FAC: 3D Representation Learning via Foreground Aware Feature Contrast (Supplementary Material)

Kangcheng Liu[1,†], Aoran Xiao[1], Xiaoqin Zhang[2], Shijian Lu[1,†], Ling Shao[3]
[1]Nanyang Technological University [2]Wenzhou University [3]UCAS-Terminus AI Lab, UCAS

In this supplementary material, experimental settings and details of experimental datasets, as well as imitations and future directions of this work are provided.

- Details of our further experimental settings in pre-training including data augmentation and hardware settings (see Section 1).
- Details of experimental datasets involved in pre-training and testing of our proposed GFC (see Section 2).
- Limitations and future directions of this work (see Section 3).

1. Further Pre-training Experimental Settings

1.1. Data Augmentation Details

We utilize four common types of data augmentation to generate augmented two different views in pre-training, including random rotation ([-180°, 180°]) along an arbitrary axis (applied independently for both two views), random scaling ([0.8, 1.2]), random flipping along X-axis or Y-axis, and random point dropout. We follow ProCo [34] in random point dropout and sample 100k points from the original point cloud for each of the two augmented views. 20k points are chosen from the same indexes to ensure a 20% overlap for the two augmented views, while the other 80k points are randomly sampled from the remaining point clouds. Our data augmentation strictly follows previous work ProCo [34] and CSC [5] for fair comparisons with them. Concretely, we follow ProCo [34] for outdoor 3D object detection on KITTI [3] and Waymo [30] and follow CSC [5] for other experimental cases for data augmentation.

1.2. Hardware Settings

We next report the hardware used in our experiments. The PCon [31], ProCo [34] and CSC [5] use data parallel on eight NVIDIA Tesla V100 GPUs with at least 16 GB GPU memory per card as reported in their papers. Limited by computational resources, we use data parallel on four NVIDIA 2080 Ti GPUs with 11 GB GPU memory per card in all experiments. For experiments in outdoor 3D object detection, we directly report the results of ProCo [34] in Table 1 and Table 2 of our main paper according to its original paper. It can be seen that GFC still outperforms the state-of-the-art approach ProCo [34] consistently even if much fewer computational resources are used. For all other experiments, we reimplement the CSC [5], ProCo [34], PCon [31] and use the same hardware and experimental settings as our proposed GFC in experiments for a fair comparison in Tables 3, 4, and 5 of our main paper. Specifically, we use data parallel on four NVIDIA 2080 Ti GPUs with 11 GB GPU memory per card.

2. Dataset Details

S3DIS. S3DIS is a large indoor point cloud scene understanding dataset across six large-scale indoor areas. The total number of scenes is 271. Area 5 is utilized for testing and other areas are used as the training set. Benefiting from Sparse convolution of Minkowski engine [1,4], we do not partition the 3D scene into small rooms. The S3DIS dataset has more than 215 million points with thirteen semantic classes. It is used to test the effectiveness of the proposed GFC for both indoor semantic segmentation and instance segmentation.

ScanNet-v2 (Sc) [2]. ScanNet-v2 is a large-scale and comprehensive 3D indoor scene understanding dataset consisting of 1,513 3D scans. The dataset has been adopted for tasks of semantic segmentation, instance segmentation, and object detection. The dataset is divided into 1,201 scans as the training set and 312 scans as the validation set. The number of the semantic category is 21 for semantic segmentation. The ScanNet-v2 [2] benchmark is used to test the effectiveness of the proposed GFC for indoor semantic segmentation, instance segmentation as well as indoor object detection. Also, it is used as the pre-training dataset for indoor scene understanding tasks and the outdoor semantic segmentation task on SemanticKITTI.

KITTI (K) [3]. KITTI [3] is a large-scale driving-scene dataset that covers sequential outdoor LiDAR point clouds. The KITTI 3D point cloud object detection dataset consists of 7481 labeled samples. The labeled 3D LiDAR scans are split into the training set with 3,712 scans and the validation set with 3,769 scans. The mean average precision (mAP) with 40 recall positions is typically utilized to evaluate the 3D object detection performance. The 3D IoU (Intersection over Union) thresholds are set as 0.7 for cars and 0.5 for cyclists and pedestrians. The KITTI [3] is used to test the effectiveness of the proposed GFC for outdoor 3D object detection.

SemanticKITTI (SK). SemanticKITTI is derived from the above-mentioned KITTI dataset [3] and annotated with point-level semantics. It is made up of more than 43 thousand (43,552) LiDAR scans. It is annotated with nineteen semantic classes. We follow the official split and use sequences 00-10 for training except sequence 08 for validation. The SemanticKITTI is used to test the effectiveness of the proposed GFC for outdoor semantic segmentation.

Waymo [30]. Waymo [30] is a large-scale driving-scene dataset that encompasses 158,361 LiDAR scans from 798 scenes for training and 40,077 LiDAR scans for validation. It is approximately twenty times larger than KITTI [3]. The whole training set (without label) is utilized for pre-training different 3D detection backbone networks. The training set of the Waymo [30] benchmark is used as the pre-training dataset for outdoor 3D object detection. Its validation set is also utilized to test the effectiveness of the proposed GFC for downstream fine-tuning in outdoor 3D object detection.

3. Limitation and Future Direction

In this Section, we discuss the limitations of our work and conduct some further discussions regarding future research directions.

3.1. Limitation

First, our designed *geometry-aware* and *feature-correlated* contrast (GFC) is more appropriate for understanding large-scale 3D scenes instead of the understanding of 3D shapes. We think that masked transformer-based approaches [6, 28] can surpass sole contrastive learning-based approaches in unsupervised representation learning for small-scale shape understanding in terms of performance, mainly because processing 3D shapes is less limited by the computational cost and memory consumption [7, 33, 35]. *Second*, as discussed in the related work, we do not take advantage of additional spatiotemporal information, which we think can be important to provide additional information to find feature consistency in self-supervised learning [29, 32]. However, we introduce a new simple but effective 3D pre-training framework that shows superiority compared with the state-of-the-art in knowledge transfer and data efficiency.

3.2. Future Direction

3D scene understanding is crucial to many tasks such as robot grasping and autonomous navigation [8–27]. In the future, we believe two directions deserve to be further explored to better unleash the potential of 3D unsupervised representation learning. The *first* is constructing large-scale 3D datasets with motion and spatio-temporal statistics for pre-training. The *second* is designing more advanced self-supervised learning techniques leveraging both geometry-aware and semantics-correlated features considering motion and spatiotemporal statistical cues.

References

[1] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019. 1

[2] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017. 1

[3] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1, 2

[4] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9224–9232, 2018. 1

[5] Ji Hou, Benjamin Graham, Matthias Nießner, and Saining Xie. Exploring data-efficient 3d scene understanding with contrastive scene contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15587–15597, 2021. 1

[6] Yaqian Liang, Shanshan Zhao, Baosheng Yu, Jing Zhang, and Fazhi He. Meshmae: Masked autoencoders for 3d mesh data analysis. In *European Conference on Computer Vision*, 2022. 2

[7] Haotian Liu, Mu Cai, and Yong Jae Lee. Masked discrimination for self-supervised learning on point clouds. In *European Conference on Computer Vision*. Springer, 2022. 2

[8] Kangcheng Liu. An integrated lidar-slam system for complex environment with noisy point clouds. *arXiv preprint arXiv:2212.05705*, 2022. 2

[9] Kangcheng Liu. An integrated lidar-slam system for complex environment with noisy point clouds. *arXiv preprint arXiv:2212.05705*, 2022. 2

[10] Kangcheng Liu. Rm3d: Robust data-efficient 3d scene parsing via traditional and learnt 3d descriptors-based semantic region merging. *IJCV*, pages 1–30, 2022. 2

[11] Kangcheng Liu. A robust and efficient lidar-inertial-visual fused simultaneous localization and mapping system with loop closure. In *2022 12th International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER)*, pages 1182–1187. IEEE, 2022. 2

[12] Kangcheng Liu. Robust industrial uav/ugv-based unsupervised domain adaptive crack recognitions with depth and edge awareness: From system and database constructions to real-site inspections. In *MM*, pages 5361–5370, 2022. 2

[13] Kangcheng Liu. Semi-supervised confidence-level-based contrastive discrimination for class-imbalanced semantic segmentation. In *2022 12th International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER)*, pages 1230–1235. IEEE, 2022. 2

[14] Kangcheng Liu. Learning-based defect recognitions for autonomous uav inspections. *arXiv preprint arXiv:2302.06093*, 2023. 2

[15] Kangcheng Liu. A lidar-inertial-visual slam system with loop detection. *arXiv preprint arXiv:2301.05604*, 2023. 2

[16] Kangcheng Liu and Muqing Cao. Dlc-slam: A robust lidar-slam system with learning-based denoising and loop closure. *IEEE/ASME Transactions on Mechatronics*, 2023. 2

[17] Kangcheng Liu and Ben M Chen. Industrial uav-based unsupervised domain adaptive crack recognitions: From system setups to real-site infrastructural inspections. *IEEE Transactions on Industrial Electronics*, 2022. 2

[18] Kangcheng Liu, Zhi Gao, Feng Lin, and Ben M Chen. Fg-net: A fast and accurate framework for large-scale lidar point cloud understanding. *IEEE Transactions on Cybernetics*, 53(1):553–564, 2022. 2

[19] Kangcheng Liu, Xiaodong Han, and Ben M Chen. Deep learning based automatic crack detection and segmentation for unmanned aerial vehicle inspections. In *2019 IEEE international conference on robotics and biomimetics (ROBIO)*, pages 381–387. IEEE, 2019. 2

[20] Kangcheng Liu and Huosen Ou. A light-weight lidar-inertial slam system with loop closing. *arXiv preprint arXiv:2212.05743*, 2022. 2

[21] Kangcheng Liu, Yanbin Qu, Hak-Man Kim, and Huihui Song. Avoiding frequency second dip in power unreserved control during wind power rotational speed recovery. *IEEE transactions on power systems*, 33(3):3097–3106, 2017. 2

[22] Kangcheng Liu, Aoran Xiao, Jiaxing Huang, Kaiwen Cui, Yun Xing, and Shijian Lu. D-lc-nets: Robust denoising and loop closing networks for lidar slam in complicated circumstances with noisy point clouds. In *IROS 2022, IEEE/RSJ International Conference on Intelligent Robots and Systems 2022*, pages 12212–12218. IEEE, 2022. 2

[23] Kangcheng Liu, Yuzhi Zhao, Zhi Gao, and Ben M Chen. Weaklabel3d-net: A complete framework for real-scene lidar point clouds weakly supervised multi-tasks understanding. In *2022 international conference on robotics and automation (ICRA)*, pages 5108–5115. IEEE, 2022. 2

[24] Kangcheng Liu, Yuzhi Zhao, Qiang Nie, Zhi Gao, and Ben M Chen. Weakly supervised 3d scene segmentation with region-level boundary awareness and instance discrimination. In *ECCV*, pages 37–55. Springer, Cham, 2022. 2

[25] Kangcheng Liu, Yuzhi Zhao, Qiang Nie, Zhi Gao, and Ben M Chen. Ws3d supplementary material. In *European Conference on Computer Vision (ECCV). Springer, Cham*, pages 37–55, 2022. 2

[26] Kangcheng Liu, Xunkuai Zhou, and Ben M Chen. An enhanced lidar inertial localization and mapping system for unmanned ground vehicles. In *2022 International Conference on Control & Automation*, pages 587–592. IEEE, 2022. 2

[27] Kangcheng Liu, Xunkuai Zhou, Benyun Zhao, Huosen Ou, and Ben M Chen. An integrated visual system for unmanned aerial vehicles following ground vehicles: Simulations and experiments. In *2022 IEEE 17th International Conference on Control & Automation*, pages 593–598. IEEE, 2022. 2

[28] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point

cloud self-supervised learning. In *European Conference on Computer Vision*. Springer, 2022. 2

[29] Yongming Rao, Benlin Liu, Yi Wei, Jiwen Lu, Cho-Jui Hsieh, and Jie Zhou. Randomrooms: Unsupervised pre-training from synthetic shapes and randomized layouts for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3283–3292, 2021. 2

[30] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2454, 2020. 1, 2

[31] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *European Conference on Computer Vision*, pages 574–591. Springer, 2020. 1

[32] Ryosuke Yamada, Hirokatsu Kataoka, Naoya Chiba, Yukiyasu Domae, and Tetsuya Ogata. Point cloud pre-training with natural 3d structures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21283–21293, 2022. 2

[33] Siming Yan, Zhenpei Yang, Haoxiang Li, Li Guan, Hao Kang, Gang Hua, and Qixing Huang. Implicit autoencoder for point cloud self-supervised representation learning. *arXiv preprint arXiv:2201.00785*, 2022. 2

[34] Junbo Yin, Dingfu Zhou, Liangjun Zhang, Jin Fang, Cheng-Zhong Xu, Jianbing Shen, and Wenguan Wang. Proposal-contrast: Unsupervised pre-training for lidar-based 3d object detection. In *European Conference on Computer Vision*, pages 574–591. Springer, 2022. 1

[35] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19313–19322, 2022. 2