講演番号	391

Self-Supervised Pre-Training for Generalizable Feature Extraction from 4D Imaging Radar Point Cloud

Takumi Takai [1)] Keisuke Yoneda [1)] Keigo Hariya [1)]

Haku Shinoda [1)] Yukiya Fukuda [1)] Naoki Suganuma [1)]

1) Kanazawa University, Graduate School of Natural Science and Technology, Division of Mechanical Science and Engineering Kakumamachi, Kanazawa-shi, Ishikawa, 920-1164, Japan (E-mail: takatakusty@stu.kanazawa-u.ac.jp)

KEY WORDS: **Software and its Underlying Technologies, Autonomous driving system platform, Machine learning,** Self-supervised learning(E3)

Automobiles play a crucial role in modern society, but challenges such as traffic accidents and driver shortages in the logistics industry remain. Autonomous driving is attracting attention as one way to address these challenges. Recently, 4D imaging radar (4DMWR) is increasingly utilized in autonomous driving for its environmental robustness, long-range detection capabilities, and ability to acquire velocity information. However, 4DMWR point clouds are inherently sparse and noisy compared to LiDAR. Consequently, directly applying 4DMWR point clouds to downstream tasks results in limited performance. To effectively exploit 4DMWR data, the acquisition of informative feature representations is essential. Masked prediction pre-training has shown strong performance in natural language processing and computer vision and has attracted increasing attention for acquiring generalizable representations. Although masked prediction has been extended to point clouds, it is not directly applicable to sparse and noisy 4DMWR point clouds. Therefore, we propose a novel self-supervised pre-training framework specifically designed for 4DMWR which learns feature representations incorporating radar-specific physical characteristics. Finally, the pre-trained model is fine-tuned for a 3D object detection task to evaluate the effectiveness of the proposed pre-training.

Figure 1 illustrates the pre-training framework. The 3D point cloud is divided into pillars and converted into a pseudo-image. Our masked prediction targets Pillar Tokens compressed into the latent space. Thus, a Pillar Tokenizer is trained to convert pillar features into discrete token representations to generate the ground-truth Pillar Tokens. The Tokenizer is trained via reconstruction from latent representations derived from 3D point clouds. In particular, by reconstructing voxel coordinates, Radar Cross Section (RCS), and point density, the model incorporates radar- specific physical characteristics into the latent representations. Subsequently, masking is applied to the pseudo-image, and a Swin Transformer-based backbone extracts features from the masked image to acquire informative feature from the 4DMWR point clouds. The model predicts three resolutions of Pillar Tokens, calculating the Mean Squared Error (MSE) in masked regions. To validate this approach, fine-tuning for 3D object detection compares models initialized with pre-trained weights against those with random parameters.

Our dataset is used for Tokenizer and pre-training, and another dataset with four classes (Car, Large Car, Pedestrian, Cyclist) for detection. Performance is evaluated using the F-score. We compare four models: the Swin Transformer and a standard CNN based backbone (SECOND), both evaluated with and without pre-training.

Table 1 shows Swin(+PT) outperforms Swin(Base) across all classes, improving the average F-score by 3.82 points. This confirms our pre-training effectively extracts features from sparse 4DMWR data. While Swin(Base) initially trails the CNN-based SECOND(Base), Swin(+PT) surpasses it, achieving a 1.02-point higher average F-score. Furthermore, pre-training improves Swin much more (+3.82 points) than SECOND (+0.53 points), indicating that it better utilizes Swin's higher capacity and attention-based modeling. Table 2 evaluates the Car class for static and dynamic objects. Pre-training improves F-scores for both static (7.23% to 12.31%) and dynamic (49.93% to 58.51%) objects. However, static object detection remains significantly lower, as 4DMWR's inherent sparsity and ambiguous shapes make distinguishing them from the background challenging.

Future work will focus on refining the pre-training mechanism to better distinguish static objects from the background. Additionally, we aim to construct a novel pre-training framework utilizing past frames as temporal information to compensate for the inherent sparsity of 4DMWR point clouds.

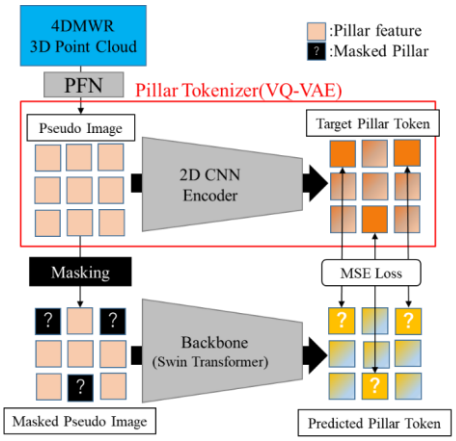

Fig.1 Overview of pre-training

Table1 Results of each model by F-score[%]

Model	Car	L-Car	Ped.	Cyclist	Avg.
SECOND(Base)	36.75	16.81	6.29	28.46	22.08
SECOD(+PT)	38.72	17.18	6.48	28.04	22.61 +0.53
Swin(Base)	31.73	12.31	5.83	27.24	19.28
Swin(+PT)	38.61	17.59	7.12	29.09	23.10 +3.82

PT: Pre-Training

Table2 Results of dynamic and static by F-score[%]

Model	Static	Dynamic
Swin(Base)	7.23	49.93
Swin(+PT)	12.31	58.51