Supplementary Material: Selection and Cross Similarity for Event-Image Deep Stereo

Hoonhee Cho and Kuk-Jin Yoon

Korea Advanced Institute of Science and Technology
{gnsgnsgml, kjyoon}@kaist.ac.kr

Abstract. Due to the limitation of space in the main paper, we provide more detailed analysis for the proposed *Selection and Cross Similarity* framework and present more experimental results in this supplementary material. Specifically, in Sec. 1, we describe more detailed implementation details for reproduction. In Sec. 2, we provide more experimental results. Lastly, in Sec. 3, we provide the discussion and future direction of this study.

1 Implementation Details

1.1 Details of Differential Event Selection Network

The proposed Differential Event Selection Network comprises a score network and image reconstruction network. In this section, we describe the implementation details of the score network, the image reconstruction network, and the Top-K algorithm.

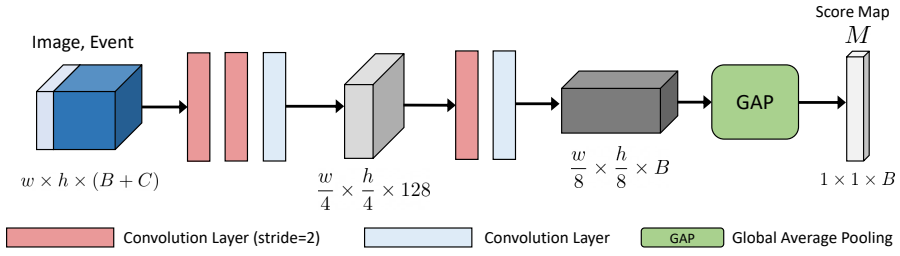

Fig. 1. The proposed network structure for score network.

Score Network. To score each bin of a voxelized event, we use image information as a condition. To consider the spatial location correlation between events

and images, we create a unified tensor through concatenation. Through a convolution layer with a stride of 2, downsampling goes through 3 times, and finally, the score for each channel is calculated by global average pooling.

Our score network is designed to be very shallow and lightweight, considering the cost. Instead, we use an image at a specific time as a condition to efficiently find events related to the scene. This mechanism is similar to [4], which generates the desired class based on a specific conditioning input. Our score network uses images with spatially dense information as conditional input and can effectively extract relevant events despite being structurally shallow design.

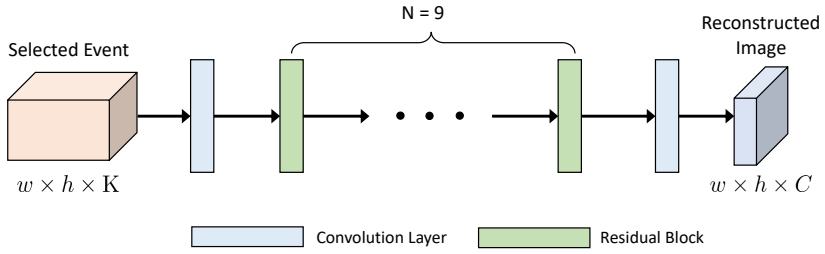

Fig. 2. The proposed network structure for image reconstruction network.

Image Reconstruction Network. We design the image reconstruction network in the differential event selection network so that the score network can be stably supervised by back-propagation from image reconstruction in the initial training. Since this image reconstruction network is not explicitly used for disparity estimation, it is only used for training and can be freely removed for inference.

Top-K Algorithm. As discussed in the main paper, the Top-K algorithm is equivalent to the following equation:

$$\arg\max_{H \in \mathcal{C}} \langle H, M\mathbf{1}^\top \rangle, \tag{1}$$

with solutions $H_{b,k} \in \{0,1\}^{B \times \mathrm{K}}$. This means that the column of matrix H is a one-hot vector that selects the index of highest score based on the score map M. Also, to consider the time sequence of event data, we introduce index-sort as a constraint as follows:

$$\begin{aligned} \mathcal{C} = \big\{ H \in \mathbb{R}^{B \times \mathrm{K}} : & H_{b,k} \geq 0, \mathbf{1}^\top H = \mathbf{1}, \\ & \sum_{i \in [B]} i H_{i,k} < \sum_{j \in [B]} j H_{j,k'} \ \ for \ \forall k < k' \big\}. \end{aligned} \tag{2}$$

This ensures that the preceding event index lies in the smaller column index of the matrix. Since event data is sequential data, keeping order is fundamental.

1.2 Details of Cross Similarity Feature Extraction.

In this subsection, we provide implementation details for Cross Similarity Feature (CSF), not covered in the main paper because of not enough space.

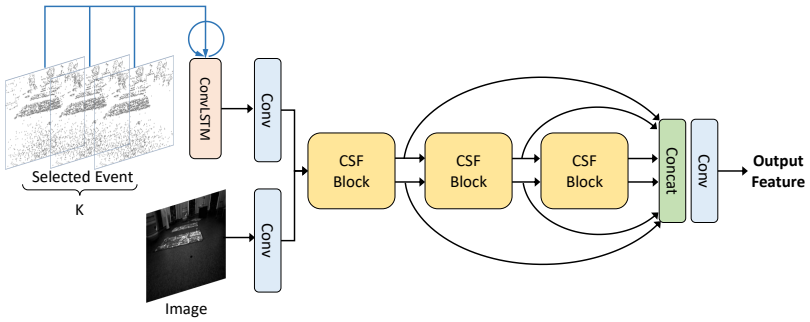

Fig. 3. The overall framework of feature extraction.

Overall Architecture of Feature Extraction. We describe the detail of the proposed feature extraction in Fig. 3. Events selected from the differential event selection network are discrete in time series. Therefore, considering temporal information, for sequential embedding, we employ ConvLSTM, proposed in [6]. After that, images and events are extracted through the convolution layer as individual features, go through a step-type CSF block, and finally completed as a unified feature. The CSF block will be explained in detail throughout next subsection.

CSF Block and increasing the receptive field. In the main paper, we discussed the generation of a Cross Similarity Feature that considers the spatial correlation of two modalities through CSF. In particular, we also consider the pixel neighborhood for the association with the event, which is temporally variable even between two frames. However, the correlation between distant events and images may be necessary depending on the motion, and we propose a strategy to increase the receptive field. Instead of increasing the size of the kernel, as shown in Fig. 4, we design a CSF block that merges features of various receptive fields, which is generated by CSF modules with multiple dilations, into one unified feature.

Various similarity function. In main paper, we exploit the widely used cosine similarity to consider the similarity between two modalities. We tried and introduced another similarity method [1,3], but it didn't work well in our setup. We remain the exploring new similarity functions as future work.

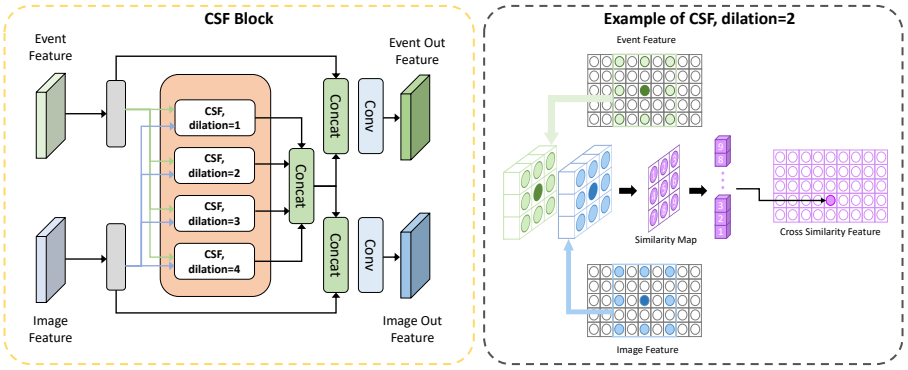

Fig. 4. Left: The detail of the CSF Block used in feature extraction. **Right:** An example of a CSF case where dilation is 2.

Table 1. Results obtained for disparity estimation on DSEC datasets. (E) implies that the event data are adopted as the input, (I) means that the image data are adopted as the input, and (E+I) means both event and image are adopted. We report the results for each sequence in three different areas (Interlaken, Thun, and Zurich City) as well as the results of all sequences averages. The best and the second best scores are **highlighted** and underlined.

Model	1PE ↓				2PE ↓				MAE ↓				RMSE ↓			
	Inter	Thun	City	All	Inter	Thun	City	All	Inter	Thun	City	All	Inter	Thun	City	All
GwcNet-g (E)	14.44	14.27	15.14	14.73	4.32	4.76	3.78	4.13	0.66	0.71	0.66	0.66	1.45	1.73	1.43	1.47
GwcNet-g (I)	5.64	6.14	7.85	6.66	1.07	1.58	1.25	1.21	0.39	0.44	0.45	0.42	0.85	1.11	0.95	0.92
Ours (E+I)	**4.86**	**5.30**	**6.62**	**5.67**	**0.87**	**1.24**	**1.05**	**0.99**	**0.36**	**0.40**	**0.41**	**0.39**	**0.79**	**1.02**	**0.87**	**0.85**

2 More Qualitative Results

2.1 The effectiveness of CSF module

We provide qualitative results for our proposed CSF module in Fig. 5. In the main paper, we have done enough quantitative ablation studies, but the qualitative results more clearly demonstrate the effectiveness of the CSF module. Compared with the model that does not use CSF, the proposed network effectively utilizes the event that hints at the boundary of the object. In particular, the CSF module effectively separates objects from the background, and sharp disparity estimation is possible.

2.2 Additional evaluation on DSEC dataset

Advantages of using the cross-modalities. We show the results of the ablation study on the use of modality in Table 1. Since we utilize GwcNet-g [2] as a baseline, the experiments on single modalities (events or images) also use the same network. Stereo disparity estimation using event modality has deficient performance compared to image-based. However, using events and images

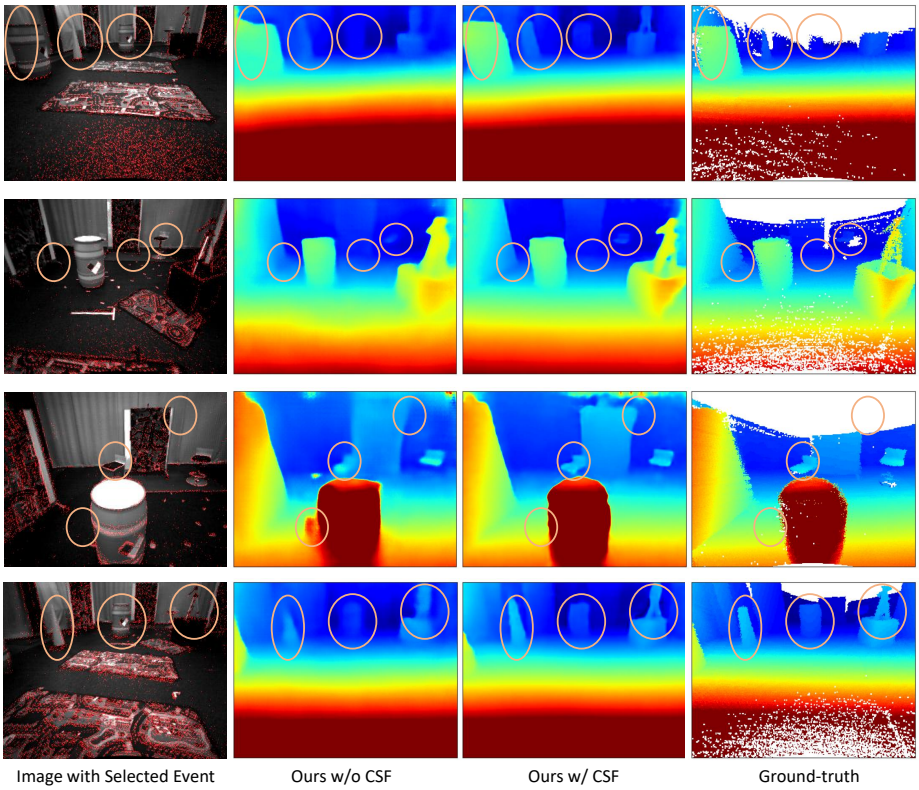

Fig. 5. Qualitative results for the use of the CSF module on the MVSEC dataset.

together, our method surpasses image-based stereo. In other words, using both modalities together is effective for stereo performance. In particular, the proposed method outperforms image-based by noticeable margins in the Zurich city sequence, which includes challenging illumination.

The results indicate that, although event data can reflect the edge and boundary information, only event-based methods cannot accurately obtain dense 3D information because events are inherently sparse and noisy. However, when spatially dense images are used together, an efficient effect is shown in stereo matching. Specifically, our method reduces from 6.66 to 5.67 in one-pixel-error for the all average sequence compared to the image-based method.

Challenging illumination scene. The advantage of using an event camera is its ability to work in challenging illumination scenes. We show the qualitative evaluation according to modality in Fig. 6. As can be seen, the event data captured in the challenging illumination (1st column) can detect objects, which are not visible in the image. Therefore, stereo depth estimation using only images in low light provides distorted or smoothed 3D information of the object (4th

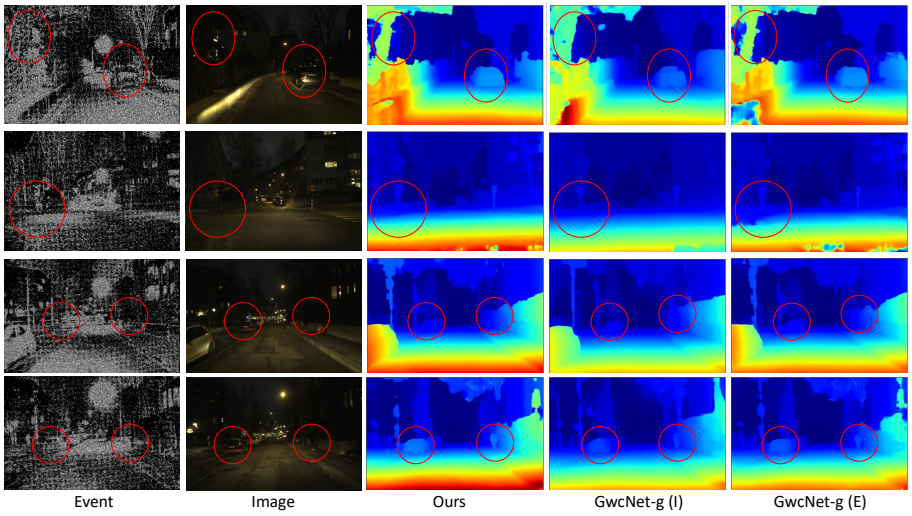

Fig. 6. Qualitative results for various modalities on the challenging illumination scene. (I) indicates that only the image is adopted as the model modality, and (E) implies that only the event data are adopted as the input. Ours means using both modalities together in our proposed method.

column). On the other hand, although an event can detect an invisible entity (5th column), it often mismatches pixel-level correspondence because it is noisy. Our model, which leads to a correlation that takes advantage of the two modalities and complements them, is also capable of sharp depth estimation even in challenging illumination circumstances.

General scene. Our fusion modality model also shows superiority in general scenes. Moreover, our model can clearly distinguish objects from the background through spatial locations correlation of events and images. The results for the general scene are shown in Fig. 7, which shows the strength of using both modalities. Image-based stereo matching has an issue in that edge-fattening at depth discontinuities and the smoothed boundary of an object. Therefore, precise stereo matching considering the separation of structure is complex. On the other hand, our method sufficiently solves the edge-flattening issue, enabling sharp depth estimation. Using event and image together leads to good results not only in challenging situations but also in general stereo matching problems.

3 Discussions

3.1 The effectiveness of event data in stereo matching

The motivation for using the event camera in stereo matching can be viewed from two perspectives.

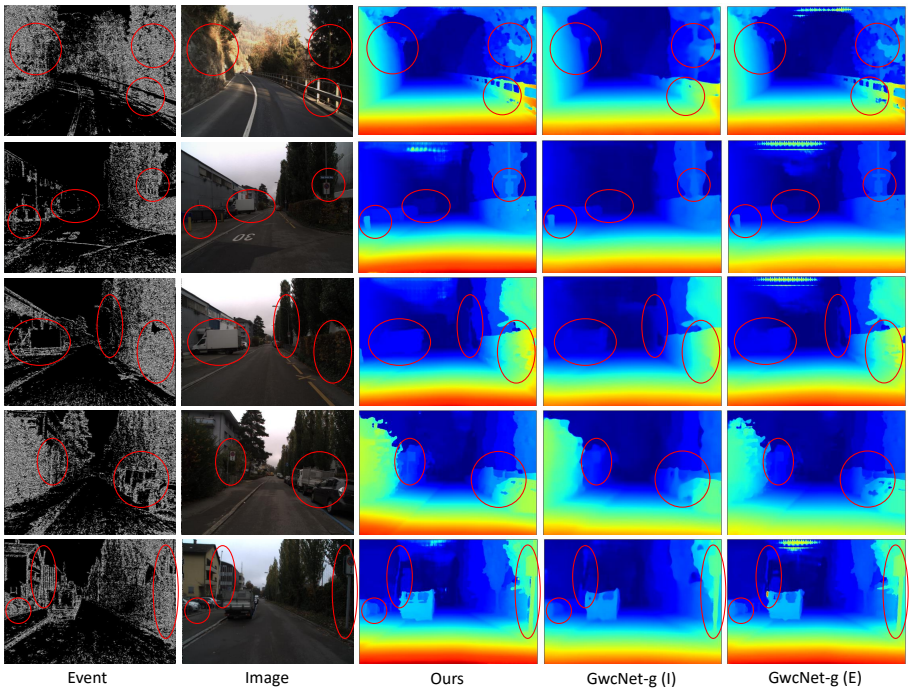

Fig. 7. Qualitative results for various modalities on the general scene.

The first is the hardware strength of the event camera. Event cameras have very low latency and cover a high dynamic range, making them intrinsically immune to motion blur and suitable for extreme lighting scenes. Therefore, the event camera can enrich the missing information as a complementary source for the shortcomings of the RGB sensor. This point of view has been mainly addressed in previous work [5, 7] and has been verified through experiments.

The second feature is that the event camera provides edge information. Since event data is mainly triggered at the boundary of the object where the intensity change occurs, it can be an ideal tool in stereo matching. However, this perspective has been less studied, so we elaborately design networks to highlight these features. In particular, we extract the refined event with a differential event selection network and consider the correlation between the two modalities through the CSF module. The experimental results show that our approach can solve the depth discontinuous in the boundary of the object, which has been a problem in the existing image-based stereo matching.

3.2 Value of this work for the community

In this paper, we propose the differential event selection network and cross similarity feature for stereo matching. From the experimental results, we demonstrate

the effectiveness of the proposed components. Especially, our approach achieve the significant performance boost than previous state-of-the-art method [5] and single modality baselines. Our approach shows advantages not only in challenging illumination conditions, but also in general scenes. Although we have focused on experiments on stereo matching, our framework is generally applicable to tasks that use images and events together. In the future work, we will extend our approach to other task, such as optical flow, object detection that require the advantage of both modalities (event and image).

References

1. Chatfield, K., Philbin, J., Zisserman, A.: Efficient retrieval of deformable shape classes using local self-similarities. In: 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops. pp. 264–271. IEEE (2009)
2. Guo, X., Yang, K., Yang, W., Wang, X., Li, H.: Group-wise correlation stereo network. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3268–3277 (2019)
3. Kim, S., Min, D., Ham, B., Ryu, S., Do, M.N., Sohn, K.: Dasc: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2103–2112 (2015)
4. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
5. Mostafavi, M., Yoon, K.J., Choi, J.: Event-intensity stereo: Estimating depth by the best of both worlds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4258–4267 (2021)
6. Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.c.: Convolutional lstm network: A machine learning approach for precipitation nowcasting. Advances in neural information processing systems **28** (2015)
7. Zhang, S., Zhang, Y., Jiang, Z., Zou, D., Ren, J., Zhou, B.: Learning to see in the dark with events. In: European Conference on Computer Vision. pp. 666–682. Springer (2020)