a blurry image, the corresponding synthetic events are accumulated into 6 event frames. 2103 samples of blurry image, event frames stack and ground-truth frame are used for training and 1111 samples are for testing. Table 1 compares the performance of MADANet and the state-of-the-art deblurring methods. MADANet is the standard model with 5 and 6 residual blocks for low and high blur level branches, respectively. MADANet+ is the deeper version containing 10 and 11 residual blocks for corresponding branches. Among the listed methods, BHA [34], LEBMD [18], and ERDN [15] are also event-aided solutions. DGN [23] provides depth-aware deblurring, whereas MBRNN [35], GSTA [45] and PVDNet [42] are video-based (multi-frame) methods. MADANet outperforms HINet by 0.4 dB with fewer parameters, but the deeper version with moderate size widens this gap to 1 dB. In addition, the Supplementary file provides a visual comparison of real event data/blurry intensity image pair with eSLNet [49].

Table 1. Deblurring results on GoPro dataset

Method	PSNR	SSIM	Params
BHA [34]	29.06	0.943	N/A
DeepDeblur [31]	29.23	0.916	11.7M
SVDN [57]	29.81	0.937	N/A
SRN [47]	30.26	0.934	6.8M
DGN [23]	30.49	0.938	11.32M
PSS-NSC [12]	30.92	0.942	2.8M
MT-RNN [36]	31.15	0.945	2.6M
DMPHN [59]	31.20	0.945	21.7M
RADN [38]	31.76	0.953	N/A
LEBMD [18]	31.79	0.949	N/A
PVDNet [42]	31.98	0.928	23.4M
SAPHN [44]	32.02	0.953	N/A
GSTA [45]	32.10	0.960	N/A
MBRNN [35]	32.16	0.953	5.42M
BANET [48]	32.44	0.957	85.6M
MPRNET [58]	32.66	0.959	20.1M
MIMO-UNet++ [7]	32.68	0.959	16.1M
HINet [6]	32.71	0.959	88.6M
ERDN [15]	32.99	0.935	N/A
MADANET	33.09	0.958	9.9M
MADANET+	33.84	0.964	16.9M

TSlowmotion dataset – The limitations of GoPro are low image quality, insufficient diversity of scenes, and moderate motion blur for that the improvement from events may be less significant. Thus, we collected more than 200 slow-motion videos with 250 FPS using SONY RX VI camera. Our dataset contains dynamic scenes of objects' moving and natural handshaking existing. Moreover, instead of focusing on outdoor scenes we also collected many indoor videos of complex light sources. The collected high frame rate (HFR) videos are combined with the Sony-slowmotion dataset collected by [19]. The blurry images synthesis process is the similar to GoPro dataset but we average 13 to 25 frames to produce heavier motion blur. We also simulate an short exposed image that follows the blurry image. To obtain short expose frame less number of images are averaged. In that way, the ratio of long and short exposure time is set to $4:1$. In order to mimic the real world under-exposed image which usually results in insufficient light and under heavy noise, the brightness of short-exposed is reduced and a certain amount of noise is added to the image. The noise can be divided into signal dependent (Poisson) and signal independent (Gaussian) [10, 16]. The Poisson noise defines the photon noise which can be approximated as a normal distribution $\mathcal{P}(\lambda) \approx \mathcal{N}(\lambda, \lambda)$ where λ is the amount of photons hitting the sensor. A zero-mean Guassian noise model can be used to represent the other noise sources. The final amount of noise can be expressed as:

$$X \sim \mathcal{N}(\lambda, \lambda) + \mathcal{N}(0, \sigma^2) = \mathcal{N}(\lambda, \lambda + \sigma^2) \ , \qquad (10)$$

The video sequences are split into 233 for training and 36 for testing that 10213 training samples and 1528 testing

Table 2. Performance comparison on TSlowMotion Dataset

Method	PSNR	SSIM	params
LSD2 (L+S) [29]	33.32	0.939	31.04M
LSD2 (L+E)	36.05	0.965	31.04M
MIMO+ [7]	**33.88**	**0.949**	14.56M
MIMO+ (L+S)	**35.07**	**0.958**	14.56M
MIMO+ (L+E)	35.99	0.965	14.56M
BANet [48]	33.35	0.942	14.09M
BANet (L+S)	34.54	0.953	14.09M
BANet (L+E)	36.48	0.967	14.09M
MADANET	33.46	0.944	9.89M
MADANET (L+S)	34.83	0.956	9.89M
MADANET (L+E)	**37.09**	**0.971**	9.89M
MADANET (L+E+S)	37.18	0.972	9.89M

samples are obtained.

Along with MADANet, other three deblurring networks, LSD2, MIMO+, and BANet, were also trained for different input types on the TSlowmotion dataset, for a fair comparison. The original LSD2 is a U-net that takes long and short frames as input. MIMO+ and BANet achieve the state-of-the-art level performance on single image deblurring and both utilize a certain number of residual blocks. We reduce the the number of the residual blocks of MIMO+ and BANet to compare the performances of the networks with similar complexity. All 4 networks are trained with ADAM optimizer with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The batch-size is set to 8 and the patch size is 512×512. The learning rate is initialized to be 1×10^{-3} and reduced when the validation loss has stopped dropping for 5 successive epochs. The maximum number of epochs is set to be 200. For MADANet, the HBRS module is first trained semi-supervised manner for 100 epochs with ADAM optimizer and we freeze this module for the main deblurring module's training process. We trained the mentioned networks with different types of inputs: single long exposure blurry frames (L), long and short frames (L+S) and long and corresponding events stack (L+E). A special case with all 3 types of input (L+E+S) is also considered, and MADANet was trained for this case. As shown in Table 2, the performances of all networks are improved by utilizing event data. MIMO+ works the best on L and L+S deblurring thanks to the advantage of multi-scale information fusion. Compare MIMO+, the blur-aware attention network BANet achieves about 0.5 dB improvements on L+E. MADANet performs the second-best when input only contains RGB image or is concatenated with the short exposure frame, and surpasses the BANet 0.5 dB on L+E. The visual comparison over MIMO+, BANet and MADANet trained with (L+E) is shown in Figure 5. MADANet can restore more details for both foreground and background compared to the other networks.

Table 3. Performance of different components of MADANet

HBRS	*Event Injection*	PSNR	SSIM	params
✓		36.87	0.970	9.89M
	✓	36.76	0.969	7.74M
✓	✓	**37.09**	**0.971**	9.89M

Table 4. Performance of different branches

High-level	*Low-level*	*Shared*	PSNR	SSIM	params
✓			36.48	0.968	6.40M
	✓		36.67	0.969	6.40M
✓	✓		**37.09**	**0.971**	9.89M
✓	✓	✓	36.64	0.968	6.64M

4.3. Ablation Study

Table 3 shows the effectiveness of different components in MADANet: HBRS and event-injection. Without HBRS module, the MADANet becomes a two branches incremental model feeding same input to each branch. The model with two branches mechanism brings performance improvements compare to single branch models BANet and MIMO+ using much less amount of parameters. Without event-injection, the 6 event frames are first super-resolved to have same size of blurry RGB image and concatenated to RGB, which yields a 9-channel input. Comparing to the MADANet with event-injection, the performance slightly drops. In this manner, proposed injection not only bypass the possible SR requirement, but also increase the reconstruction accuracy. On the other hand, for a noisy and not perfectly registered data, we may expect from SR process to further increse the detoriation.

Table 4 shows ablation study on different branches in MADANet. The performance of low-level and high-level branches is directly measured on the output of the two auxiliary decoders after each branch. The low-level branch produce better objective scores compare to high-level branch since high level focus on local high blur regions, while the low-level branch prioritizes the rest of image. A MADANet with two branches of shared weights is also trained. The best performance is achived by fusing the outputs from two branches with different weights as expected.

4.4. Testing on Real Events

In order to evaluate how well the algorithm handles real-world event data, a group of image/event pairs was captured by a camera-rig consisting of a conventional intensity camera with global shutter mode and Prophesee Gen4.0 with a high-speed event sensor. The recorded RGB and event data are at resolution 1280x720, and a low light commercial alignment algorithm is applied on the downsampled version of RGB and event frames. This way, low-resolution event data is realistically represented and used as input to the competing algorithms as blurry images are used in their original size. This dataset, TReal, is only used for test

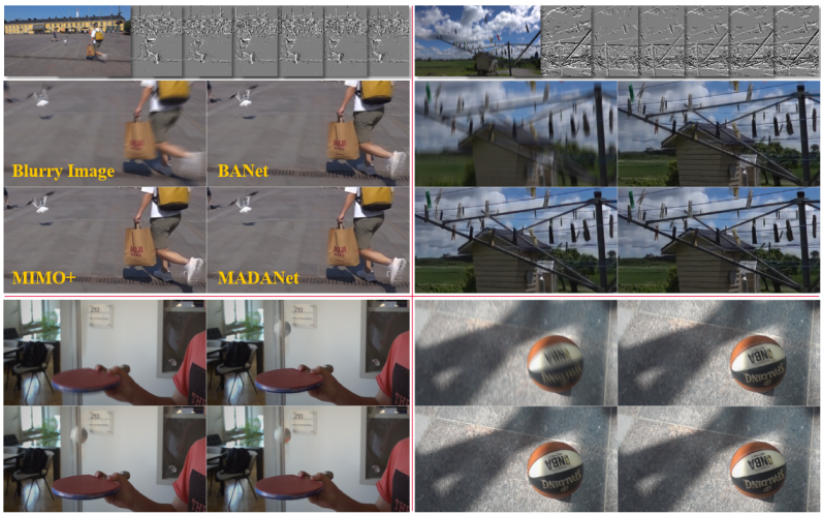

Figure 5. The visualized results from TSlowmotion dataset. BANet, MIMO+ and MADANet are trained with same type of data: blurry image + events as inputs. The first row shows two examples of input data.

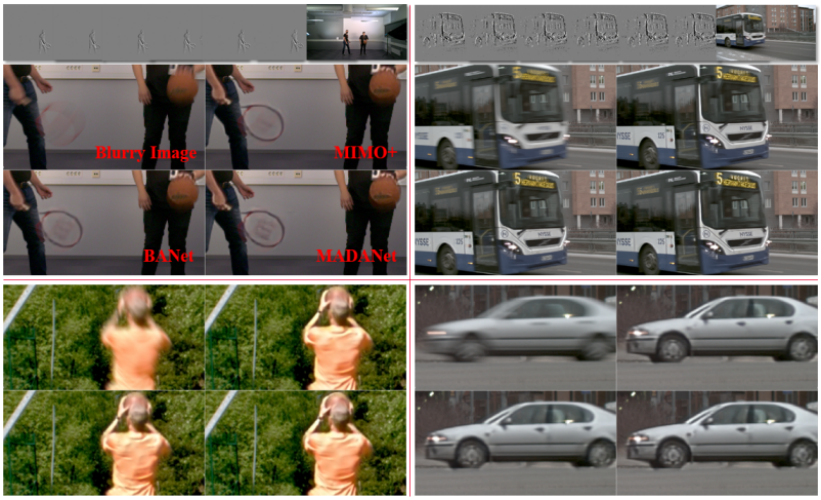

Figure 6. BANet, MIMO+ and MADANet are tested on real captured blurry image and events data. The three models are trained with L+E input data from TSlowmotion dataset and the first row shows two examples of testing inputs.

purposes, BANet, MIMO+, and MADANet are all trained with L+E input data from TSlowmotion dataset. Figure 6 shows the visual results tested by the three models and our method outperforms the other two methods. Even if the alignment is not perfect, Figure 8 of the Supplementary shows using event data clearly improves the deblurring results compared to using only the single frame or perfectly registered short/long frame pair.

5. Conclusion

In this paper, we present a novel approach for event-aided deblurring that can effectively recover a sharp image from a blurry image distorted by different motion blur. By localizing the high blur region with High Blur Region Segmentation (HBRS) module, a high blur mask and the complement of the mask can emphasize high and low-level deblurring in our double-branch deblurring network. Unlike other event-aided deblurring method, the event data is accumulated into number of frames and fused with feature maps at lower dimention. Our method achieves the state-of-the-art performance on the benchmark GoPro dataset with very limited number of parameters. Extensive experiments on our TSlowmotion dataset and visualized deblur results on real-world data demonstrate that our method outperforms other methods. Although MADANet is a single image-deblurring network, the proposed motion segmentation-assisted double attention technology can be extended to video deblurring, as future work.

References

[1] Miika Aittala and Frédo Durand. Burst image deblurring using permutation invariant convolutional neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 731–747, 2018. 1, 3

[2] Tim Brooks and Jonathan T Barron. Learning to synthesize motion blur. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6840–6848, 2019. 6

[3] Michael Cannon. Blind deconvolution of spatially invariant image blurs with phase. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(1):58–63, 1976. 2

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

[5] Meng Chang, Huajun Feng, Zhihai Xu, and Qi Li. Low-light image restoration with short-and long-exposure raw pairs. *IEEE Transactions on Multimedia*, 2021. 1, 3

[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, 6

[7] 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. 1, 2, 6, 7

[8] Myungsub Choi, Heewon Kim, Bohyung Han, Ning Xu, and Kyoung Mu Lee. Channel attention is all you need for video frame interpolation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10663–10671, 2020. 6

[9] Rob Fergus, Barun Singh, Aaron Hertzmann, Sam T Roweis, and William T Freeman. Removing camera shake from a single photograph. In *ACM SIGGRAPH 2006 Papers*, pages 787–794. 2006. 2

[10] Alessandro Foi, Mejdi Trimeche, Vladimir Katkovnik, and Karen Egiazarian. Practical poissonian-gaussian noise modeling and fitting for single-image raw-data. *IEEE Transactions on Image Processing*, 17(10):1737–1754, 2008. 6

[11] Guillermo Gallego, Tobi Delbruck, Garrick Michael Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew Davison, Jorg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 3

[12] Hongyun Gao, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Dynamic scene deblurring with parameter selective sharing and nested skip connections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3848–3856, 2019. 6

[13] Daniel Gehrig, Henri Rebecq, Guillermo Gallego, and Davide Scaramuzza. Asynchronous, photometric feature tracking using events and frames. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 750–765, 2018. 3, 4

[14] Dong Gong, Jie Yang, Lingqiao Liu, Yanning Zhang, Ian Reid, Chunhua Shen, Anton Van Den Hengel, and Qinfeng Shi. From motion blur to motion flow: A deep learning solution for removing heterogeneous motion blur. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2319–2328, 2017. 2

[15] 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. 6

[16] Samuel W Hasinoff. Photon, poisson noise., 2014. 6

[17] 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

[18] 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, 6

[19] 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 (CVPR)*, June 2019. 6

[20] Jürgen Kogler, Christoph Sulzbachner, and Wilfried Kubinger. Bio-inspired stereo vision system with silicon retina imagers. In *International Conference on Computer Vision Systems*, pages 174–183. Springer, 2009. 3, 4

[21] Jan Kotera, Jiří Matas, and Filip Šroubek. Restoration of fast moving objects. *IEEE Transactions on Image Processing*, 29:8577–8589, 2020. 3

[22] Deepa Kundur and Dimitrios Hatzinakos. Blind image deconvolution. *IEEE signal processing magazine*, 13(3):43–64, 1996. 2

[23] Lerenhan Li, Jinshan Pan, Wei-Sheng Lai, Changxin Gao, Nong Sang, and Ming-Hsuan Yang. Dynamic scene deblurring by depth guided model. *IEEE Transactions on Image Processing*, 29:5273–5288, 2020. 6

[24] Siqi Li, Yutong Feng, Yipeng Li, Yu Jiang, Changqing Zou, and Yue Gao. Event stream super-resolution via spatiotemporal constraint learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4480–4489, 2021. 2, 3

[25] Yuelong Li, Mohammad Tofighi, Junyi Geng, Vishal Monga, and Yonina C. Eldar. Efficient and interpretable deep blind image deblurring via algorithm unrolling. *IEEE Transactions on Computational Imaging*, 6:666–681, 2020. 2

[26] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128 x 128 120db 30mw asynchronous vision sensor that responds to relative intensity change. In *2006 IEEE International Solid State Circuits Conference-Digest of Technical Papers*, pages 2060–2069. IEEE, 2006. 1

[27] Min Liu and Tobi Delbruck. Adaptive time-slice block-matching optical flow algorithm for dynamic vision sensors. 2018. 3, 4

[28] Moritz B Milde, Olivier JN Bertrand, Harshawardhan Ramachandran, Martin Egelhaaf, and Elisabetta Chicca. Spik-

ing elementary motion detector in neuromorphic systems. *Neural computation*, 30(9):2384–2417, 2018. 3

[29] Janne Mustaniemi, Juho Kannala, Jiri Matas, Simo Särkkä, and Janne Heikkilä. Lsd _2–joint denoising and deblurring of short and long exposure images with cnns. *arXiv preprint arXiv:1811.09485*, 2018. 1, 3, 7

[30] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2

[31] 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. 1, 2, 6

[32] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020. 4, 5

[33] Liyuan Pan, Yuchao Dai, and Miaomiao Liu. Single image deblurring and camera motion estimation with depth map. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2116–2125. IEEE, 2019. 2

[34] 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, 6

[35] Dongwon Park, Dong Un Kang, and Se Young Chun. Blur more to deblur better: Multi-blur2deblur for efficient video deblurring. *arXiv preprint arXiv:2012.12507*, 2020. 6

[36] Dongwon Park, Dong Un Kang, Jisoo Kim, and Se Young Chun. Multi-temporal recurrent neural networks for progressive non-uniform single image deblurring with incremental temporal training. In *European Conference on Computer Vision*, pages 327–343. Springer, 2020. 6

[37] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016. 4, 6

[38] Kuldeep Purohit and AN Rajagopalan. Region-adaptive dense network for efficient motion deblurring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11882–11889, 2020. 6

[39] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. Esim: an open event camera simulator. In *Conference on Robot Learning*, pages 969–982. PMLR, 2018. 6

[40] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *European Conference on Computer Vision*, pages 184–201. Springer, 2020. 2

[41] Bongki Son, Yunjae Suh, Sungho Kim, Heejae Jung, Jun-Seok Kim, Changwoo Shin, Keunju Park, Kyoobin Lee, Jinman Park, Jooyeon 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)*, pages 66–67. IEEE, 2017. 1

[42] Hyeongseok Son, Junyong Lee, Jonghyeop Lee, Sunghyun Cho, and Seungyong Lee. Recurrent video deblurring with blur-invariant motion estimation and pixel volumes. *ACM Transactions on Graphics (TOG)*, 40(5):1–18, 2021. 6

[43] Yunjae Suh, Seungnam Choi, Masamichi Ito, Jeongseok Kim, Youngho Lee, Jongseok Seo, Heejae Jung, Dong-Hee Yeo, Seol Namgung, Jongwoo Bong, et al. A 1280× 960 dynamic vision sensor with a 4.95-μm pixel pitch and motion artifact minimization. In *2020 IEEE international symposium on circuits and systems (ISCAS)*, pages 1–5. IEEE, 2020. 1, 2, 3

[44] 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. 6

[45] Maitreya Suin and AN Rajagopalan. Gated spatio-temporal attention-guided video deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7802–7811, 2021. 6

[46] Jian Sun, Wenfei Cao, Zongben Xu, and Jean Ponce. Learning a convolutional neural network for non-uniform motion blur removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 769–777, 2015. 2

[47] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8174–8182, 2018. 1, 2, 6

[48] Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, and Chia-Wen Lin. Banet: Blur-aware attention networks for dynamic scene deblurring. *arXiv preprint arXiv:2101.07518*, 2021. 1, 2, 4, 6, 7

[49] 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. 2, 6

[50] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[51] Oliver Whyte, Josef Sivic, Andrew Zisserman, and Jean Ponce. Non-uniform deblurring for shaken images. *International journal of computer vision*, 98(2):168–186, 2012. 2, 3

[52] Patrick Wieschollek, Bernhard Schölkopf, Hendrik PA Lensch, and Michael Hirsch. End-to-end learning for image burst deblurring. In *asian conference on computer vision*, pages 35–51. Springer, 2016. 1, 3

[53] Fang Xu, Lei Yu, Bishan Wang, Wen Yang, Gui-Song Xia, Xu Jia, Zhendong Qiao, and Jianzhuang Liu. Motion deblur-

ring with real events. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2583–2592, 2021. 2

[54] Li Xu and Jiaya Jia. Depth-aware motion deblurring. In *2012 IEEE International Conference on Computational Photography (ICCP)*, pages 1–8. IEEE, 2012. 2, 3

[55] Li Xu, Shicheng Zheng, and Jiaya Jia. Unnatural l0 sparse representation for natural image deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1107–1114, 2013. 2

[56] Lu Yuan, Jian Sun, Long Quan, and Heung-Yeung Shum. Image deblurring with blurred/noisy image pairs. In *ACM SIGGRAPH 2007 papers*, pages 1–es. 2007. 1, 3

[57] Yuan Yuan, Wei Su, and Dandan Ma. Efficient dynamic scene deblurring using spatially variant deconvolution network with optical flow guided training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3555–3564, 2020. 6

[58] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14821–14831, 2021. 1, 2, 6

[59] 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. 6

[60] Shuang Zhang, Ada Zhen, and Robert L Stevenson. Deep motion blur removal using noisy/blurry image pairs. *Journal of Electronic Imaging*, 30(3):033022, 2021. 1, 3

[61] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging*, 3(1):47–57, 2016. 5