Event Camera Data Dense Pre-training

Yan Yang[2], Liyuan Pan[1], and Liu Liu[3]

[1] School of CSAT, Beijing Institute of Technology, Beijing, China
Corresponding author
[2] BDSI, Australian National University, Canberra, Australia
[3] KooMap Dept., Huawei, Beijing, China
Yan.Yang@anu.edu.au Liyuan.Pan@bit.edu.cn liuliu33@huawei.com

Abstract. In this supplementary material, we present content omitted from the main paper due to space constraints. The supplementary material is organized into three sections. Specifically, i) in Sec. 1, we provide the pipeline for synthesizing our E-TartanAir dataset. The details of the E-TartanAir dataset are also provided; ii) in Sec. 2, we give details of our pre-training process and details of transferring the pre-trained network to downstream tasks; iii) in Sec. 3, we conduct additional analyses on our model components.

1 Synthetic Dataset

Table 1: *Details of our E-TartanAir dataset.*

Scene name	Number of samples
office2	8534
endofworld	8423
hospital	17772
gascola	13199
westerndesert	6312
abandonedfactory	15498
amusement	7750
office	9870
ocean	7757
oldtown	8397
japanesealley	4417
seasonsforest	4629
seasonsforest_winter	10304
neighborhood	28733
soulcity	13228
abandonedfactory_night	14523
carwelding	3772
seasidetown	6535

We curate a synthetic event camera dataset, E-TartanAir, by synthesizing data from the TartanAir dataset [18]. The synthesis is performed at a resolution of 480×640,

Table 2: *Hyperparameters for the E-TartanAir dataset synthesis.*

Hyperparameter	Value
dvs_exposure	duration 0.033
pos_thres	0.2
neg_thres	0.2
sigma_thres	0.03
cutoff_hz	30
leak_rate_hz	0.1
shot_noise_rate_hz	Uniform$(0, 0.25)$

utilizing sequences from the TartanAir dataset. Approximately 180K samples are generated, employing EMA-VFI [21] for frame interpolations and V2E [10] for converting frames into events. Details of the dataset are provided in Tab. 1. The hyperparameters of V2E are given in Tab. 2.

2 Experiment Details

2.1 Pre-training Processes

We perform pre-training of the network on the E-TartanAir dataset using a batch size of 1024 for 300 epochs. Event data is converted to frames for pre-training, utilizing a six-channel voxel grid [23]. Our pre-training hyperparameters adhere to [4,14], and they are detailed in Tab. 3. Specifically, a cosine scheduler with a warm-up epoch of 10, a peak learning rate (peak_lr) of 2×10^{-3}, and a minimum learning rate (min_lr) of 1×10^{-6} are employed. With a cosine schedule, weight decay is increased from 4×10^{-2} to 4×10^{-1} during pre-training. For EMA [8], the momentum is set to 0.992 and scaled to 1 at the end of pre-training using a cosine scheduler. The projection head architecture from [14] is utilized, and an attention pooling layer is incorporated for projection heads in context-level and image-level similarity mining. The number of prototypes (i. e., output dimension) for all projection heads is fixed at 2^{15}. During pre-training, a multi-crop strategy with 2 global crops and 8 local crops is employed [3,4,14]. The resolutions of the global and local crops are respectively set to 192×192 and 96×96.

When comparing with DINOv2, the timm codebase is used to interpolate ViT-S/14 to ViT-S/16 for consistency with other baseline methods.

2.2 Semantic Segmentation

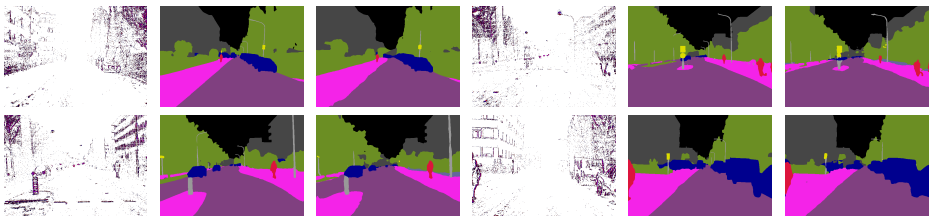
We perform fine-tuning on our backbone network with a UperNet decoder [19] for semantic segmentation. The network is optimized using both cross-entropy and Dice losses [16], with their weights set to 1, as suggested in [17]. During inference, a test time augmentation with HorizontalFlip and Scale (1,1.5,2.0) is used. Receipts of the fine-tuning process are presented in Tab. 4, and samples of semantic segmentation results from our method are provided in Fig. 1.

Table 3: *Hyperparameters for pre-training.*

Hyperparameters	Value
optimizer	AdamW
batch size	1024
epochs	300
peak_lr	2×10^{-3}
min_lr	1×10^{-6}
lr scheduler	cosine
lr warm-up epochs	10
weight decay	4×10^{-1}
minimum weight decay	4×10^{-2}
weight decay scheduler	cosine
momentum	0.992
momentum end	1.0
momentum scheduler	cosine
drop path rate	1×10^{-1}

Table 4: *Fine-tuning hyperparameters for semantic segmentation on the DDD17 and DSEC datasets.*

Hyperparameters	DDD17	DSEC
optimizer	AdamW	AdamW
peak_lr	2.5×10^{-4}	2.5×10^{-4}
min_lr	0	0
weight decay	5×10^{-2}	5×10^{-2}
batch size	8	8
epochs	100	50
warmup epochs	10	10
lr scheduler	cosine	cosine
drop path rate	1×10^{-1}	1×10^{-1}



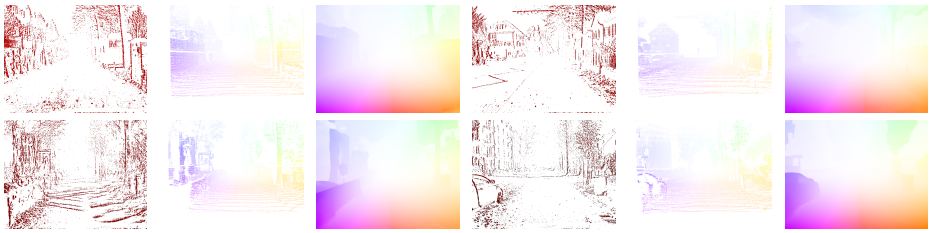
(a) (b) (c) (d) (e) (f)

Fig. 1: *Qualitative results of semantic segmentation. (a) and (d): event images. Red and blue pixels depict positive and negative events, respectively. (b) and (e): ground-truth semantic segmentation labels. (c) and (f): our model predictions.*

As a finding in the paper and the evolving feature pyramid structure for using ViT with the UperNet, a learning rate smaller than [20] leads to better fine-tuning performance, we re-benchmark the baseline methods in Tab. 1 of the main paper, by performing a binary search on the learning rate, and ablating with 4 stages feature pyramid

Table 5: *Fine-tuning hyperparameters for optical flow estimation on the MVSEC dataset and DSEC dataset.*

Hyperparameters	MVSEC	DSEC
optimizer	AdamW	AdamW
peak_lr	2×10^{-4}	2×10^{-4}
min_lr	0	1×10^{-6}
weight decay	1×10^{-4}	1×10^{-4}
batch size	6	6
epochs	150	150
warmup epochs	1	1.5
lr scheduler	cosine	cosine
gradient clipping	1	1



(a) (b) (c) (d) (e) (f)

Fig. 2: *Qualitative results of optical flow estimation. (a) and (d): event images. Red and blue pixels depict positive and negative events, respectively. (b) and (e): ground-truth optical flow. (c) and (f): our model predictions.*

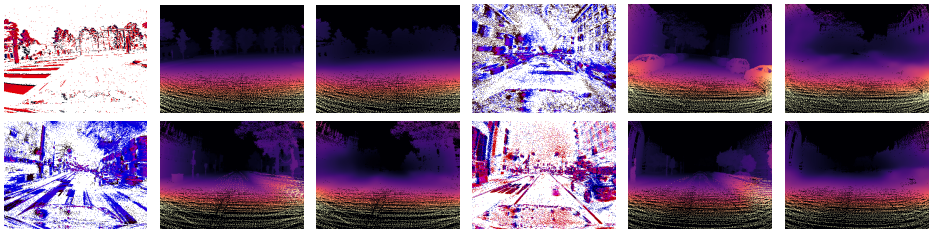
Table 6: *Fine-tuning hyperparameters for depth estimation on the MVSEC dataset.*

Hyperparameters	Value
optimizer	Adam
peak_lr	2×10^{-4}
weight decay	1×10^{-4}
batch size	16
iterations	5.6K
warm-up epochs	0
lr scheduler	cosine

structures [1] and simple feature pyramids [12]. The best performance is reported for the baseline methods.

2.3 Optical Flow Estimation

For the fine-tuning of our pre-trained backbone network for optical flow estimation, we employ TMA [13] architectures. The context encoder and feature encoder are constructed from the first two layers, comprising 4 transformer blocks, selected from the pre-trained backbone network. The optimization is performed using the sequence L1



(a) (b) (c) (d) (e) (f)

Fig. 3: *Qualitative results of depth estimation. (a) and (d): event images. Red and blue pixels depict positive and negative events, respectively. (b) and (e): ground-truth depth. (c) and (f): our model predictions.*

loss with a gamma value of 0.8. For the MVSEC dataset, we perform intermediate fine-tuning with the DSEC dataset. Then, we validate on the testing data selected by [2], after training on the remaining data [22]. The hyperparameters for fine-tuning on the MVSEC dataset and DESC dataset are detailed in Tab. 5, primarily adopted from [13]. Samples of optical flow estimations from our method can be found in Fig. 2. Note that the sample predictions are generated from the evaluation folds specified in [13], considering that the ground truth optical flow of the testing folds on the DSEC dataset is not publicly released. Consistent with Sec. 2.2, the baseline methods are re-benchmarked, to mine their best performance.

2.4 Depth Estimation

In the light of [15], we conduct fine-tuning on our pre-trained backbone network for depth estimation, and replace half of channels of BatchNorm layers to GroupNorm layers. The fine-tuning hyperparameters for depth estimation are provided in Tab. 6. The network optimization, borrowed from [7], is performed using the scale-invariant loss [5] and multi-scale scale-invariant gradient matching loss [9], with respective weights set to 1 and 0.25, following [7]. Throughout training, the network is trained to predict normalized log depth, and the predictions are inverted back to metric depth during testing. Samples of depth estimations from our method are presented in Fig. 3. Following [7], the baseline methods are benchmarked.

3 Discussions

We present additional ablations conducted on the DSEC semantic segmentation dataset [6, 17] to further validate our model components. We set the pre-training backbone and dataset to the Swin-T/7 and E-TartanAir dataset, except where otherwise indicated.

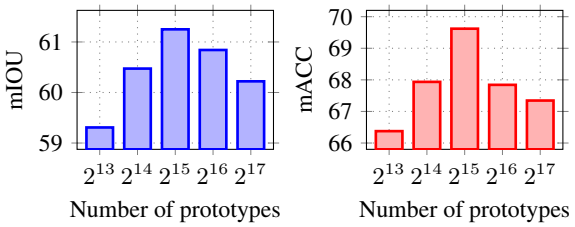


Fig. 4: *Comparison of the number of prototypes used in $\mathcal{H}_s^c$ and $\mathcal{H}_t^c$.*

3.1 Number of Prototypes

Fig. 4 investigates the impact of the number of prototypes (i. e., output dimension) used in the projection heads $\mathcal{H}_s^c$ and $\mathcal{H}_t^c$. We observe that using 2^{15} as the number of prototypes yields the best performance. Further increasing the number of prototypes introduces excessive complexity to the pre-training task, resulting in a performance decrease.

3.2 Loss Weight

We ablate the weight λ_1 in Eq. 5 of the main paper for pre-training our network. Note that loss weight is automatically set per the number of local and global crops used in pre-training, and an L1 normalization is performed on λ_1 and λ_2 [14], *e.g.*, $\lambda_1 = \frac{2}{2+2\times8+2} = 0.1$. The ablation for λ_1 before the normalization is given in Fig. 5. Having $\lambda_1 = 2.0$ (before applying the normalization) finds the best performance.

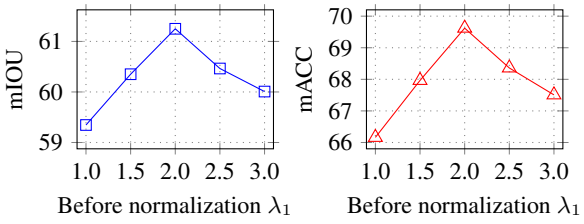


Fig. 5: *Comparison of loss weight.*

3.3 Object Recognition

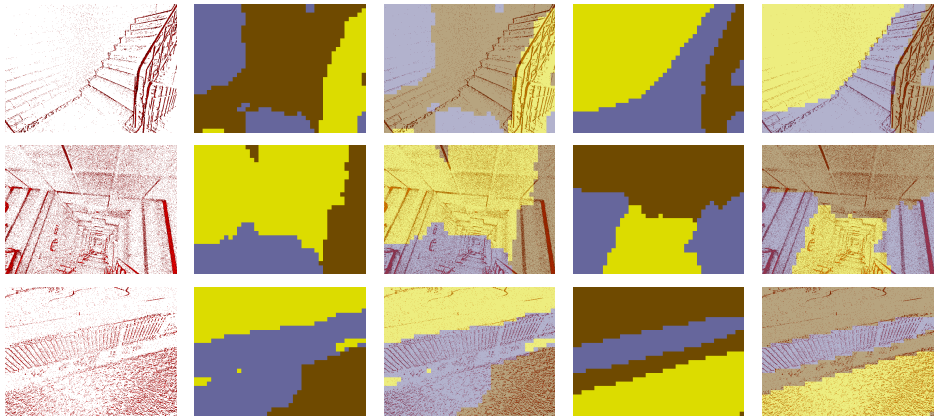
To show the ability of our method for event-based object recognition, we fine-tune our approach and compare with the past state-of-the-art method, ECDP [20], on the largest event object recognition dataset, N-ImageNet [11]. With a ViT-S/16 backbone, ECDP achieves 64.836% top-1 and 86.296% top-5 accuracies, and our method achieves 66.066% top-1 and 87.254% top-5 accuracies. Despite our pre-training framework primarily focusing on dense prediction tasks, we demonstrate state-of-the-art performance in object recognition.

3.4 Additional Visualizations

Fig. 6 gives sample results of patches belonging to different contexts. Our method successfully mines contexts in an event image, and groups patches with the same semantics.

References

1. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: BERT pre-training of image transformers. In: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net (2022), https://openreview.net/forum?id=p-BhZSz59o4
4
2. Barchid, S., Mennesson, J., Djeraba, C.: Exploring joint embedding architectures and data augmentations for self-supervised representation learning in event-based vision. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023. pp. 3903–3912. IEEE (2023). https://doi.org/10.1109/CVPRW59228.2023.00405, https://doi.org/10.1109/CVPRW59228.2023.00405 5



(a) (b) (c) (d) (e)

Fig. 6: *Sample results of patches belonging to different contexts on the E-TartanAir dataset. (a): input event images. (b): mined context labels (without enforcing the context-level similarity). (c): mined context labels (enforcing the context-level similarity). (d) and (e): blends of the event image with context labels from (b) and (c) for visualization purposes, respectively.*

3. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (2020), https://proceedings.neurips.cc/paper/2020/hash/70feb62b69f16e0238f741fab228fec2-Abstract.html 2
4. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 9630–9640. IEEE (2021). https://doi.org/10.1109/ICCV48922.2021.00951, https://doi.org/10.1109/ICCV48922.2021.00951 2
5. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada. pp. 2366–2374 (2014), https://proceedings.neurips.cc/paper/2014/hash/7bccfde7714a1ebadf06c5f4cea752c1-Abstract.html 5
6. Gehrig, M., Aarents, W., Gehrig, D., Scaramuzza, D.: DSEC: A stereo event camera dataset for driving scenarios. IEEE Robotics Autom. Lett. **6**(3), 4947–4954 (2021). https://doi.org/10.1109/LRA.2021.3068942, https://doi.org/10.1109/LRA.2021.3068942 5
7. Hamaguchi, R., Furukawa, Y., Onishi, M., Sakurada, K.: Hierarchical neural memory network for low latency event processing. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 22867–22876. IEEE (2023). https://doi.org/10.1109/CVPR52729.2023.02190, https://doi.org/10.1109/CVPR52729.2023.02190 5
8. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.B.: Momentum contrast for unsupervised visual representation learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. pp. 9726–9735. Computer

Vision Foundation / IEEE (2020). https://doi.org/10.1109/CVPR42600.2020.00975, https://doi.org/10.1109/CVPR42600.2020.00975 2
9. Hidalgo-Carrió, J., Gehrig, D., Scaramuzza, D.: Learning monocular dense depth from events. In: Struc, V., Fernández, F.G. (eds.) 8th International Conference on 3D Vision, 3DV 2020, Virtual Event, Japan, November 25-28, 2020. pp. 534–542. IEEE (2020). https://doi.org/10.1109/3DV50981.2020.00063, https://doi.org/10.1109/3DV50981.2020.00063 5
10. Hu, Y., Liu, S., Delbrück, T.: v2e: From video frames to realistic DVS events. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021. pp. 1312–1321. Computer Vision Foundation / IEEE (2021). https://doi.org/10.1109/CVPRW53098.2021.00144, https://openaccess.thecvf.com/content/CVPR2021W/EventVision/html/Hu_v2e_From_Video_Frames_to_Realistic_DVS_Events_CVPRW_2021_paper.html 2
11. Kim, J., Bae, J., Park, G., Zhang, D., Kim, Y.M.: N-imagenet: Towards robust, fine-grained object recognition with event cameras. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2146–2156 (October 2021) 6
12. Li, Y., Mao, H., Girshick, R.B., He, K.: Exploring plain vision transformer backbones for object detection. In: Avidan, S., Brostow, G.J., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IX. Lecture Notes in Computer Science, vol. 13669, pp. 280–296. Springer (2022). https://doi.org/10.1007/978-3-031-20077-9_17, https://doi.org/10.1007/978-3-031-20077-9_17 4
13. Liu, H., Chen, G., Qu, S., Zhang, Y., Li, Z., Knoll, A., Jiang, C.: Tma: Temporal motion aggregation for event-based optical flow. In: ICCV (2023) 4, 5
14. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P., Li, S., Misra, I., Rabbat, M.G., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. CoRR **abs/2304.07193** (2023). https://doi.org/10.48550/arXiv.2304.07193, https://doi.org/10.48550/arXiv.2304.07193 2, 6
15. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 12159–12168. IEEE (2021). https://doi.org/10.1109/ICCV48922.2021.01196, https://doi.org/10.1109/ICCV48922.2021.01196 5
16. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M.J.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Cardoso, M.J., Arbel, T., Carneiro, G., Syeda-Mahmood, T.F., Tavares, J.M.R.S., Moradi, M., Bradley, A.P., Greenspan, H., Papa, J.P., Madabhushi, A., Nascimento, J.C., Cardoso, J.S., Belagiannis, V., Lu, Z. (eds.) Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support - Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, 2017, Proceedings. Lecture Notes in Computer Science, vol. 10553, pp. 240–248. Springer (2017). https://doi.org/10.1007/978-3-319-67558-9_28, https://doi.org/10.1007/978-3-319-67558-9_28 2
17. Sun, Z., Messikommer, N., Gehrig, D., Scaramuzza, D.: ESS: learning event-based semantic segmentation from still images. In: Avidan, S., Brostow, G.J., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXIV. Lecture Notes in Computer Science, vol. 13694, pp. 341–357. Springer (2022). https://doi.org/10.1007/978-3-031-19830-4_20, https://doi.org/10.1007/978-3-031-19830-4_20 2, 5

18. Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., Scherer, S.A.: Tartanair: A dataset to push the limits of visual SLAM. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2020, Las Vegas, NV, USA, October 24, 2020 - January 24, 2021. pp. 4909–4916. IEEE (2020). https://doi.org/10.1109/IROS45743.2020.9341801, https://doi.org/10.1109/IROS45743.2020.9341801 1
19. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part V. Lecture Notes in Computer Science, vol. 11209, pp. 432–448. Springer (2018). https://doi.org/10.1007/978-3-030-01228-1_26, https://doi.org/10.1007/978-3-030-01228-1_26 2
20. Yang, Y., Pan, L., Liu, L.: Event camera data pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10699–10709 (October 2023) 3, 6
21. Zhang, G., Zhu, Y., Wang, H., Chen, Y., Wu, G., Wang, L.: Extracting motion and appearance via inter-frame attention for efficient video frame interpolation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 5682–5692. IEEE (2023). https://doi.org/10.1109/CVPR52729.2023.00550, https://doi.org/10.1109/CVPR52729.2023.00550 2
22. Zhu, A.Z., Thakur, D., Özaslan, T., Pfrommer, B., Kumar, V., Daniilidis, K.: The multi vehicle stereo event camera dataset: An event camera dataset for 3d perception. CoRR **abs/1801.10202** (2018), http://arxiv.org/abs/1801.10202 5
23. Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K.: Unsupervised event-based learning of optical flow, depth, and egomotion. CoRR **abs/1812.08156** (2018), http://arxiv.org/abs/1812.08156 2