UniPAD: A Universal Pre-training Paradigm for Autonomous Driving

Supplementary Material

1. Experimental Settings

Feature Projection. For the fine-tuning model, we follow [4] to use an extra three conv-layers for feature projection, while the pre-training model employs another single conv-layer for projection. To share the projection parameters, we use the same three conv-layers for both the pre-training and fine-tuning models.

Point-based Pre-training. To compare with occupancy-based self-supervised learning, we integrate ALSO [2] into our framework, utilizing volumetric inputs instead of Bird's Eye View (BEV) representation. We sample 4096 query points for each scene, comprising three types: randomly sampled empty query points between the LiDAR sensor and surface points, empty query points positioned just before surface points within a threshold of 0.1, and full query points located behind surface points within a threshold of 0.1. Subsequently, the sampled query points serve as occupancy targets (empty for 0 and full for 1) for supervising voxels within a radius of 1.0. We employ 4 linear layers with a hidden dimension of 64 on each voxel to predict occupancy. The binary cross-entropy loss is used for the estimated occupancy and targets during the training process.

For the MAE-based model, we adopt block-wise masking with a masking size of 8 to selectively remove portions of the input point clouds. After the encoder, masked regions are padded with zeros and combined with visible features to form regular dense voxel features. Finally, a 3D conv-layer with a kernel size of 3 is applied to predict the local coordinates of point clouds in the masked voxels. The reconstruction loss is computed using the L2 Chamfer Distance, following the approach in [10].

For the contrastive-based model, we follow [7] to contrast image and point features to integrate 2D knowledge into 3D points. The ConvNeXt-small [8] and VoxelNet [9] are adopted as the image encoder and the point encoder, respectively. We project point clouds onto the multi-view images, sampling 512 points per image. Bilinear interpolation is employed to extract image features from the projected point cloud coordinates, while trilinear interpolation is used to fetch point features from voxel features based on the point cloud positions. Separate linear layers are applied to project the features accordingly. Paired pixel-point features are considered positives, while others are treated as negatives. We apply InfoNCE loss with a temperature of 0.07 to encourage the alignment of positives and discourage negatives.

Different View Transformations. We investigate various view transformation strategies and make minor adjustments to seamlessly integrate them into our volumetric representations. In the case of BEVformer [6], we randomly initialize voxel-based queries and use a learned positional embedding [1] based on the voxel coordinates. The voxel queries are projected onto multi-view images, and image features are incorporated into the queries through deformable attention [11]. Given that one voxel query may project onto different view images, we employ average pooling on the output features to handle this scenario. For both BEVDet [3] and BEVDepth [5], image features are projected onto 3D space and pooled to obtain voxel features. In the view transformation of BEVDepth, we additionally apply depth supervision based on the depth of projected point clouds in the image coordinates. Specifically, we treat depth estimation as a classification task by discretizing depth into several bins and use binary cross-entropy as the loss function for supervision.

2. Supplementary Experiments

Table 1 shows supplementary ablation studies of the effectiveness of the masking, sampling, and loss. The first and fourth rows demonstrate a 2.5 NDS improvement over the baseline through our rendering-based pre-training, even without additional point clouds as input. In the second and third rows, as well as the fifth and sixth rows, the effectiveness of masking in enhancing representation learning during pre-training is demonstrated. The third and fourth rows illustrate the efficacy of the proposed depth-aware sampling. In the third and sixth rows, it's emphasized that integrating additional information, such as depth for geometric constraints, further improves the learned features.

3. Qualitative Results

In Figure 1, we present 3D detection results in camera space and BEV (Bird's Eye View) space with LiDAR point clouds. Our model can predict accurate bounding boxes for nearby objects and also shows the capability of detecting objects from far distances.

References

[1] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1

Table 1. Ablation studies of the masking, sampling, and loss.

Masking	Sampling		Loss		NDS
	Random	Depth-aware	RGB	Depth	
					25.2
		✓	✓		26.7
✓		✓	✓		27.9
✓	✓		✓		27.4
		✓	✓	✓	31.7
✓		✓	✓	✓	**32.9**

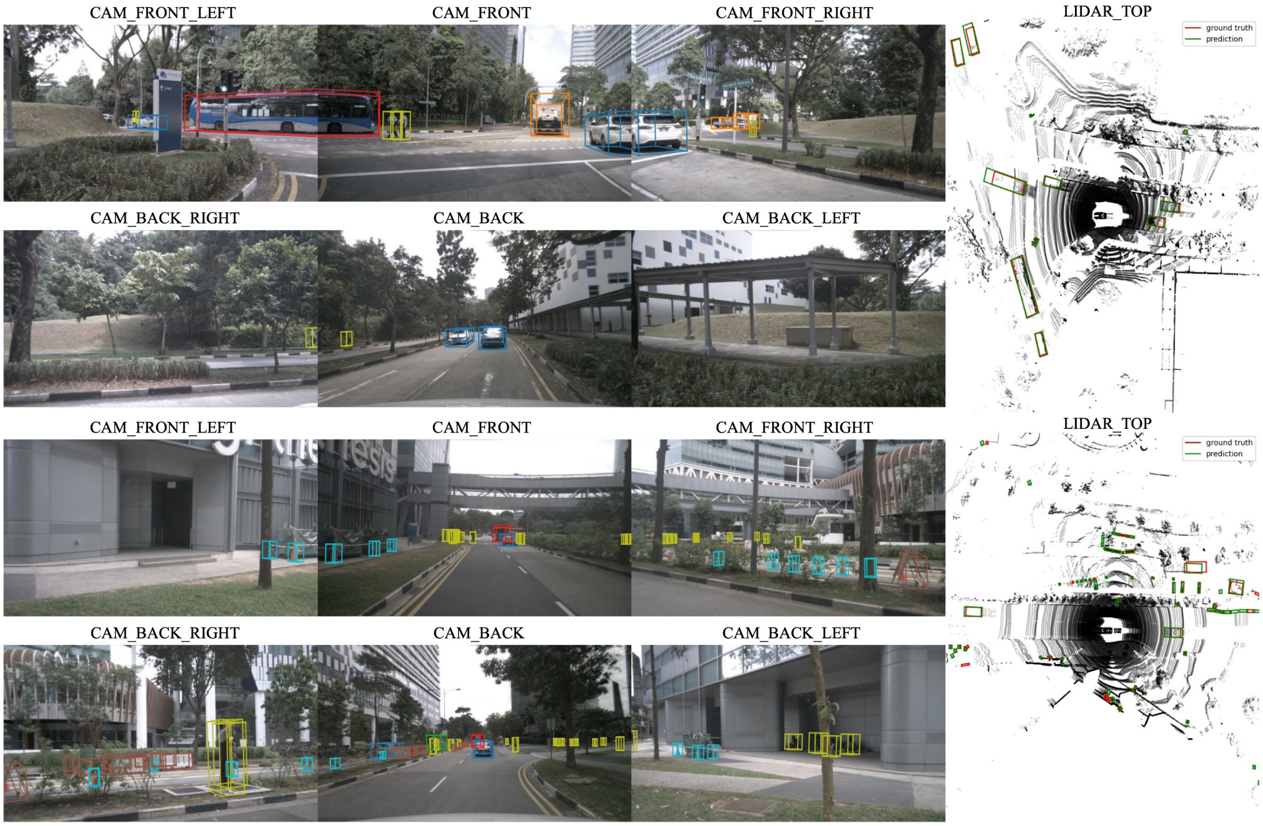

Figure 1. Illustration of the detection results. The predictions are shown on multi-view images and bird's eye view with LiDAR points.

[2] Alexandre Boulch, Corentin Sautier, Björn Michele, Gilles Puy, and Renaud Marlet. ALSO: automotive lidar self-supervision by occupancy estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 1

[3] Junjie Huang, Guan Huang, Zheng Zhu, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *CoRR*, abs/2112.11790, 2021. 1

[4] Yanwei Li, Yilun Chen, Xiaojuan Qi, Zeming Li, Jian Sun, and Jiaya Jia. Unifying voxel-based representation with transformer for 3d object detection. In *Advances in Neural Information Processing Systems*, 2022. 1

[5] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. 1

[6] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *Proceedings of the European Conference on Computer Vision*, 2022. 1

[7] Yueh-Cheng Liu, Yu-Kai Huang, HungYueh Chiang, Hung-Ting Su, Zhe Yu Liu, Chin-Tang Chen, Ching-Yu Tseng, and Winston H. Hsu. Learning from 2d: Pixel-to-point knowledge transfer for 3d pretraining. *CoRR*, abs/2104.04687, 2021. 1

[8] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1

[9] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embed-

ded convolutional detection. *Sensors*, 18(10), 2018. 1

[10] Honghui Yang, Tong He, Jiaheng Liu, Hua Chen, Boxi Wu, Binbin Lin, Xiaofei He, and Wanli Ouyang. GD-MAE: generative decoder for MAE pre-training on lidar point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 1

[11] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021. 1