Seeing 3D Through 2D Lenses: 3D Few-Shot Class-Incremental Learning via Cross-Modal Geometric Rectification – Supplementary Materials –

Tuo Xiang[1] Xuemiao Xu[1,2,3,4] Bangzhen Liu[1*] Jinyi Li[1] Yong Li[1*] Shengfeng He[5]
[1]South China University of Technology [2]State Key Laboratory of Subtropical Building Science
[3]Guangdong Provincial Key Lab of Computational Intelligence and Cyberspace Information
[4]Ministry of Education Key Laboratory of Big Data and Intelligent Robot [5]Singapore Management University

In the supplementary material, we provide more details on hyperparameter selection (Sec.1), and further analyze the computational complexity of the core module and overall of our model (Sec.2).

1. More Analysis on Hyperparameter Selection

We conduct several experiments on the S2C setting by modifying the value of hyperparameters to investigate the impact of different masking ratios and self-attention layer combinations.

As shown in Table 1, Masking Ratio M_R denotes the ratio in the attention weight matrix being masked by row, and N_{sa} represents the number of layers performing self-attention. The SAGR module is non-sensitive to the selection of hyperparameters. This may potentially be due to the less reliance on specific attention patterns in the cross-modal integration of 3D and 2D information. We introduce M_R and N_{sa} as a regularization technique, serving as a complement to the main cross entropy loss.

Table 1. Selection of masking ratio and self-attention layers in the Structure-Aware Geometric Rectification module.

Parameters	Masking Ratio (M_R)				
	0.9	0.7	0.5	0.3	0.1
$N_{sa}=4$	78.10	78.21	78.40	78.02	78.47
$N_{sa}=3$	78.01	77.98	78.43	78.34	78.24
$N_{sa}=2$	**78.63**	78.19	78.30	78.57	78.29

In addition, the involvement of any CLIP layers already improves performance, as shown in Table 2. We sample across the 12-layer CLIP at different intervals, resulting in four sets of results with varying numbers of sampled layers. Each set includes different sampling starting points. The numbers in the table indicate the indices of CLIP's Transformer layers, ranging from 0 to 11. The results reinforce our key finding that 2D features are valuable in 3D tasks. Compared to not incorporating 2D information, integrating 2D and 3D features via the cross attention mechanism helps the model learn more robust and discriminative representations, thus better leveraging the prior knowledge of large pretrained models to overcome catastrophic forgetting and overfitting. We choose three layers as this offers the best trade-off: using more layers introduces redundancy, while fewer layers may lack sufficient cross-modal cues. $[0, 4, 8]$ corresponds to the best empirical outcome.

Table 2. Impact of varying L_r.

Selection of L_r	$AA\uparrow$
$\emptyset$	75.19
{1, 3, 5, 7, 9, 11}	77.72
{0, 2, 4, 6, 8, 10}	78.16
{1, 4, 7, 10}	77.90
{0, 3, 6, 9}	78.21
{1, 5, 9}	78.19
{0, 4, 8}	**78.63**
{1, 7}	77.98
{0, 6}	77.87

2. Computational Complexity

Computational complexity is crucial for the deployment in industrial applications. Below, we report the estimated resource consumption in terms of FLOPs, number of parameters, total runtime on S2C task, and inference speed for our model and FILP-3D [1]. Settings of the experiment can be found in the implementation details section of the main paper. Table 3 shows that our approach can significantly improve performance while introducing acceptable computational overhead. In future work, we will further explore lightweight 3DFSCIL models.

Table 3. Computational complexity analysis.

	FLOPs (G)	Params (M)	Runtime (h)	Infer Speed (n/s)
FILP-3D	103.32	121.14	2.65	47
Ours	155.98	185.85	3.07	38
-SAGR	63.22	48.72	/	/

References

[1] Wan Xu, Tianyu Huang, Tianyu Qu, Guanglei Yang, Yiwen Guo, and Wangmeng Zuo. Filp-3d: Enhancing 3d few-shot class-incremental learning with pre-trained vision-language models. *arXiv preprint arXiv:2312.17051*, 2023. 1