Masked Autoencoder for Self-Supervised Pre-training on Lidar Point Clouds - Supplementary material

Georg Hess[1,2] Johan Jaxing[1] Elias Svensson[1] David Hagerman[1]

Christoffer Petersson[1,2] Lennart Svensson[1]

[1]Chalmers University of Technology [2]Zenseact

georghe@chalmers.se

1. Baseline hyperparameters

In this section, we present hyperparameters for the 3D OD model. For training the detector, we use the same loss functions as in the original SST implementation [1], but modify hyperparameters for the nuScenes dataset. Unless stated otherwise, the same set of parameters are used for pre-training, e.g., the voxelization parameters in Table 1, the voxel encoder in Table 2, and the SST encoder in Table 3. Table 4 specifies parameters used for downstream task training only.

Parameter	Value
Voxel size (m)	$0.5 \times 0.5 \times 8$
Max #point	∞
Max #points/voxel	∞
Max #voxels	∞
Point cloud range - x	[-50 m, 50 m]
Point cloud range - y	[-50 m, 50 m]
Point cloud range - z	[-3 m, 5 m]
Voxel grid shape (x,y,z)	(200,200,1)

Table 1: Parameters used for voxelization.

Parameter	Value
First linear layer	64 output channels
Second linear layer	128 output channels

Table 2: Parameters used for voxel encoder.

2. Pre-training hyperparameters

In this section we present hyperparameters used for the decoder and reconstruction head using during pre-training, see Tables 5 and 6.

Parameter	Value
Window size	16×16
Padding levels (train)	[30, 60, 100, 200, 250]
Padding levels (test)	[30, 60, 100, 200, 256]
#Layers	8
Input dimension	128
FFN hidden dimension	256
#Heads	8
#Attached conv. layers	3
Conv. kernel size	3×3
Conv. stride	1
Conv. padding (per layer)	(1,1,2)
Conv. in/out channels	128
Linear projection dim	384

Table 3: Parameters used for SST encoder. Note that the convolution layers are not used during pre-training. The padding levels refer to the grouping of windows when passing them through an encoder block, which allows for more efficient computations.

3. Results with two sweeps

3.1. Data efficiency

Table 8 shows the data efficiency results when pre-training and fine-tuning were done with two aggregated sweeps. Similar to the results for 10 sweeps in Table 1, we see that our Voxel-MAE brings a substantial performance increase compared to the fully supervised baseline. The baseline reaches 43.6 mAP and 55.19 NDS when using the entire training dataset. The model pre-trained with Voxel-MAE outperforms this baseline when using only 60% of the annotated data with 43.77 mAP and 55.29 NDS.

Same as for the experiments with 10 sweeps the pre-trained models consistently improve upon their baseline in

Parameter	Value
Class loss	FocalLoss($\gamma = 2.0$,$\alpha = 0.25$)
Bounding box loss	SmoothL1Loss($\beta = 1/9$)
Direction loss	CrossEntropyLoss
Bounding box target weight $x, y, z, w, l, h, \theta, v_x, v_y$	1,1,1,1,1,1,1,0.1,0.1
IOU class assignment threshold	0.6
IOU background assignment threshold	0.3
max NMS evaluations	1,000
NMS IOU threshold	0.2
score threshold	0.05
min bbox size	0
max NMS predictions	500
$(\beta_{loc}, \beta_{cls}, \beta_{dir})$	(1,1,0.2)

Table 4: Parameters used for the detection head. NMS stands for non-maximum suppression and is used during evaluation to filter predictions.

Parameter	Value
Window size	16×16
Padding levels (train)	[30, 60, 100, 200, 250]
Padding levels (test)	[30, 60, 100, 200, 256]
#Blocks	8
Input dimension	128
FFN hidden dimension	256
#Heads	8
#Empty voxels	0.1·#voxels

Table 5: Parameters for the decoder used during pre-training. The padding levels refer to the grouping of windows when passing them through a decoder block, which allows for more efficient computations.

Parameter	Value
Empty voxel loss	BinaryCrossEntropy
Number of points loss	SmoothL1($\beta = 1$)
#Predicted points (Chamfer)	10
#Max GT points (Chamfer)	100
α_c	1
α_{np}	1
α_{occ}	1

Table 6: Parameters for the reconstruction head used during pre-training.

Voxel size (m)	Encoder depth	mAP	NDS
0.25	6	31.43	50.76
0.30	6	35.93	52.60
0.50	6	42.79	**55.54**
0.50	8	**43.60**	55.19
0.70	6	41.73	54.69
0.70	8	31.31	54.21

Table 7: Performance on the nuScenes validation dataset for a model using 2 sweeps and without any pre-training.

terms of mAP and NDS regardless of dataset fraction. Further, the largest improvements can be found for models fine-tuned on 20% of the annotations, indicating the effectiveness of our method when the amount of unlabeled data is large compared to the annotated one. However, also when using all available annotations, pre-training can increase detection performance.

3.2. Encoder depth and voxel size

In Table 7 we study how baseline performance varies with different number of encoder layers and voxel size. For reference, the original SST model was tuned toward the Waymo Open dataset and used 6 encoder layers and a voxel size of $0.32 \times 0.32 \times 6$ m. However, for nuScenes, we found better performance with 8 encoder layers and a voxel size of $0.5 \times 0.5 \times 8$ m.

References

[1] Lue Fan, Ziqi Pang, Tianyuan Zhang, Yu-Xiong Wang, Hang Zhao, Feng Wang, Naiyan Wang, and Zhaoxiang Zhang. Embracing single stride 3d object detector with sparse transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8458–8468, June 2022.

Dataset fraction	Pre-trained	mAP	NDS	ped.	car	truck	bus	barrier	T.C.	trailer	moto.
0.2	✗	35.54	47.79	62.5	73.6	35.8	41.7	49.4	31.2	14.8	29.5
	✓	**39.95**	**51.60**	**69.1**	**75.7**	**40.4**	**48.1**	**54.6**	**39.3**	**18.2**	**33.6**
0.4	✗	38.99	51.41	66.7	76.8	39.8	48.2	51.9	34.9	17.9	33.4
	✓	**43.15**	**54.46**	**71.4**	**77.5**	**43.0**	**54.2**	**57.5**	**41.0**	**20.7**	**38.4**
0.6	✗	41.29	53.28	69.1	77.4	41.3	50.8	54.9	37.4	20.0	38.7
	✓	**43.77**	**55.29**	**72.1**	**77.8**	**44.0**	**54.5**	**57.4**	**43.8**	**21.8**	**40.7**
0.8	✗	42.26	54.24	69.2	**77.8**	40.8	53.6	55.4	40.2	19.4	39.6
	✓	**43.96**	**55.57**	**72.2**	**77.8**	**43.0**	**55.0**	**57.9**	**42.6**	**22.1**	**41.7**
1.0	✗	43.60	55.19	69.9	**78.9**	**43.1**	**55.7**	56.7	39.5	**21.4**	41.5
	✓	**44.62**	**56.00**	**72.1**	78.3	43.0	54.5	**57.3**	**43.2**	21.2	**44.9**

Table 8: mAP, NDS, and AP per class on the nuScenes validation data for pre-trained and randomly initialized models when varying the amount of *labeled* data. Pre-training and fine-tuning is done with *two* aggregated point cloud sweeps without intensity information. ped.=pedestrian. T.C.=traffic cone. moto.=motorcycle.