Supplementary Material for 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

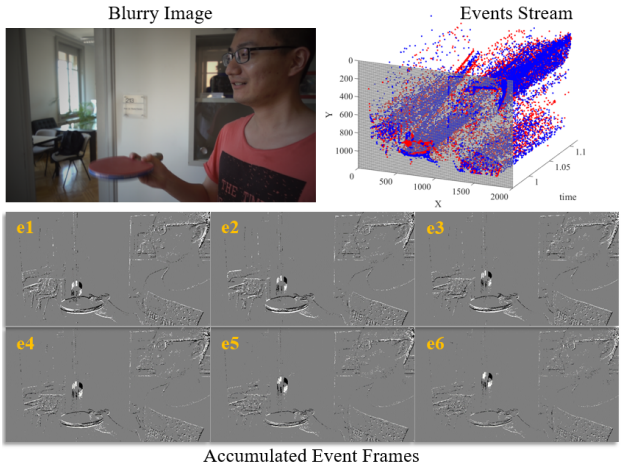

Figure 1. Standard RGB camera vs. event camera data capturing: (a) The top shows the blurry image and the corresponding event stream (red and blue marks indicate positive and negative polarity, respectively). (b) Below are the 6 event frames calculated from the event stream.

1. Event Frame Generation

As one pair are shown as an example in Figure 1, the blurry image and the event stream are captured during exposure time T. The event stream is divided uniformly into $n = 6$ chunks along the time axis. After each chunk is integrated over time, the result is quantized to produce a ternary 2-D signal, which is called an event frame. Mathematically speaking, each event frame, $\mathbf{e^i}$ is obtained as follows,

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

where $Q(h) = sgn(h)$ and $\varepsilon_{x,y}(t)$ is individual event occurs at time t, and pixel location (x, y). Instantaneous moments on a trajectory are captured on the event frames, which include important information for the deblurring network.

2. Auxiliary Decoders for MADANet

In the network design, the first branch was designed to focus on large blurs due to fast motions, etc., and the second branch can handle the rest. To ensure the network behaves in this predefined manner, a special training mechanism was created. As well as the global loss function that measures an error metric between the output image and ground truth image as a whole, two more loss functions are added to target specific areas on two branches. In the case of the high blur region deblurring branch, the error metric can only be computed for the masked region, whereas for the other branch, the error can only be computed for the complement of this mask region. Nevertheless, the output of both branches is still a feature map, not a latent image. Due to this reason, two auxiliary decoders have been developed for training, and each has been replaced in parallel with the actual decoding module. As a result, three output images are produced during the training, two of which are auxiliary images. For the actual output, the global loss function is minimized, while for the other branches, only the loss function for that region is determined.

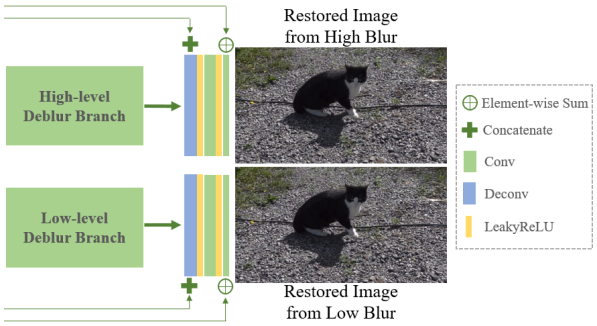

Figure 2. An auxiliary decoder after each branch is employed to enforce different branch learn different level of deblurring.

The two auxilary decoders have identical structures as shown in Figure 2. We use $\mathcal{M}_{\mathcal{D}h}$ to indicate the auxilary decoder after high-level deblur branch. The first layer of $\mathcal{M}_{\mathcal{D}h}$ is a $c = 16$ neuron transposed convolution layer

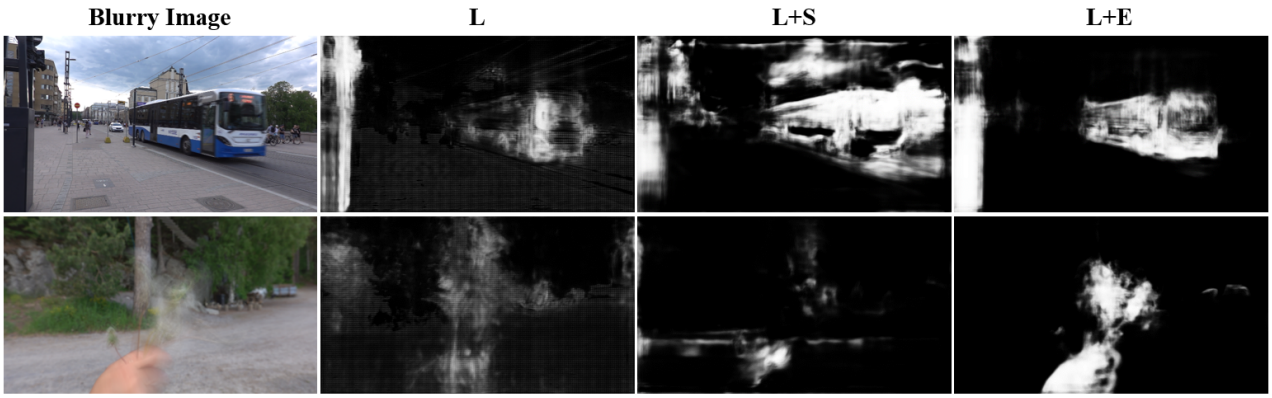

Figure 3. Masks predicted by HBRS modules for different input types; L, L+S and L+E.

which takes the feature maps from high-level deblur branch $\mathbf{F_H}$. Then, the 16 channels output of this layer is also concatenated with the $c = 16$ channels feature map from the first layer of the encoder. Finally, $c = 16$-neuron and 3-neuron convolution layers complete the decoder part. The filter size of all the layers in auxiliary decoder is 3×3. The structure of auxilary decoder after low-level deblur branch, $\mathcal{M}_{\mathcal{D}L}$ is identical to $\mathcal{M}_{\mathcal{D}h}$.

3. High Blur Region Segmentation Maps

To localize the blur regions resulting from highly relative motion to the camera, we developed an event-aided High Blur Region Segmentation Module (HBRS). The mask (or attention map) $\mathbf{A}$ created by HBRS provides the probability measure of the likelihood that every pixel lies within the high blur region. In Figure 4, one blurry image and its corresponding estimated attention map, $\mathbf{A}$, are shown. The blurry image is corrupted by deblurring caused by both moving objects and camera movement. The green box shows a stable object which is located very close to the camera. Therefore, the camera movement causes a higher level of blur (compared to other objects with higher depth) for the corresponding pixels on 2D image plane. On the other hand, the red box on the map shows a moving object (bus).

In spite of being designed to handle event frames/blurry image inputs, the network is still able to provide satisfactory results for either single image input or short/blurry input pair cases. In our experiments, we demonstrate the deblurring results for different inputs, namely; Long (L), Long and Short (L+S), Long and Event Frames (L+E). Figure 3 shows the masks predicted by HBRS for different types of inputs. Event-aided network is able to do more accurate separation compared to other input types.

Figure 4. A blurry image from TSlowmotion dataset and the corresponding estimated attention map, $\mathbf{A}$. The pixels with higher blur level due to relative high speed motion to camera are localize in $\mathbf{A}$.

4. More Visual Results

In this section, we presents more visualized results comparison from the benchmark GoPro dataset [3], TSlowmotion dataset and Real Events dataset. Figure 6 presents a comparison of state-of-the-art methods on GoPro dataset with extended examples compared to main paper. Figure 7 gives ablation study results when all the sota methods are re-trained with same type of data (L+E) on TSlowmotion dataset. In addition, Figure 8 shows the visual comparison of output of MADANets trained on and tested on different input types.

5. eSLNet vs MADANet

The recently proposed event-aided deblurring network, eSL-Net [5] is a model-based (unfolding network) intensity image recovery technique. Therefore, it is not straightforward to use it as RGB frame deblurring. However, MADANET can still be used to deblur intensity images. We have given MADANet the released eSL-Net real-time data, image/event pairs. As can be seen in Figure XX, MADANET clearly outperforms eSL-Net, even if we did not re-train it for grayscale images.

Figure 5. eSLNet vs MADANet

References

[1] 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. 4

[2] 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. 4

[3] 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

[4] 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. 4

[5] 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

Figure 6. Visual results from GoPro dataset: the proposed MADANet and the state-of-the-art deblurring methods, BANet [4], MIMO+ [2] and HINet [1].

Figure 7. The visualized results from TSlowmotion dataset. BANet, MIMO+ and MADANet are trained with same type of data: blurry image + events as inputs.

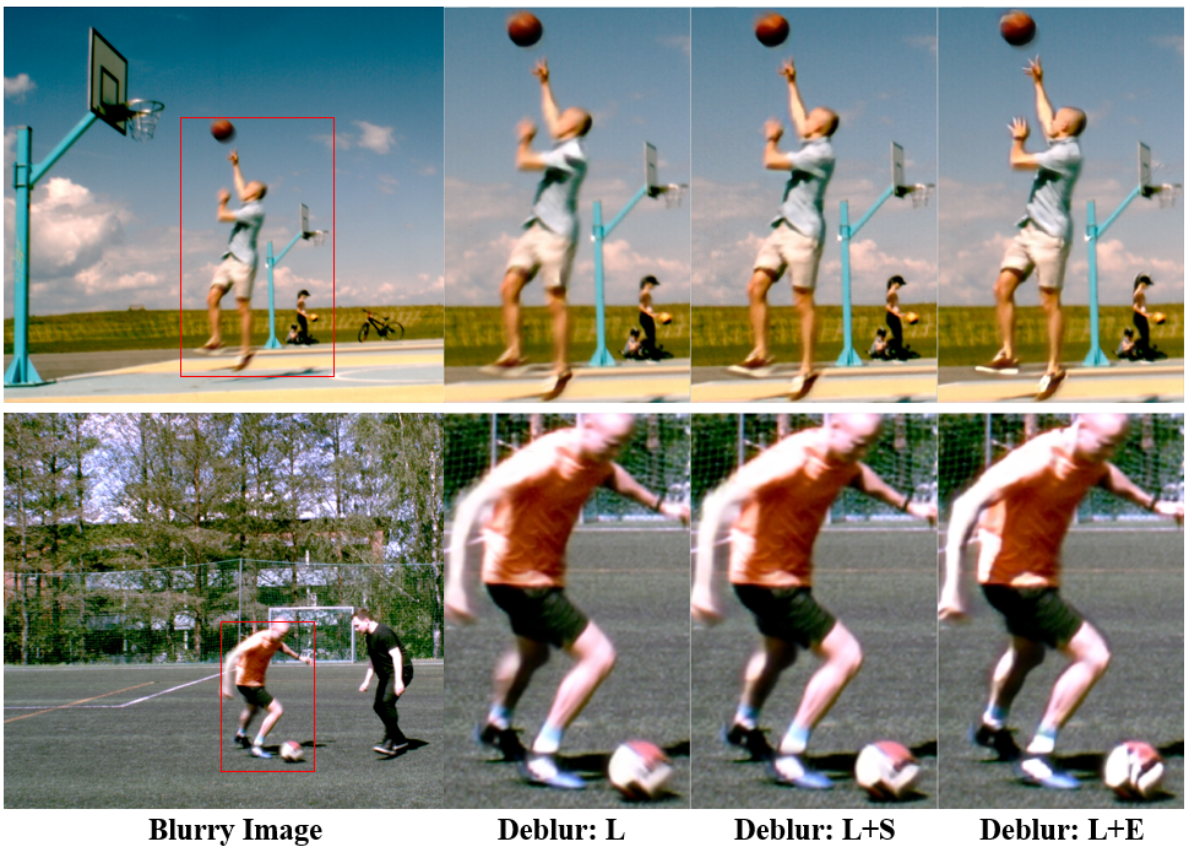

Figure 8. Visual results comparison of MADANet trained with different inputs: L, L+S and L+E where long and short images are averaged from high frame rate videos and event frames are quantized from real captured event stream.