Animation from Blur: Multi-modal Blur Decomposition with Motion Guidance

Zhihang Zhong[1,3], Xiao Sun[2], Zhirong Wu[2], Yinqiang Zheng[1], Stephen Lin[2], and Imari Sato[1,3]

[1] The University of Tokyo, zhong@is.s.u-tokyo.ac.jp
[2] Microsoft Research Asia
[3] National Institute of Informatics

1 Video Results

We present the video results as results.mp4. We first show an overview of our multi-modal framework, which uses three different interfaces for the same sample to realize blur decomposition, including guidance from prediction network, motion of video and user annotation. Then, we show the video results including comparison samples on B-Aist++, GenBlur and real-world data.

2 Quantization of Guidance

As discussed in the manuscript, guidance with two directions is essentially sufficient to represent the forward and backward movement of each individual blurred region. We present the validation loss curves of the decomposition model using 2, 4, 8, 16 directions in Fig. 1 (a). This indicates that the use of 2 directions has been able to solve the problem of directional ambiguity to some extent. We find that 4 directions bring more gain than 2 directions, and then more directions such as 8 or 16 bring limited gain. Considering the ease of using guidance (*e.g.*, user annotation) and the performance, we adopt 4 directions for other experiments.

We show the guidance representation for the case of 4 directions in Fig. 1 (b). Guidance is simplified as 5 classes for each pixel, including 4 quadrants and static, *i.e.*, quadrant I $(+1, -1)$, quadrant II $(-1, -1)$, quadrant III $(-1, +1)$, quadrant IV $(+1, +1)$, and origin $(0, 0)$, based on the motion direction. Regarding the case of 2, 8 and 16 directions, we use the similar 1-bit, 3-bit, 4-bit representation ($+1$ or -1 for each bit, all 0 for static) to mark the directions.

3 Network and Implementation Details

The architecture details of a stage network in our 2-stage decomposition network are illustrated in Fig. 2, which follows the design of [8]. The difference is that we replace the non-parametric up-sampling interpolation with a deconvolution layer. The architecture of our prediction network follows the cVAE-GAN of [10].

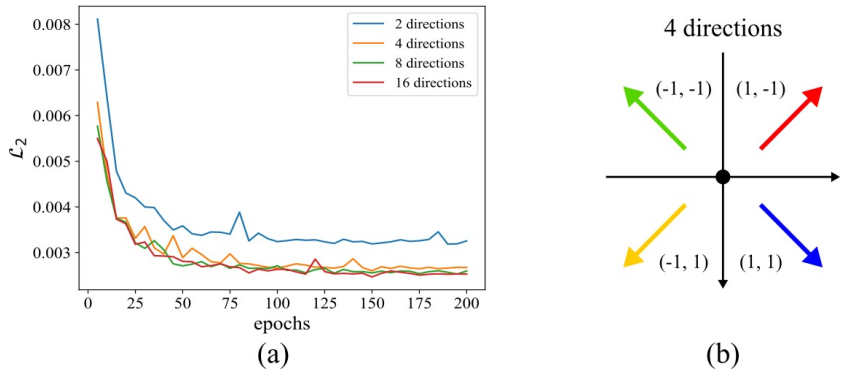

Fig. 1: **Number of guidance directions.** (a) shows the validation loss curves of the model using guidance with 2, 4, 8, and 16 directions, respectively. (b) shows the guidance representation for the case of 4 directions that we used to present the visual results.

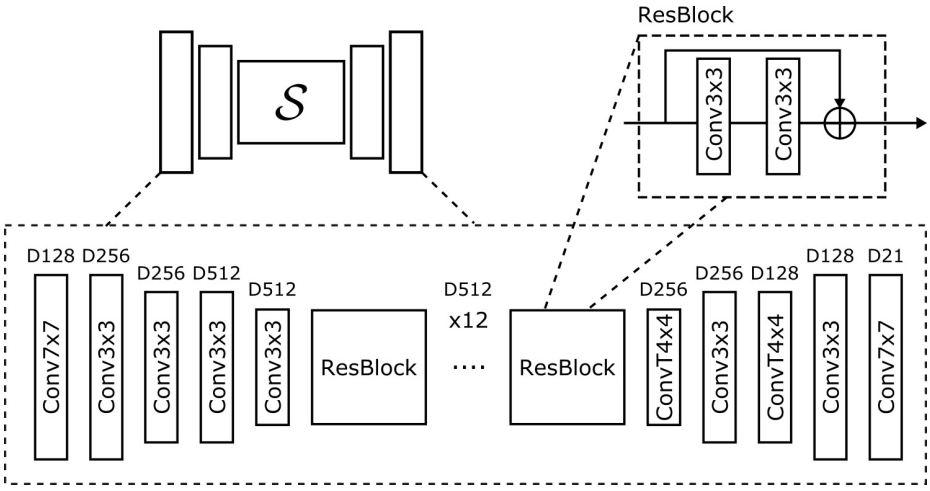

Fig. 2: **The architecture details of a network stage in our 2-stage decomposition network.** The design of our stage network follows the commonly used encoder-decoder structure. It consists of two down-sampling 2d convolution steps, 12 ResBlocks [2] bottleneck, and two up-sampling 2d deconvolution steps. D# denotes the number of channels after each convolution layer. The activation (ReLU [1]) and batch norm layers have been omitted in this figure for better clarity.

The difference is that we adopt a Gumble-Softmax [3] layer to the estimated

Fig. 3: **Multi-modal predictions on a sample of GenBlur dataset.**

guidance before calculating GAN loss. Please see the their source code for more details.

We use Pytorch [5] to implement our blur decomposition and guidance prediction networks. We trained the blur decomposer on GenBlur with a batch size of 16 for 500 epochs and on B-Aist++ with a batch size of 8 for 400 epochs. The learning rate is adjusted by a cosine scheduler with an initial learning rate of 2×10^{-4} and 1×10^{-4} for the two datasets, respectively. Flipping and random cropping (256×256) are applied for data augmentation on GenBlur. For B-Aist++, cropping is applied to frame the person in the images, which are then resized to 192×192. To prevent overfitting, the dancers in the training set and the dancers in the test set of B-Aist++ are not intersected. The number of extracted frames is fixed to 7. Furthermore, motivated by the directional ambiguity, we augment each training sample by adding its inverse direction sample $(I_b, G_{\text{inv}}, \mathbf{I}_{\text{inv}})$ to the training data. Thanks to the ability of our network to handle directional ambiguity, including the inverse samples strengthens the dependency between G and $\mathbf{I}$ and effectively increases the diversity of training samples. The setting of our guidance prediction network basically follows the cVAE-GAN of [10]. The reconstruction loss is replaced by a Cross-Entropy loss. Empirically, $\lambda_1, \lambda_2, \lambda_3$ are set as 0.1, 10, 0.1, respectively. For the comparisons, both the blurry image-based method [4] and video-based method [7] are retrained on our datasets with the same number of epochs for fairness. We will release the datasets and source code to the community.

4 Additional Qualitative Results

We additionally provide an example of multi-modal predictions on the GenBlur dataset, as illustrated in Fig. 3. Besides, we provide our guidance predictions and blur decomposition results for real-world data, along with the results of [4] in Fig. 4.

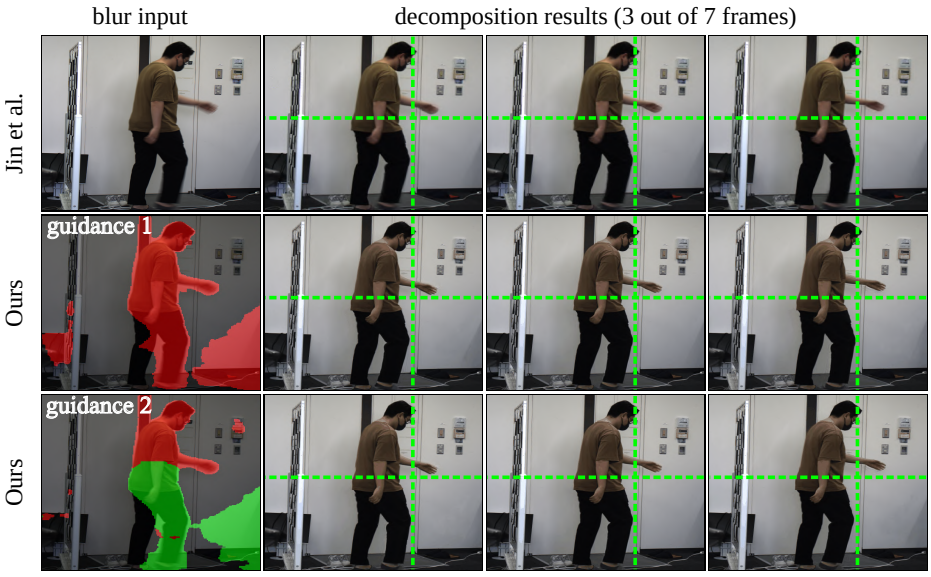

Fig. 4: **Real-world evaluations with the guidance prediction network.**

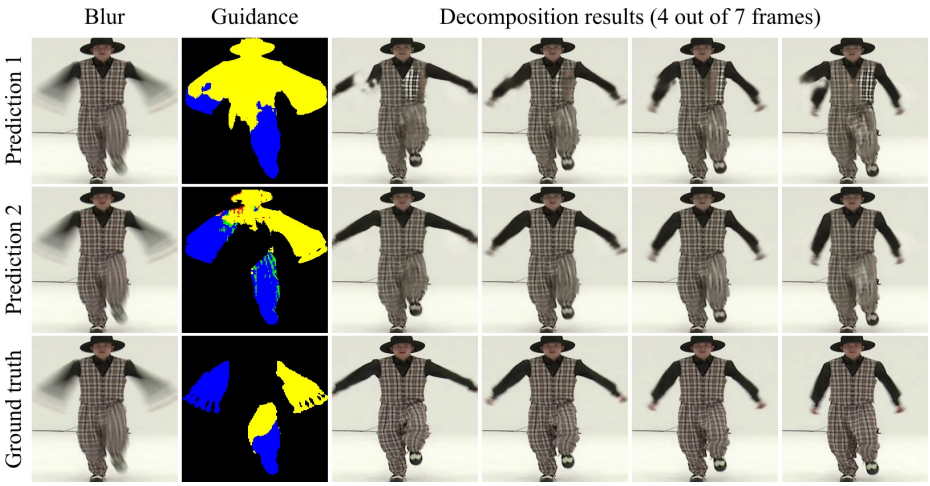

Fig. 5: **Failure case of our method.** The guidance prediction network may produce wrong guidance in the face of extremely severe blur. In addition, it is very challenging for the decomposition network to restore the details of complex texture in the case of severe blur.

5 Failure Case

We show the failure cases of our approach in the face of severe blur, as illustrated in Fig. 5. Severe blur can cause difficulties for the guidance prediction network

and thus may produce incorrect guidance like the first row (Note the left arm) of Fig. 5. In addition, it is very challenging for the decomposition network to restore the details of complex texture (Note the stripes on the right pants) in the case of severe blur. However, this does not affect the fact that the proposed guidance can help resolve the problem of directional ambiguity.

6 Limitations

The current simple form of motion guidance cannot handle the case with large and complex motion blur. Using a more elaborate guidance that takes into account different motion intensities may improve this problem. Regarding the training data, we did not consider the camera noise when synthesizing the data. This may impair the performance of the model on real-world data to some extent. Using a beam-splitter acquisition system [6,9] to collect a real-world blur decomposition dataset will be a promising future work.

References

1. Agarap, A.F.: Deep learning using rectified linear units (relu). arXiv preprint arXiv:1803.08375 (2018)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
3. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. arXiv preprint arXiv:1611.01144 (2016)
4. Jin, M., Meishvili, G., Favaro, P.: Learning to extract a video sequence from a single motion-blurred image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6334–6342 (2018)
5. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. arXiv preprint arXiv:1912.01703 (2019)
6. Rim, J., Lee, H., Won, J., Cho, S.: Real-world blur dataset for learning and benchmarking deblurring algorithms. In: European Conference on Computer Vision. pp. 184–201. Springer (2020)
7. Shen, W., Bao, W., Zhai, G., Chen, L., Min, X., Gao, Z.: Blurry video frame interpolation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5114–5123 (2020)
8. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. Advances in Neural Information Processing Systems **32**, 7137–7147 (2019)
9. Zhong, Z., Gao, Y., Zheng, Y., Zheng, B.: Efficient spatio-temporal recurrent neural network for video deblurring. In: European Conference on Computer Vision. pp. 191–207. Springer (2020)
10. Zhu, J.Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E.: Multimodal image-to-image translation by enforcing bi-cycle consistency. In: Advances in neural information processing systems. pp. 465–476 (2017)