# Supplementary Material: Learning Adaptive Dense Event Stereo from the Image Domain

Hoonhee Cho, Jegyeong Cho, and Kuk-Jin Yoon
Visual Intelligence Lab., KAIST, Korea
{gnsgnsgml, j2k0618, kjyoon}@kaist.ac.kr

## Abstract

*Due to the limitation of space in the main paper, we provide more details about the proposed ADES and present more experimental results in this supple. Specifically, in Sec. 2, we describe the split of the DSEC target datasets. In Sec. 4, we describe more detailed implementation details for reproduction. In Sec. 5, we provide more qualitative results. Lastly, in Sec. 6, we provide additional experiments and future works.*

## 1. Dataset Licenses

In this work, we use the DSEC [6], MVSEC [11], KITTI [8], and SceneFlow [7] datasets. Each dataset is published under the following URL: (1) DSEC https://dsec.ifi.uzh.ch/. (2) MVSEC https://daniilidis-group.github.io/mvsec/ (3) KITTI https://www.cvlibs.net/datasets/kitti/ (4) SceneFlow https://lmb.informatik.uni-freiburg.de/resources/datasets/SceneFlowDatasets.en.html.

## 2. DSEC Dataset Split

Table A. Summary of DSEC dataset splits.

| Set | Sequences | Name |
|---|---|---|
| Train | Zurich City | zurich_city_00_a, zurich_city_01_a, zurich_city_02_a, zurich_city_04_a, zurich_city_05_a, zurich_city_06_a |
| | Interlaken | interlaken_00_c, interlaken_00_d, interlaken_00_e |
| Test | Zurich City | zurich_city_07_a, zurich_city_08_a |
| | Interlaken | interlaken_00_f, interlaken_00_g |

In this section, we describe more details about the split of the DSEC datasets. Since the disparity ground-truth of the test set in the DSEC dataset is not provided, we re-split the train set of DSEC to train and test set. As shown in Table A, we utilize Zurich City sequences, which are mostly recorded during the day, and the Interlaken sequences containing challenging illumination scenes.

**Algorithm 1** Motion-invariant Consistency Loss

**Input:** Input event stream $E_l^{\hat{t},T}$ and $E_r^{\hat{t},T}$
**Output:** Motion-invariant Consistency Loss

Random sample the $s$ from $\{s_1, s_2, \ldots, s_n\}$ and the $sign$ from $\{plus, minus\}$ with the uniform probability $1/n$ and $1/2$, respectively.
$\tau = s \times T$ ▷ Scale of Perturbation
**if** $sign == plus$ **then**
  $\hat{T} = T + \tau$ ▷ Perturbed Time
**else**
  $\hat{T} = T - \tau$
**end if**
Convert the event stream pairs $E_l^{\hat{t},T}, E_r^{\hat{t},T}$ accumulated during time $T$ until $\hat{t}$ into $V_l^{\hat{t},T}, V_r^{\hat{t},T}$. In the same way, the motion perturbed event stream pairs $E_l^{\hat{t},\hat{T}}, E_r^{\hat{t},\hat{T}}$ accumulated during perturbed time $\hat{T}$ until $\hat{t}$ into $V_l^{\hat{t},\hat{T}}, V_r^{\hat{t},\hat{T}}$.
$D_l^{\hat{t}} = \text{model}(V_l^{\hat{t},T}, V_r^{\hat{t},T})$
$\tilde{D}_l^{\hat{t}} = \text{model}(V_l^{\hat{t},\hat{T}}, V_r^{\hat{t},\hat{T}})$ ▷ Perturbed Disparity
$\mathcal{L}_{target}^{consistency} = L_1(D_l^{\hat{t}}, \tilde{D}_l^{\hat{t}})$ ▷ Consistency Loss
**return** $\mathcal{L}_{target}^{consistency}$

## 3. Motion-invariant Consistency Module

We provide details about the implementation of motion-invariant consistency module (MCM) in Algorithm 1. We set the samples for Algorithm 1 to $\{0.0125, 0.025, 0.0375, 0.05\}$.

## 4. Implementation Details

In this section, we provide more details about the implementation of smudge generation in Sec. 3.3 of the main paper. In the training phase, we set the number of components and compactness of superpixel algorithm [1] as 100 and 10, respectively. Then, we select 3 regions out of the components and add the smudge effect. To generate the realistic

Table B. Cross-domain comparisons with other stereo methods to MVSEC target domains. The 2-pixel error (%), 3-pixel error (%), end-point-error, and root mean square error are adopted for evaluation.

| | **KITTI-to-MVSEC** | | | | **SceneFlow-to-MVSEC** | | | |
|---|---|---|---|---|---|---|---|---|
| Method | 2PE | 3PE | EPE | RMSE | 2PE | 3PE | EPE | RMSE |
| | E2VID [9] on target domain | | | | | | | |
| AANet | 89.3 | 79.4 | 32.2 | 39.2 | 65.2 | 53.8 | 19.5 | 29.9 |
| PSMNet | 77.9 | 71.2 | 19.5 | 26.6 | 67.9 | 57.9 | 25.3 | 48.5 |
| | EventGAN [12] on source domain | | | | | | | |
| AANet | 76.5 | 64.3 | 39.3 | 19.9 | 60.8 | 49.1 | 5.8 | 12.3 |
| PSMNet | 60.5 | 40.8 | 29.4 | 15.5 | 75.0 | 56.2 | 6.2 | 13.4 |
| | **ADES (Ours)** | | | | | | | |
| AANet | 40.1 | 21.0 | 2.7 | 4.3 | 37.9 | 24.1 | 2.7 | 4.1 |
| PSMNet | 27.6 | 17.0 | 1.9 | 3.5 | **25.1** | **14.8** | **1.8** | **3.2** |

smudge effects, we utilize the diverse Blur and Optical Distortion transforms from library [3]. We utilize one of Blur, MotionBlur, and GaussianBlur in the selected regions with a uniform probability and apply OpticalDistortion with a probability of 0.3. The range of the Blur kernel is set to (15, 30), and the limit of distortion and shift in the OpticalDistortion are set to (0.2, 0.5) and (0.05, 0.12), respectively.

## 5. More Qualitative Results

We provide the more qualitative results using the KITTI dataset as source domain in Fig. A and Fig. B. The base network is PMSNet [4]. As can be seen, our ADES framework estimates the sharp and accurate disparity even trained from the image domain. On the other hand, existing image-to-event [12] or event-to-image [9] works do not resolve the gaps from domain and modality at once and thus generate inaccurate disparity containing artifacts.

## 6. Experiments on MVSEC Dataset

We conduct additional experiments using the MVSEC dataset as the target domain. Following the prior works [2, 5, 10], we use the *indoor flying* sequences. *indoor flying 1* is used for training, and *indoor flying 2* is used for testing. Table B reports the results on the MVSEC dataset for networks trained from various source domains. As shown in the table, the performance of existing stereo networks drastically decreases. The reason is that there is a difference in the environment because KITTI is the outdoor dataset, while MVSEC is the indoor dataset. This transition from outdoor to indoor makes the domain gap larger in the stereo setting. The main problem is the significant difference in baseline, *i.e.*, MVSEC is 10 cm, and KITTI is 54 cm. This problem causes a large difference in disparity distribution: MVSEC has a maximum disparity of 36, and KITTI has a maximum disparity of about 192. Even if the network is trained with many image data, in most cases of KITTI or SceneFlow, the disparity is above 100, so performance degradation in the MVSEC dataset is inevitable. Similarly, it can be seen that the performance of our ADES framework decreases when looking at 2PE compared to experiments on the DSEC dataset as the target domain (Table 1 of the main paper). Though, thanks to our self-supervision pipeline, proposed normalization, and motion-invariant consistency, ADES shows better performance than previous stereo networks by a large margin, *e.g.*, 2PE decreases from 67.9 to 25.1 comparing PSMNet+E2VID and our PSMNet+ADES in SceneFlow-to-MVSEC. Our framework still has room for improvement in significant differences of baseline with the disparity between the two domains. In future work, we will devise a way for adaptation even when the distribution difference of disparity between cross-domain is significantly large.

## References


[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012. 1

[2] Soikat Hasan Ahmed, Hae Woong Jang, SM Nadim Uddin, and Yong Ju Jung. Deep event stereo leveraged by event-to-image translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 882–890, 2021. 2

[3] Alexander Buslaev, Vladimir I Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A Kalinin. Albumentations: fast and flexible image augmentations. *Information*, 11(2):125, 2020. 2

[4] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018. 2

[5] Hoonhee Cho and Kuk-Jin Yoon. Event-image fusion stereo using cross-modality feature propagation. In *36th AAAI Conference on Artificial Intelligence (AAAI 22)*. Association for the Advancement of Artificial Intelligence, 2022. 2

[6] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, 2021. 1

[7] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016. 1

[8] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3061–3070, 2015. 1

[9] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with

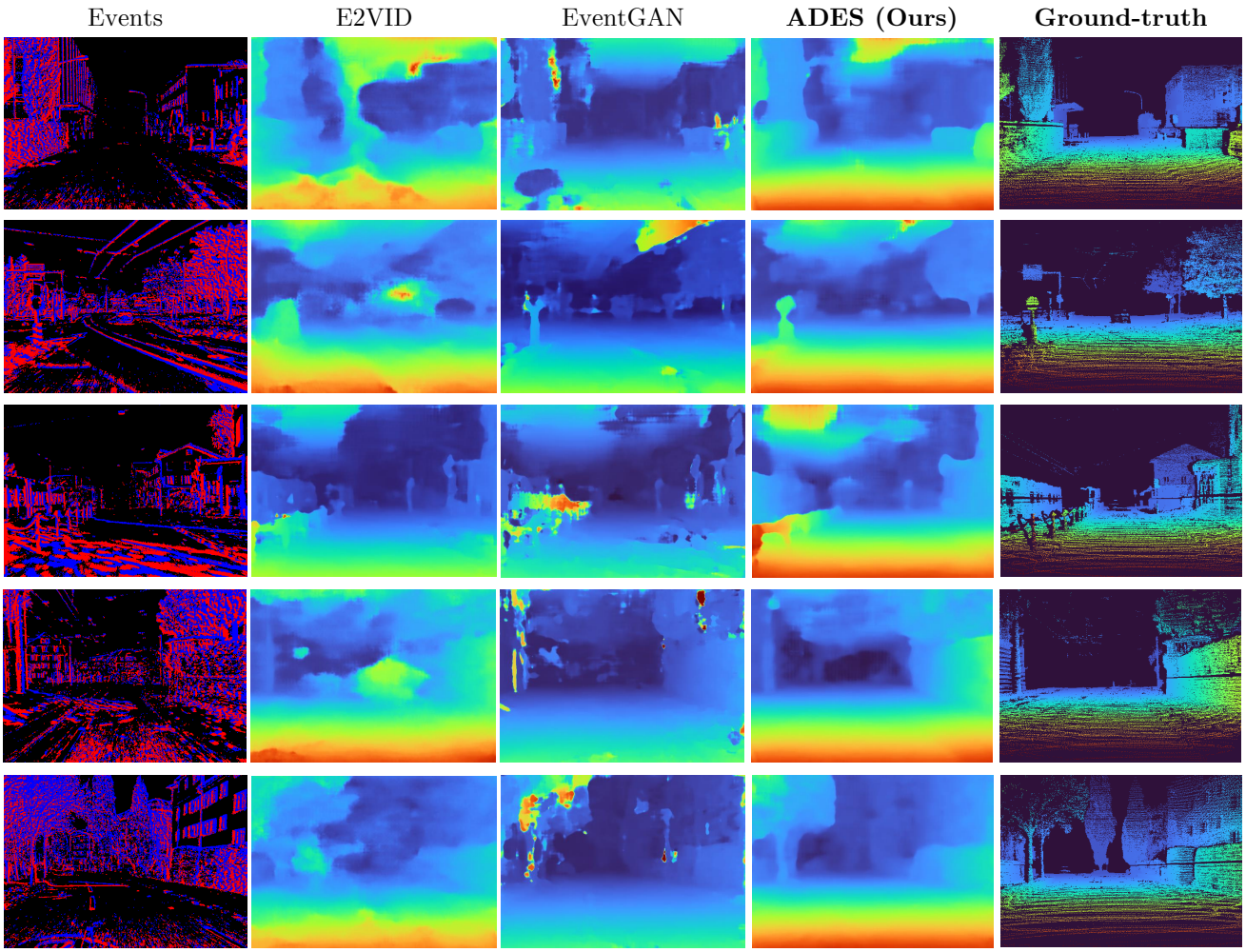


Figure A. Qualitative results for the proposed method with other methods on Zurich City sequences. Compared to EventGAN [12] and E2VID [9], our method can predict accurate and sharp disparity maps.

an event camera. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):1964–1980, 2019. 2, 3, 4

[10] Stepan Tulyakov, Francois Fleuret, Martin Kiefel, Peter Gehler, and Michael Hirsch. Learning an event sequence embedding for dense event-based deep stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1527–1537, 2019. 2

[11] Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robotics and Automation Letters*, 3(3):2032–2039, 2018. 1

[12] Alex Zihao Zhu, Ziyun Wang, Kaung Khant, and Kostas Daniilidis. Eventgan: Leveraging large scale image datasets for event cameras. In *2021 IEEE International Conference on Computational Photography (ICCP)*, pages 1–11. IEEE, 2021. 2, 3, 4

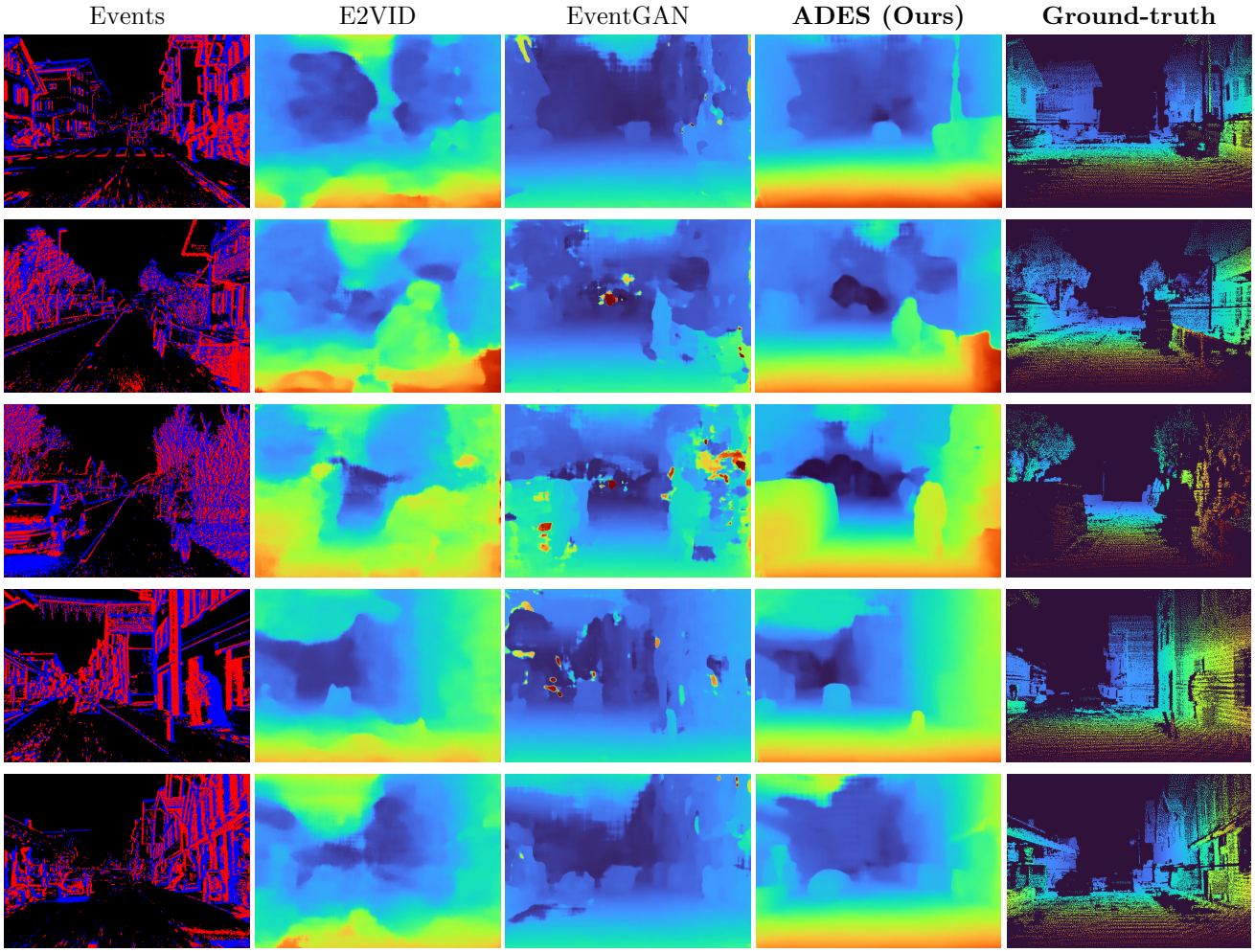


Figure B. Qualitative results for the proposed method with other methods on Interlaken sequences. Compared to EventGAN [12] and E2VID [9], our method can predict accurate and sharp disparity maps.