Appendix

Contents

The following items are included in the supplementary material:

- Results on the test split, computation comparison of different fusion modules and additional ablation study in Section A.
- Pre-training configuration in Section B.
- Fine-tuning architecture and configuration in Section C.
- Reconstruction visualization results in Section D.

A Additional Experiments

A.1 Additional Results on the Nuscenes Test Split

The 3D object detection results on test set of the nuScenes are reported in Table A. In the multi-modal setting, our UniM^{2}AE boosts the BEVFusion [12] by 0.4 NDS and achieve competitive results compared with the SOTA multi-modal detectors. For the detectors that are solely single-modal, our LiDAR-only UniM^{2}AE-L outperforms the baseline model TransFusion-L [1] by 2.4/2.0 mAP/NDS improvement, indicating the generalization of our self-supervised methods. Concerning that our MAE framework isn't specifically designed for the LiDAR-only detector, the UniM^{2}AE lags slightly behind GeoMAE [13], which introduces extra loss functions for the characteristics of the point cloud.

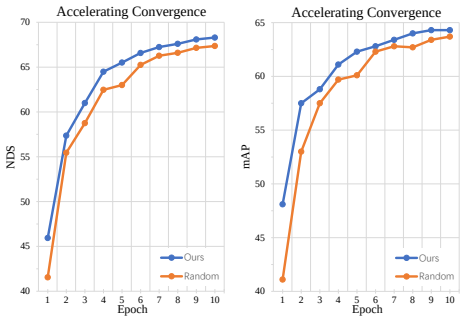

Fig. A: 3D object detection results on the nuScenes validation split. Our UniM^{2}AE accerlerates the model convergence and ultimately improve the performance.

A.2 Additional Abalation Study

As shown in Fig. A, we compare the performance of detectors trained from scratch and pre-trained with our UniM^{2}AE for 10 epochs. Our pre-training method significantly accelerates model convergence and finally stabilises it at a higher score when utilizing the entire dataset.

Table A: Performance of the 3D object detection on the nuScenes dataset test split. † means applying the pre-trained MMIM to the downstream task.

Method	Modality	mAP	NDS
PointPillar [8]	L	40.1	55.0
CenterPoint [15]	L	60.3	67.3
VoxelNeXt [3]	L	64.5	70.0
LargeKernel3D [2]	L	65.4	70.6
GeoMAE [13]	L	67.8	**72.5**
TransFusion-L [1]	L	65.5	70.2
UniM^{2}AE-L	L	**67.9**	72.2
UVTR-M [9]	C+L	67.1	71.1
TransFusion [1]	C+L	68.9	71.7
VFF [10]	C+L	68.4	72.4
DSVT [14]	C+L	68.4	72.7
BEVFusion [12]	C+L	70.2	72.9
UniM^{2}AE†	C+L	**70.3**	**73.3**

Table B: Computation comparison of fusion modules

Fusion modules	Voxel Size(m)	#param	FLOPs	mAP↑
Convolution	(0.15, 0.15, 8)	0.24M	124.5G	68.2
MSMD	(0.075, 0.075, 2)	5.61M	211.5G	69.3
MMIM	(0.15, 0.15, 4)	0.36M	191.1G	69.7

A.3 Extra Computation in MMIM

Table B lists the mAP, FLOPs and module size(#param) for fusion modules in BEVFusion-SST(Convolution) [12], MSMDFusion(MSMD) [7] and UniM^{2}AE†. Compared to the convolution-based fusion module, MMIM does not double the FLOPs accordingly, although the number of input voxels is twice as large, while there is a large performance improvement. Meanwhile, MMIM achieves better performance with less computational cost and fewer parameters compared to MSMD, which validates the effectiveness of MMIM.

B Pre-training Details

To fairly compare UniM^{2}AE with the single-modal MAE methods (*i.e.* GreenMIM [6] and Voxel-MAE [5]), a consistent pre-training configuration shown in Table C is adapted during the pre-training process, where point cloud is abbreviated as PC. The detailed hyperparameters of MAE methods we used in this work are as follows.

Table C: Pre-training Configuration

Config	Value
Optimizer	AdamW
Base lr	5e-4
Weight decay	0.001
Batch size	32
Lr schedule	cosine annealing
Warmup iterations	1000
PC augmentation	random flip, resize
Image augmentation	crop, resize, random flip
Total epochs	200

Table D: Fine-tuning Configuration

Config	Value	
	Detection	Segmentation
PC range -x	[-54.0m, 54.0m]	[-51.2m, 51.2m]
PC range -y	[-54.0m, 54.0m]	[-51.2m, 51.2m]
PC range -z	[-3.0m, 5.0m]	[-3.0.m, 5.0m]
Optimizer	AdamW	AdamW
Base lr	1e-4	1e-4
Weight decay	0.01	0.01
Batch size	4	4

B.1 UniM^{2}AE Hyperparameters

Generally, we employ the configuration of the encoder and decoder presented in GreenMIM [6] and Voxel-MAE [5] with adaptive modification to better suit multi-modal self-supervised pre-training. The image size is set to [256, 704] and the point cloud range is restrict in [-50m, 50m] for x-, y-axes, [-3m, 5m] for z-axes. At the same time, the volume grid shape is set to [200, 200, 2]. Specifically, to align multi-view images and LiDAR point cloud, only random flipping, resizing and cropping are used in the image augmentation, discarding other data augmentation methods originally applied to Masked Image Modeling.

For the Spatial Cross-Attention during the image to 3D volume projection, the number of deformable attention blocks is set to 6 with 256 hidden channels and $\mathcal{N}_{ref}$ is set to 4. In the Multi-modal Interaction Module (MMIM), we stack 3 deformable self-attention blocks comprising 8 heads in each block and the number of reference point is 4.

B.2 Baseline Hyperparameters

In the experiments on data efficiency, we compare our UniM^{2}AE with single-modal MAE methods, whose implementation follow their publicly released codes with minimal changes. Since GreenMIM [6] doesn't employ Swin-T [11] as their backbone while pre-training, we replaced their original Swin-B with Swin-T [11], and rigorously follow the other settings for pre-training. Additionally, for fair comparison, we use the same data augmentation in the camera branch of UniM^{2}AE. For the Voxel-MAE [5], pre-training are done with intensity information in data efficiency experiment. The rest of the setup is the same as the Voxel-MAE [5].

C Fine-tuning Details

In the pre-training phase, we first adhere to the process shown in Fig. 2 to obtain the pre-trained backbone and MMIM. We then evaluate our multi-modal self-supervised pre-training framework by fine-tuning two state-of-the-art detectors [1, 12], denoted as TransFusion-L-SST and BEVFusion-SST, whose LiDAR backbones are replaced by SST [4], as illustrated in Fig. B. Detail configuration is presented in Table D, where point cloud range is abbreviated as PC range.

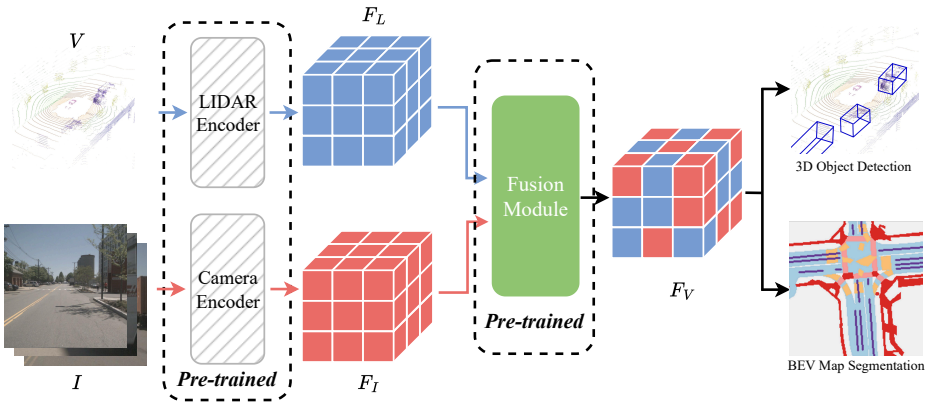

Fig. B: Fine-tuning overview. The LiDAR and camera backbone are initialized with the weights pre-trained by UniM^{2}AE in the finetuning phase. As for UniM^{2}AE†, the fusion module is additionally replaced with our pre-trained MMIM. Depending on the downstream task, different head as well as training settings are adopted accordingly.

In the 3D object detection task(denoted as Detection), we separately set the voxel size to [0.5m, 0.5m, 8m] in the LiDAR-only method and [0.5m, 0.5m, 4m] in the multi-modal method. During the pre-training we first transfer the weights of UniM^{2}AE LiDAR encoder to TransFusion-L-SST and fintune it. We follow the TransFusion [1] training schedule and the results obtained by fine-tuning is denoted as UniM^{2}AE-L. As for the multi-modal strategies, the weights of LiDAR encoder in UniM^{2}AE-L and camera encoder pre-trained by UniM^{2}AE are loaded to finetune the BEVFusion-SST following the BEVFusion [12] training schedule. Furthermore, we replace the fusion module in BEVFusion-SST by the pre-trained MMIM and get UniM^{2}AE†.

In the BEV map segmentation task(denoted as Segmentation in Table D), unlike the previous two-stage training schedule in the 3D object detection, we directly fine-tune the multi-modal BEVFusion-SST with [0.2m, 0.2m, 4m] voxel size for 24 epochs and the camera-only BEVFusion [12] for 20 epochs, The changes regarding the backbone and fusion modules are the same as for the 3D detection task. For the camera-only detectors, all configuration is aligned with camera-only BEVFusion [12].

D Visualization

In Fig. C, we provide examples of reconstruction visualizations. Our UniM^{2}AE is able to reconstruct the masked LiDAR point clouds and corresponding multi-view images, accurately reflecting semantic and geometric understanding. For ease of observation, part of the point cloud(framed by a red box) is zoomed in and shown below.

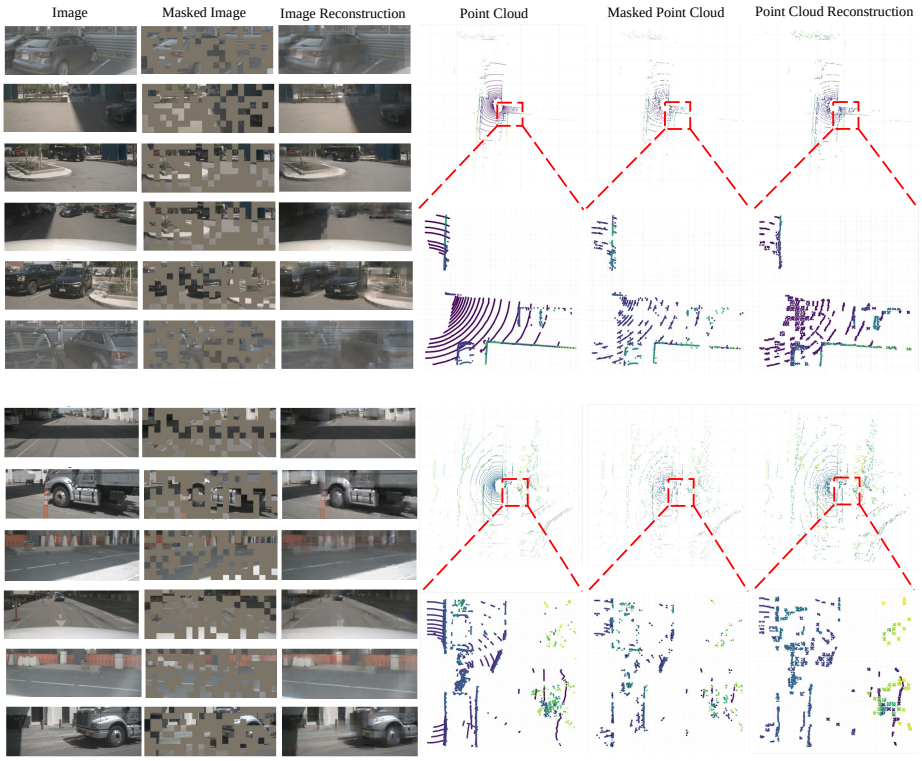

Fig. C: Visualization of reconstruction results. The reconstruction for two different scenes is presented, including 6 images and a point cloud. For ease of observation, we zoom in the point cloud at [0m, 15m] for x-axes and [-7.5m, 7.5m] for y-axes.

References

1. Bai, X., Hu, Z., Zhu, X., Huang, Q., Chen, Y., Fu, H., Tai, C.L.: Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1090–1099 (2022) 1, 2, 3, 4
2. Chen, Y., Liu, J., Zhang, X., Qi, X., Jia, J.: Largekernel3d: Scaling up kernels in 3d sparse cnns. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13488–13498 (2023) 2
3. Chen, Y., Liu, J., Zhang, X., Qi, X., Jia, J.: Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21674–21683 (2023) 2
4. Fan, L., Pang, Z., Zhang, T., Wang, Y.X., Zhao, H., Wang, F., Wang, N., Zhang, Z.: Embracing single stride 3d object detector with sparse transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8458–8468 (2022) 3

5. Hess, G., Jaxing, J., Svensson, E., Hagerman, D., Petersson, C., Svensson, L.: Masked autoencoder for self-supervised pre-training on lidar point clouds. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 350–359 (2023) 2, 3
6. Huang, L., You, S., Zheng, M., Wang, F., Qian, C., Yamasaki, T.: Green hierarchical vision transformer for masked image modeling. Advances in Neural Information Processing Systems **35**, 19997–20010 (2022) 2, 3
7. Jiao, Y., Jie, Z., Chen, S., Chen, J., Ma, L., Jiang, Y.G.: Msmdfusion: Fusing lidar and camera at multiple scales with multi-depth seeds for 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21643–21652 (2023) 2
8. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12697–12705 (2019) 2
9. Li, Y., Chen, Y., Qi, X., Li, Z., Sun, J., Jia, J.: Unifying voxel-based representation with transformer for 3d object detection. Advances in Neural Information Processing Systems **35**, 18442–18455 (2022) 2
10. Li, Y., Qi, X., Chen, Y., Wang, L., Li, Z., Sun, J., Jia, J.: Voxel field fusion for 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1120–1129 (2022) 2
11. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021) 3
12. Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D.L., Han, S.: Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 2774–2781. IEEE (2023) 1, 2, 3, 4
13. Tian, X., Ran, H., Wang, Y., Zhao, H.: Geomae: Masked geometric target prediction for self-supervised point cloud pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13570–13580 (2023) 1, 2
14. Wang, H., Shi, C., Shi, S., Lei, M., Wang, S., He, D., Schiele, B., Wang, L.: Dsvt: Dynamic sparse voxel transformer with rotated sets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13520–13529 (2023) 2
15. Yin, T., Zhou, X., Krahenbuhl, P.: Center-based 3d object detection and tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11784–11793 (2021) 2