# Supplemental Material

**Evaluation of Deep Learning based 3D-Point-Cloud Processing Techniques for Semantic Segmentation of Neuromorphic Vision Sensor Event-Streams**

Tobias Bolten[1], Felix Lentzen[1], Regina Pohle-Fröhlich[1] and Klaus D. Tönnies[2]
[1]*Institute of Pattern Recognition, Niederrhein University of Applied Sciences, Reinarzstr. 49, Krefeld, Germany*
[2]*Department of Simulation and Graphics, University of Magdeburg, Universitätsplatz 2, Magdeburg, Germany*
{*tobias.bolten, felix.lentzen, regina.pohle*}*@hs-niederrhein.de, klaus@isg.cs.uni-magdeburg.de*

This document provides additional details, example visualizations, precise network configurations and more quantitative as well as qualitative results in addition to the main paper.

## 1 Event Counts in "DVS-OUTLAB"

Refer to Section 4.2 of main paper

The DVS-OUTLAB dataset (Bolten et al., 2021) used in the evaluation is composed of two different parts, namely, of staged scenes with focus on objects and of scenes that mainly contain environmental influences. The object scenes are plain sensor recordings with the spatial resolution of $768px \times 512px$. Figure 1a illustrates the event count contained in these raw recordings within 60ms temporal windows.

Section 4.2 of the main paper explains the processing pipeline used for the generation of the space-time event clouds. Here, a Patch-Of-Interest method was introduced, since the pre-mentioned event count was too high for processing if the entire spatial resolution was included. The effects of filtering and sampling in this pipeline are shown in detail in Figure 2. Details are given for the entire dataset, the staged object scenes and the environmental influences that formed the basis of the decision in implementing these steps of the processing pipeline.

## 2 Event-to-Frame-Encodings

Refer to Section 4.3 of main paper

Compared to the main paper, additional examples of event-to-frame encodings are given in Figure 3. Frames are randomly selected from the test set of the DVS-OUTLAB dataset and were used in the 2D Mask-R-CNN baseline evaluation.

## 3 Network configurations and training

Refer to Section 4.5 and 4.6 of main paper

The exact settings, e.g. with respect to the radii selection, number of events in the set-abstraction logic, size of the MLP in the network layers are summarized in relation to the performed meta-parameter optimization in Tables 1 and 2. This should eliminate any ambiguities in terms of comprehensibility and reproducibility.

An overview of the course of the training is given in Figure 4.

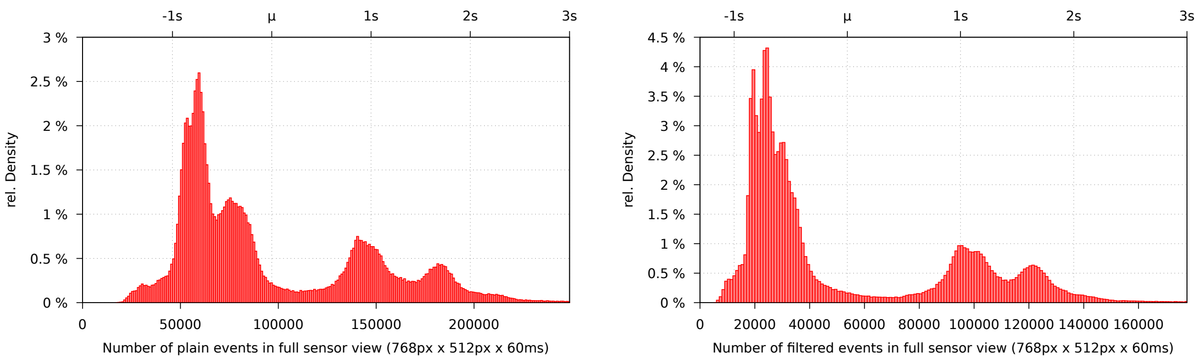
## 4 Detailed per-class F1 scores

Refer to Section 4.6 and 4.7 of main paper

To increase readability and clarity of the evaluation given in the main paper, the label classes were grouped into the three categories background, objects and environmental influences. In addition, due to the very uneven distribution of the events per class, a weighted F1 measure was used. For completeness, the plain F1 scores for all 10 classes are listed in Tables 3 and 4 for all conducted experiments.

For further clarification, qualitative results of the segmentations are given in Figure 5. Here, the processed Patches-Of-Interest are projected into 2D frames and the resulting labeling is represented by false colors.

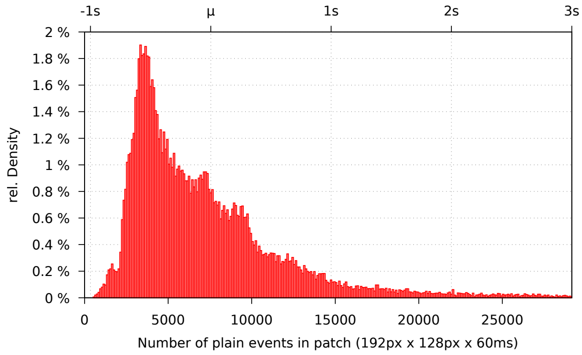
Please refer to the main paper:
Bolten, T., Lentzen, F., Pohle-Fröhlich, R., and Tönnies, K. D.: Evaluation of Deep Learning based 3D-Point-Cloud Processing Techniques for Semantic Segmentation of Neuromorphic Vision Sensor Event-Streams. International Conference on Computer Vision Theory and Applications (VISAPP), 2022



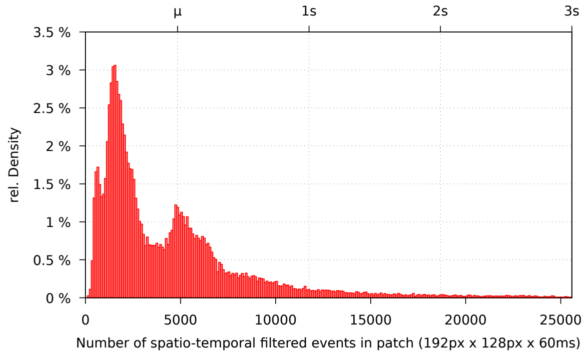

(a) Only Staged scenes, plain, complete sensor view in 60ms

(b) Only staged scenes, time-filtered, complete sensor view in 60ms

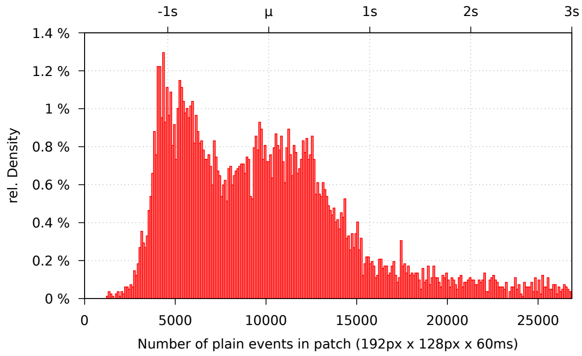
Figure 1: Event count comparison in DVS-OUTLAB database, complete sensor view



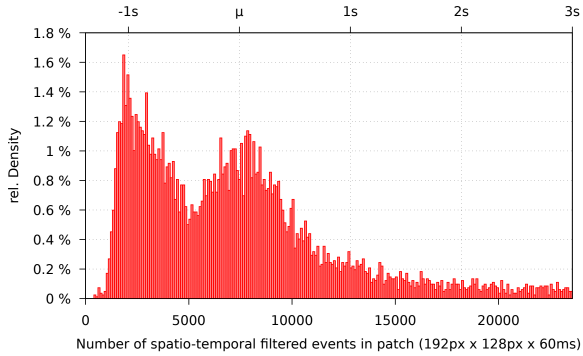

(a) Complete DVS-OUTLAB, plain, patched in 60ms



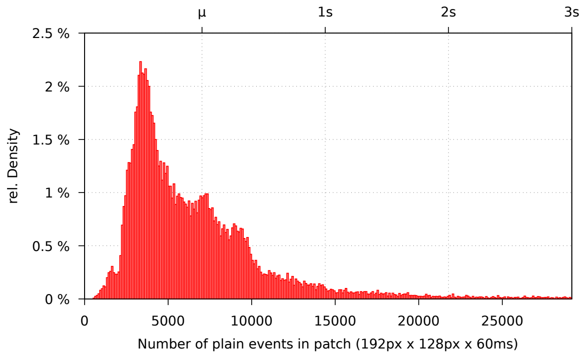

(b) Complete DVS-OUTLAB, time-filtered, patched in 60ms



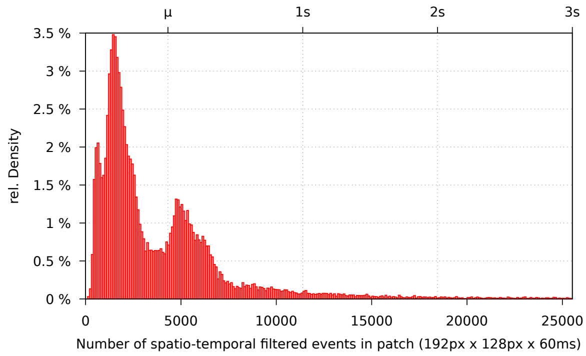

(c) Only Staged scenes, plain, patched in 60ms




(d) Only staged scenes, time-filtered, patched in 60ms

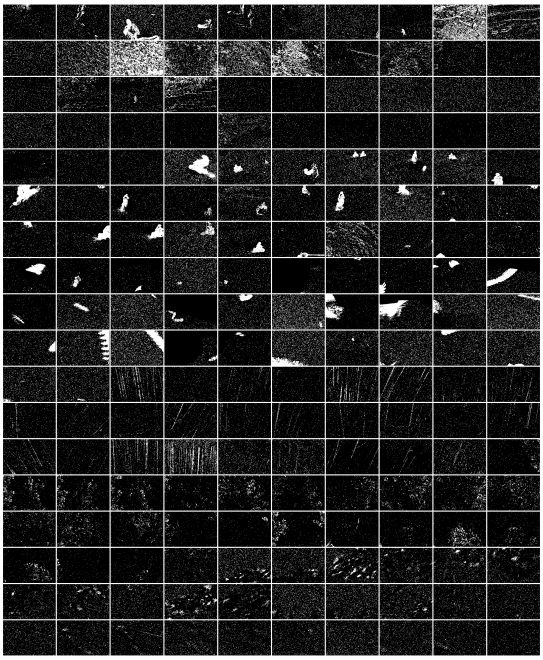



(e) Only environmental influences, plain, patched in 60ms

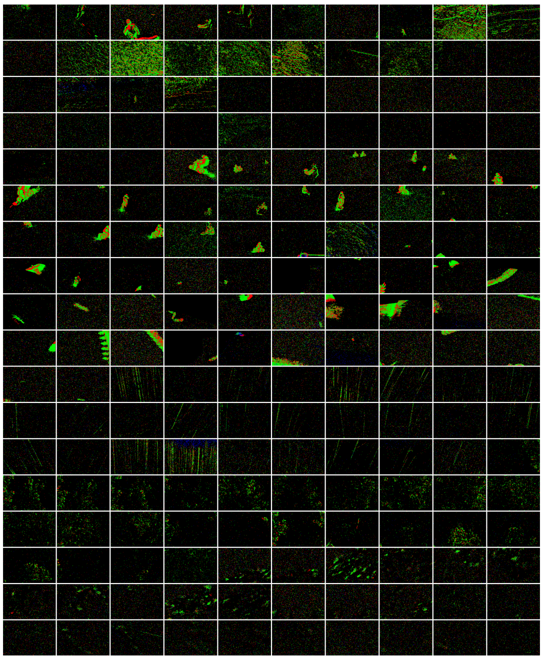



(f) Only environmental influences, time-filtered, patched in 60ms

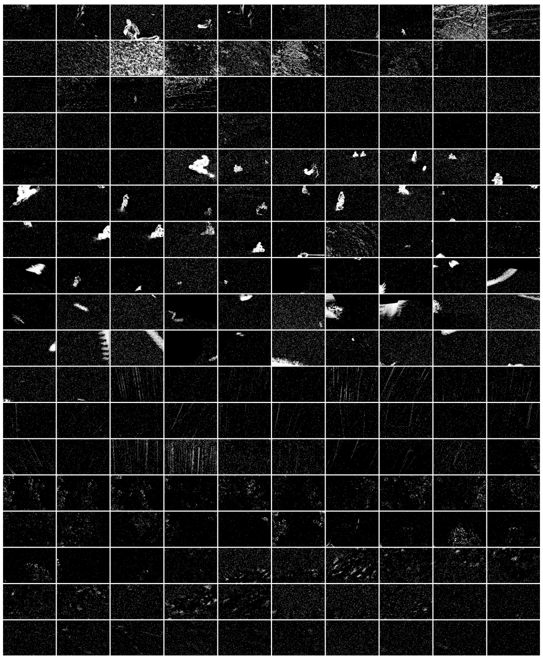
Figure 2: Event count comparison in DVS-OUTLAB database, per patch



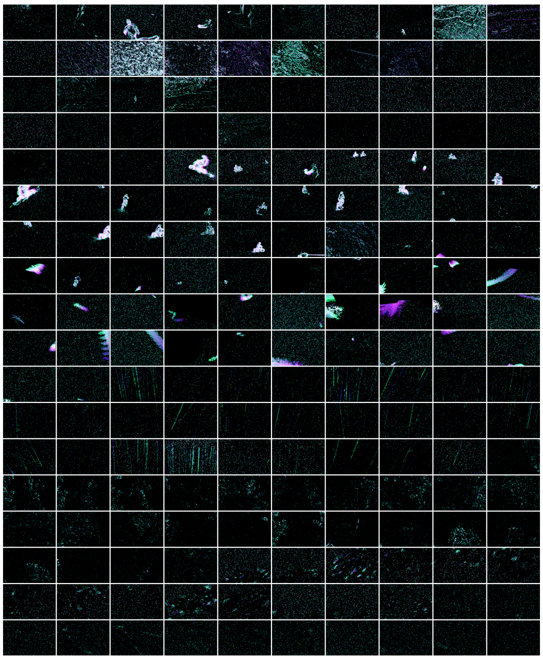
(a) Binary



(b) Polarity



(c) Frequency



(d) MTC

Figure 3: Examples of the different frame representations used, images are randomly selected out of the test-set of DVS-OUTLAB dataset (best viewed in color and digital zoomed)

| | |
|---|---|
| PointNet++(1024, 3L): | $SA(1024,0.1,[32,32,64]) \to SA(256,0.3,[64,64,128]) \to SA(16,0.8,[128,128,256]) \to FP([256,256]) \to FP([256,128]) \to FP([128,128,128,128,10])$ |
| PointNet++(1024, 4L): | $SA(1024,0.1,[32,32,64]) \to SA(256,0.2,[64,64,128]) \to SA(64,0.4,[128,128,256]) \to SA(16,0.8,[256,256,512]) \to FP([256,256]) \to FP([256,256]) \to FP([256,128]) \to FP([128,128,128,128,10])$ |
| PointNet++(1024, 5L): | $SA(1024,0.1,[32,32,64]) \to SA(512,0.2[64,64,128]) \to SA(256,0.4,[128,128,256]) \to SA(64,0.6,[256,256,512]) \to SA(16,0.8,[512,512,1024]) \to FP([256,256]) \to FP([256,256]) \to FP([256,256]) \to FP([256,128]) \to FP([128,128,128,128,10])$ |
| PointNet++(1024, 6L): | $SA(1024,0.1,[32,32,64]) \to SA(512,0.2,[64,64,128]) \to SA(256,0.3,[128,128,256]) \to SA(128,0.4,[256,256,512]) \to SA(64,0.6,[512,512,1024]) \to SA(16,0.8,[1024,1024,2048]) \to FP([256,256]) \to FP([256,256]) \to FP([256,256]) \to FP([256,256]) \to FP([256,128]) \to FP([128,128,128,128,10])$ |

(a) Layer experiments

| | |
|---|---|
| PointNet++(512, 3L): | $SA(512,0.1,[32,32,64]) \to SA(256,0.3,[64,64,128]) \to SA(16,0.8,[128,128,256]) \to FP([256,256]) \to FP([256,128]) \to FP([128,128,128,128,10])$ |
| PointNet++(1024, 3L): | $SA(1024,0.1,[32,32,64]) \to SA(256,0.3,[64,64,128]) \to SA(16,0.8,[128,128,256]) \to FP([256,256]) \to FP([256,128]) \to FP([128,128,128,128,10])$ |
| PointNet++(2048, 3L): | $SA(2048,0.1,[32,32,64]) \to SA(256,0.3,[64,64,128]) \to SA(16,0.8,[128,128,256]) \to FP([256,256]) \to FP([256,128]) \to FP([128,128,128,128,10])$ |
| PointNet++(3072, 3L): | $SA(3072,0.1,[32,32,64]) \to SA(256,0.3,[64,64,128]) \to SA(16,0.8,[128,128,256]) \to FP([256,256]) \to FP([256,128]) \to FP([128,128,128,128,10])$ |

(b) Point-Count experiments

| | |
|---|---|
| $S^T_{\text{cube}}$ PointNet++(2048, 3L): | $SA(2048,0.1,[32,32,64]) \to SA(256,0.3,[64,64,128]) \to SA(16,0.8,[128,128,256]) \to FP([256,256]) \to FP([256,128]) \to FP([128,128,128,128,10])$ |
| $S^T_{\text{tScaled}}$ PointNet++(2048, 3L): | $SA(2048,9.6,[32,32,64]) \to SA(256,28.8,[64,64,128]) \to SA(16,76.8,[128,128,256]) \to FP([256,256]) \to FP([256,128]) \to FP([128,128,128,128,10])$ |
| $S^T_{\text{native}}$ PointNet++(2048, 3L): | $SA(2048,9.6,[32,32,64]) \to SA(256,28.8,[64,64,128]) \to SA(16,76.8,[128,128,256]) \to FP([256,256]) \to FP([256,128]) \to FP([128,128,128,128,10])$ |

(c) Input-scaling experiments

| | |
|---|---|
| Time weight 1 PointNet++(2048, 3L): | $SA(2048,9.6,[32,32,64]) \to SA(256,28.8,[64,64,128]) \to SA(16,76.8,[128,128,256]) \to FP([256,256]) \to FP([256,128]) \to FP([128,128,128,128,10])$ |
| Time weight 3.2 PointNet++(2048, 3L): | Weighted spatio-temporal distance: $d_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + 3.2 \cdot (t_i - t_j)^2}$ |
| Time weight 20 PointNet++(2048, 3L): | Weighted spatio-temporal distance: $d_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + 20 \cdot (t_i - t_j)^2}$ |

(d) Spatio/Temporal weighting experiments

Table 1: PointNet++ Network configurations

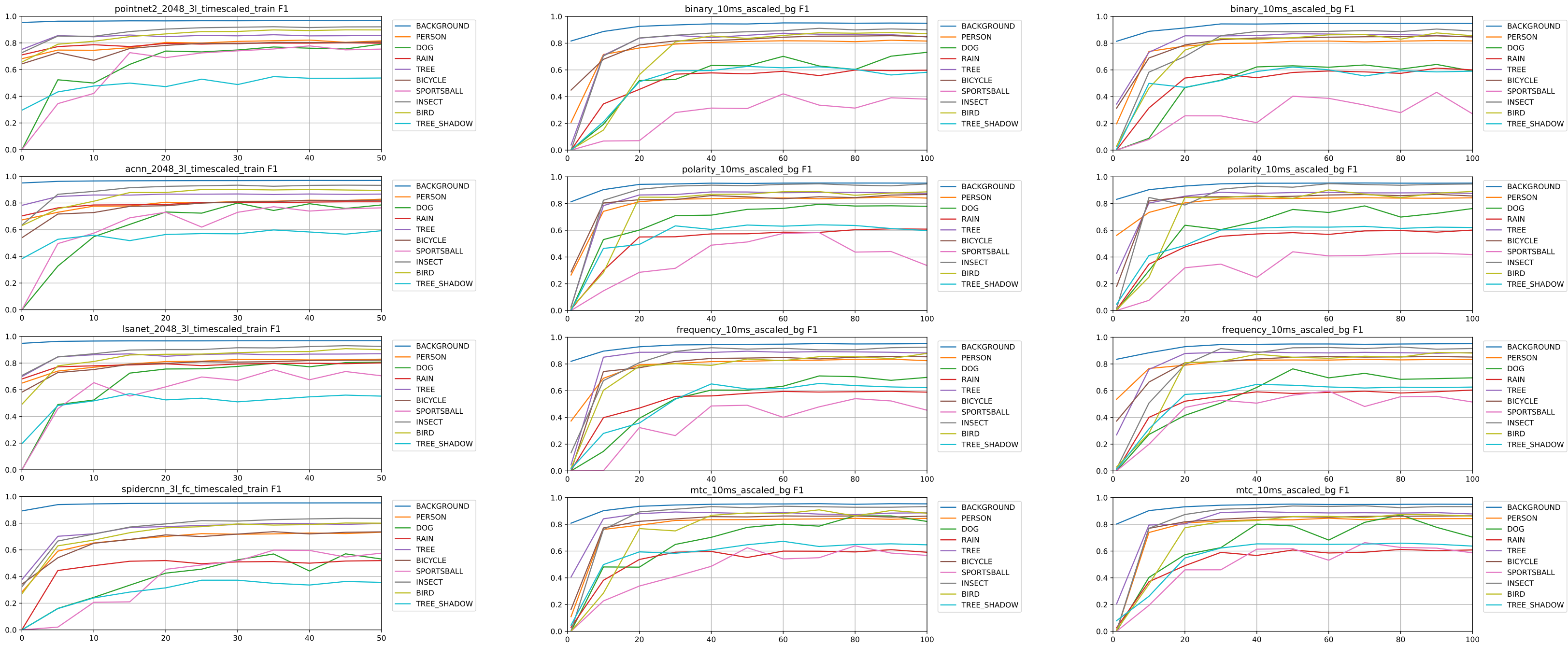
| | |
|---|---|
| $192 \times 128 \times 1$<br>PointNet++(2048, 3L): | $SA(2048, 9.6, [32, 32, 64]) \rightarrow SA(256, 28.8, [64, 64, 128]) \rightarrow SA(16, 76.8, [128, 128, 256]) \rightarrow FP([256, 256]) \rightarrow$<br>$FP([256, 128]) \rightarrow FP([128, 128, 128, 128, 10])$ |

(a) PointNet compare configuration

| | |
|---|---|
| A-CNN(2048, 3L): | $MR(2048, [[0.0, 4.8], [4.8, 9.6]], [[32, 32, 64], [64, 64, 128]]) \rightarrow$<br>$MR(256, [[7.2, 14.4], [21.6, 28.8]], [[64, 64, 128], [128, 128, 256]]) \rightarrow$<br>$MR(16, [[19.2, 38.4], [57.6, 76.8]], [[128, 128, 256], [256, 256, 512]]) \rightarrow$<br>$FP([256, 256]) \rightarrow FP([256, 128]) \rightarrow FP([128, 128, 128, 128, 10])$ |
| LSANet(2048, 3L): | $LSA(2048, 9.6, [32, 32, 64], [32, 32]) \rightarrow LSA(256, 28.8, [64, 64, 128], [32, 32]) \rightarrow LSA(16, 76.8, [128, 128, 256], [32, 32]) \rightarrow$<br>$FP([256, 256]) \rightarrow FP([256, 128]) \rightarrow FP([128, 128, 128, 128, 10])$ |
| SpiderCNN(*, 3L): | $BallQuery(9.6) \rightarrow SpiderConv(32) \rightarrow SpiderConv(64) \rightarrow SpiderConv(128) \rightarrow Top\text{-}2 \rightarrow FC([256, 256, 128, 10])$ |

(b) Network successor configurations

Table 2: Network-Variant configurations




(a) 3D point cloud methods

(b) Mask-R-CNN @ ResNet50-Backbone, evaluated every 10th epoch

(c) Mask-R-CNN @ ResNet101-Backbone, evaluated every 10th epoch

Figure 4: Per-class F1 score in network trainings processes

| Config | BACKGROUND | PERSON | DOG | RAIN | TREE | BICYCLE | SPORTSBALL | INSECT | BIRD | TREE_SHADOW | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 Layer | 0.95 | **0.76** | **0.59** | **0.8** | 0.7 | **0.75** | 0.41 | 0.88 | 0.83 | 0.28 | **0.7** |
| 4 Layer | 0.95 | 0.74 | 0.53 | **0.8** | 0.73 | 0.73 | 0.49 | **0.89** | 0.83 | 0.22 | 0.69 |
| 5 Layer | 0.95 | 0.75 | 0.56 | **0.8** | 0.76 | **0.75** | **0.5** | **0.89** | **0.84** | 0.22 | **0.7** |
| 6 Layer | **0.96** | 0.75 | 0.52 | **0.8** | **0.77** | 0.72 | 0.36 | **0.89** | 0.81 | **0.31** | 0.69 |

(a) Layer experiments

| Config | BACKGROUND | PERSON | DOG | RAIN | TREE | BICYCLE | SPORTSBALL | INSECT | BIRD | TREE_SHADOW | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|
| n = 512 | 0.95 | 0.74 | 0.49 | 0.71 | 0.7 | 0.7 | 0.2 | 0.86 | 0.82 | 0.23 | 0.64 |
| n = 1024 | 0.95 | **0.76** | 0.59 | 0.8 | 0.7 | **0.75** | 0.41 | 0.88 | 0.83 | 0.28 | 0.7 |
| n = 2048 | **0.96** | **0.76** | 0.6 | **0.86** | 0.79 | **0.75** | **0.6** | **0.9** | **0.84** | **0.37** | 0.74 |
| n = 3072 | **0.96** | **0.76** | **0.64** | **0.86** | **0.8** | **0.75** | **0.6** | 0.89 | **0.84** | 0.35 | **0.75** |

(b) Point-Count experiments

| Config | BACKGROUND | PERSON | DOG | RAIN | TREE | BICYCLE | SPORTSBALL | INSECT | BIRD | TREE_SHADOW | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $S^T_{\text{cube}}$ | 0.96 | 0.76 | 0.6 | **0.86** | 0.79 | 0.75 | 0.6 | 0.9 | 0.84 | 0.37 | 0.74 |
| $S^T_{\text{tScaled}}$ | **0.97** | **0.83** | **0.78** | 0.8 | 0.86 | **0.82** | **0.65** | 0.92 | 0.88 | 0.52 | **0.8** |
| $S^T_{\text{native}}$ | **0.97** | 0.82 | 0.73 | 0.85 | **0.87** | 0.8 | 0.63 | **0.93** | **0.89** | **0.53** | **0.8** |

(c) Input-scaling experiments

| Config | BACKGROUND | PERSON | DOG | RAIN | TREE | BICYCLE | SPORTSBALL | INSECT | BIRD | TREE_SHADOW | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Time Weight 1 | **0.97** | **0.82** | **0.73** | 0.85 | **0.87** | 0.8 | **0.63** | **0.93** | 0.89 | **0.53** | **0.8** |
| Time Weight 3.2 | **0.97** | **0.82** | 0.69 | **0.86** | 0.86 | **0.81** | **0.63** | **0.93** | **0.9** | 0.49 | 0.79 |
| Time Weight 20 | 0.96 | 0.8 | 0.7 | **0.86** | 0.81 | 0.79 | 0.6 | 0.91 | 0.86 | 0.31 | 0.76 |

(d) Spatio/Temporal weighting experiments

Table 3: Detailed per-class F1-Scores in PointNet++ parameter optimization experiments

| Config | BACKGROUND | PERSON | DOG | RAIN | TREE | BICYCLE | SPORTSBALL | INSECT | BIRD | TREE_SHADOW | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|
| PointNet++ | **0.97** | **0.83** | 0.78 | 0.8 | 0.86 | **0.82** | **0.65** | 0.92 | 0.88 | 0.52 | **0.8** |
| A-CNN | **0.97** | **0.83** | 0.77 | **0.81** | **0.87** | **0.82** | 0.58 | **0.93** | **0.89** | **0.58** | **0.8** |
| LSANet | **0.97** | **0.83** | **0.82** | **0.81** | **0.87** | **0.82** | 0.57 | 0.92 | **0.89** | 0.54 | **0.8** |
| SpiderCNN | 0.95 | 0.74 | 0.51 | 0.53 | 0.8 | 0.73 | 0.41 | 0.83 | 0.76 | 0.32 | 0.66 |

(a) 3D Network variants

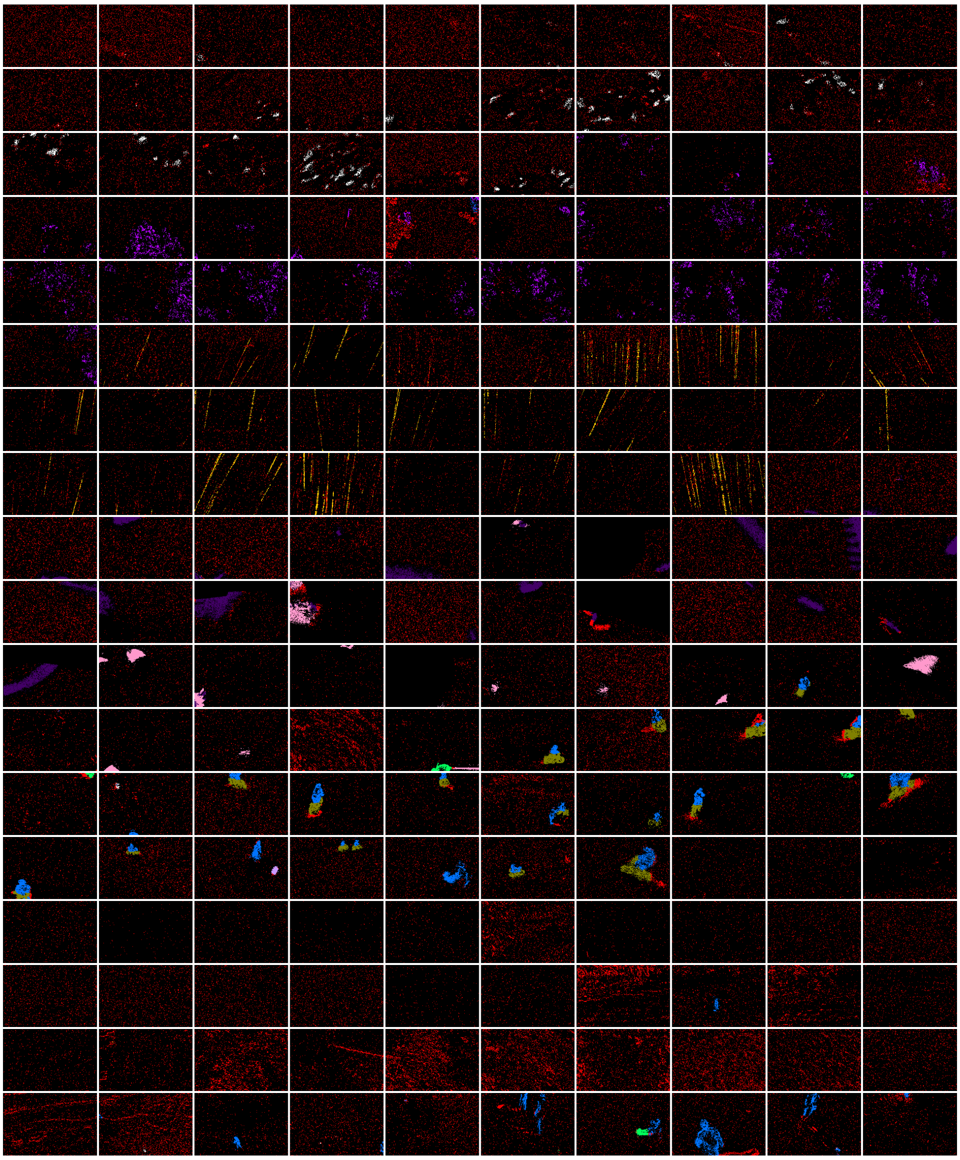
| Config | BACKGROUND | PERSON | DOG | RAIN | TREE | BICYCLE | SPORTSBALL | INSECT | BIRD | TREE_SHADOW | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Binary | **0.95** | 0.82 | 0.73 | 0.6 | 0.85 | 0.85 | 0.38 | 0.9 | 0.87 | 0.58 | 0.75 |
| Polarity | **0.95** | **0.84** | 0.78 | **0.61** | 0.88 | **0.87** | 0.34 | **0.95** | **0.89** | 0.6 | 0.77 |
| Frequency | **0.95** | 0.82 | 0.7 | 0.59 | 0.88 | 0.85 | 0.45 | 0.93 | 0.88 | 0.62 | 0.77 |
| MTC | **0.95** | **0.84** | **0.82** | 0.59 | **0.89** | 0.86 | **0.57** | 0.93 | 0.88 | **0.65** | **0.8** |

(b) Mask-R-CNN @ ResNet50-Backbone

| Config | BACKGROUND | PERSON | DOG | RAIN | TREE | BICYCLE | SPORTSBALL | INSECT | BIRD | TREE_SHADOW | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Binary | **0.95** | 0.82 | 0.59 | 0.6 | 0.85 | 0.85 | 0.27 | 0.89 | 0.86 | 0.59 | 0.73 |
| Polarity | **0.95** | **0.84** | **0.76** | 0.6 | 0.88 | **0.86** | 0.42 | **0.95** | **0.89** | 0.62 | 0.78 |
| Frequency | **0.95** | 0.83 | 0.7 | 0.6 | **0.89** | 0.85 | 0.52 | 0.92 | 0.88 | 0.63 | 0.78 |
| MTC | **0.95** | **0.84** | 0.7 | **0.61** | 0.88 | **0.86** | **0.59** | 0.93 | 0.86 | **0.64** | **0.79** |

(c) Mask-R-CNN @ ResNet101-Backbone

Table 4: Detailed per-class F1-Scores for network variants and 2D-Mask-R-CNN baseline



(a) Labeled Patches-Of-Interest projected into 2D frames

BACKGROUND PERSON DOG BICYCLE SPORTSBALL RAIN INSECT BIRD TREE TREE SHADOW

(b) False-color legend

Figure 5: False-color examples for PointNet++ semantic segmentation results

## REFERENCES


Abdulla, W. (2017). Mask R-CNN for object detection and instance segmentation on keras and tensorflow. https://github.com/matterport/Mask_RCNN.

Bolten, T., Pohle-Fröhlich, R., and Tönnies, K. D. (2021). DVS-OUTLAB: A neuromorphic event-based long time monitoring dataset for real-world outdoor scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1348–1357.

Chen, L., Li, X., Fan, D., Cheng, M., Wang, K., and Lu, S. (2019). LSANet: Feature learning on point sets by local spatial aware layer. *arXiv*, abs/1905.05442.

He, K., Gkioxari, G., Dollar, P., and Girshick, R. (2017). Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Komarichev, A., Zhong, Z., and Hua, J. (2019). A-CNN: Annularly convolutional neural networks on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7421–7430.

Qi, C. R., Yi, L., Su, H., and Guibas, L. J. (2017). PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 5105–5114, Red Hook, NY, USA. Curran Associates Inc.

Xu, Y., Fan, T., Xu, M., Zeng, L., and Qiao, Y. (2018). SpiderCNN: Deep learning on point sets with parameterized convolutional filters. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., editors, *Computer Vision – ECCV 2018*, pages 90–105, Cham. Springer International Publishing.