Pattern Recognition Letters
journal homepage: www.elsevier.com

Data-efficient 3D Instance Segmentation by Transferring Knowledge from Synthetic Scans *Supplementary Material*

Xiaodong Wu[a,b], Ruiping Wang[a,b,**], Xilin Chen[a,b]

[a] *Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100049, China*
[b] *University of Chinese Academy of Sciences, Beijing 100049, China*

ABSTRACT

This is a supplementary material for "Data-efficient 3D Instance Segmentation by Transferring Knowledge from Synthetic Scans", which offers additional implementation details, experimental results, and discussions due to the space limitation in the main paper.

© 2023 Elsevier Ltd. All rights reserved.

Table 1. Instance segmentation results on the ScanNet Limited Annotation Benchmark (200 points). Models are initialized from scratch or from a model pre-trained on a synthetic dataset (InteriorNetScan).

	AP	AP@50	AP@25
From scratch	28.9	48.8	63.1
Finetune InteriorNetScan	**33.2**	**53.3**	**68.9**

1. Experiments on the limited annotation benchmark

In practical applications, besides scenarios with a limited number of annotated scenes, there are also cases where there are fewer annotated points within point clouds. To verify the effectiveness of synthetic data in such scenarios, we conducted experiments on the ScanNet Limited Annotation (LA) benchmark. Specifically, for each scene in the training set, only 200 points were annotated. We followed a methodology similar to PointGroup, with the distinction that our clustering was performed solely within shifted coordinate space. The ScoreNet proposed in PointGroup was not utilized; instead, we calculated the average semantic score of points belonging to the same instance as instance scores. We report AP, AP@50, and AP@25 as evaluation metrics. The results of the experiment are presented in Table 1. It's evident that instance segmentation performance improved after pretraining with synthetic data.

2. Experiments on the full ScanNet benchmark

We have already validated the effectiveness of synthetic data on the Data-efficient ScanNet benchmark. In this section, we further validate its performance on the ScanNet full dataset. In our experiments here, we followed the experimental settings of PointGroup [1] without bells and whistles. We utilized a constant learning rate and SGD optimizer. The model parameters were initialized from scratch, initialized from real data pre-training (3RScan [2]), initialized from synthetic data pre-training (InteriorNetScan), and initialized from target-aware pre-training using synthetic data. To simplify the experiment, we conducted instance clustering only in shifted coordinate space. Table 2 presents the experimental results. Comparing the first and third rows (from scratch and fine-tune on a model pre-trained without our strategy) reveals that synthetic data is also effective on the ScanNet full dataset, albeit with relatively limited performance improvement. Contrasting the second and fourth rows (fine-tune from 3RScan and fine-tuning from a model pre-trained on InteriorNetScan with our strategy) demonstrates that models pre-trained on synthetic data with our method outperform real data. Comparing the third and fourth rows (without and with our target-aware pre-training strategy), it can be seen that after employing the target-aware pre-training strategy, performance is further enhanced. This indicates that introducing target domain data during pre-training to adapt to the target domain's data distribution is beneficial.

3. Effectiveness of target-aware Pre-training

In our experiments, using a model trained with target-aware pre-training has shown performance improvements on down-

**Corresponding author
e-mail: xiaodong.wu@vipl.ict.ac.cn (Xiaodong Wu), wangruiping@ict.ac.cn (Ruiping Wang), xlchen@ict.ac.cn (Xilin Chen)

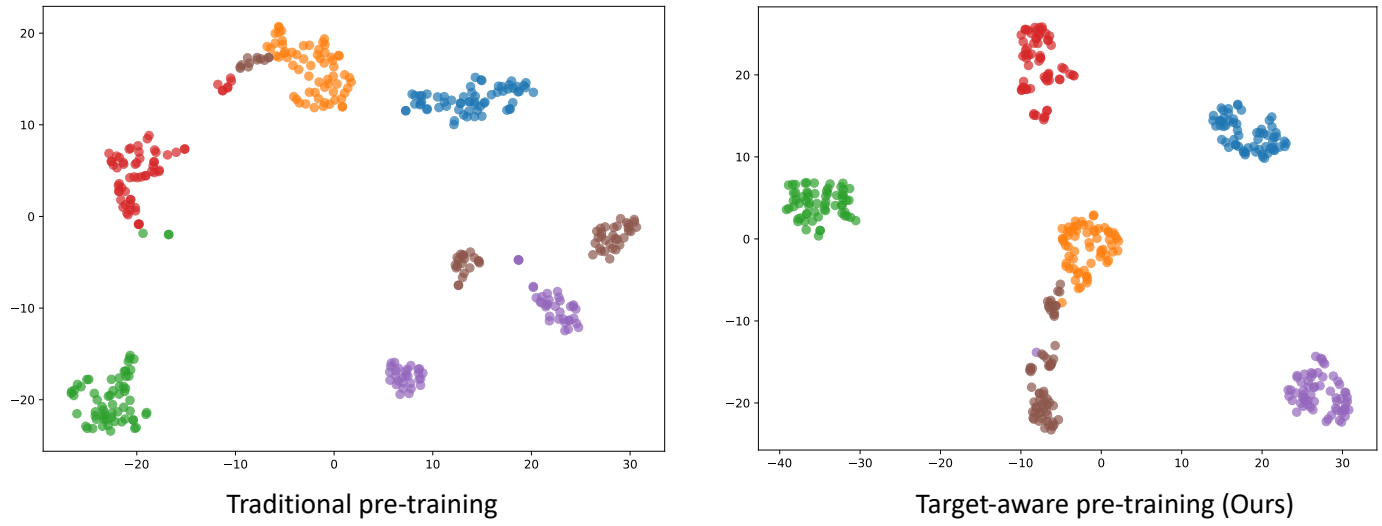

Fig. 1. t-SNE visualization of point cloud features. Our method (right) adopts target-aware pre-training. Traditional pre-training (left) does not take this strategy. Different colors represent different semantic categories.

Table 2. Comparison of instance segmentation results on ScanNet v2 full dataset. Models are initialized from scratch or from a model pre-trained on a real dataset (3RScan) or from a model pre-trained on a synthetic dataset (InteriorNetScan) with or without target-aware pre-training.

	AP	AP@50	AP@25
From scratch (Baseline)	38.0	58.2	70.5
Fine-tune from 3RScan	40.7	59.8	71.4
Without target-aware pre-training	40.2	59.3	71.3
With target-aware pre-training	**41.1**	**60.6**	**72.6**

Table 3. Ablation study results investigating the quantity of target data utilized in target-aware pre-training. During the pre-training process, 10% and 20% of the ScanNet data are respectively mixed with the synthetic dataset InteriorNetScan for pre-training. The downstream task is fine-tuning on 20% labeled ScanNet and the instance segmentation results (AP, AP@50 and AP@25) on ScanNet validation set are reported.

	AP	AP@50	AP@25
Target-aware Pre-training (10% ScanNet)	37.6	56.2	68.5
Target-aware Pre-training (20% ScanNet)	**38.4**	**57.4**	**68.6**

stream tasks. This is mainly attributed to the introduction of unlabeled target domain data and appropriate pseudo-label training, enabling the pre-trained model to adapt to the distribution of the target domain data, thereby reducing the difficulty of knowledge transfer. To validate this point, we visualized the feature space of the pre-trained models. Figure 1 displays the t-SNE visualization of features extracted by models trained using target-aware pre-training and traditional pre-training on point clouds from synthetic scenes. InteriorNetScan is used for pre-training. We utilized the output of the last layer of the U-Net backbone as the visualization input. Different colors represent different semantic categories. It can be observed that the features extracted by target-aware pre-training exhibit noticeable clustering of target domain data, indicating that our approach enables the pre-trained model to have stronger discriminative abilities on the target domain data.

4. Ablation study on target domain data quantity

In target-aware pre-training, we utilized unlabeled target domain data to characterize the data distribution of the target domain. Here, we further analyzed the impact of the amount of target domain data used in the pre-training process on the performance in the target domain. Specifically, we employed 10% and 20% of unlabeled ScanNet data separately in the target-aware pre-training process for InteriorNetScan. To ensure the same number of iterations and reduce experimental time, we trained them for 160 epochs and 80 epochs, respectively. After obtaining the trained models, we evaluated their performance on the 20% ScanNet task. The results obtained are shown in Table 3. It can be observed that increasing the amount of target domain data allows the model to perform better in the target domain task. This is attributed to the larger dataset enabling the model to more accurately characterize the data distribution of the target domain.

5. Analysis on virtual scans

In the process of constructing the dataset, we acquired point cloud data by conducting virtual scans on synthesized scenes. This method of generating point cloud data simulates the real point cloud acquisition process and produces similarly incomplete point clouds. Another approach to generate point cloud data from synthetic data involves directly sampling point clouds on the surface of object CAD models. To examine the differences between these two methods, we conducted experiments on SceneNetScan. For each scene in SceneNetScan, we generated a sample-point-based point cloud that was identical to the scene settings. In order to reduce training time, we generated 5000 3D point clouds using each method, resulting in SceneNetScan-5000 and SceneNetSample-5000, respectively.

Fig. 2. Visualization of the synthetic point clouds in InteriorNetScan.

Table 4. Comparison of pre-training on synthetic point cloud collected by virtual scanning and sampling. Models are pre-trained on SceneNetScan-5000 and ScenenNetSample-5000 respectively and fine-tuned on ScanNet 10%. Performances on the validation set of ScanNet are reported.

	AP	AP@50	AP@25
SceneNetSample-5000	27.0	45.2	60.1
SceneNetScan-5000	27.0	45.3	60.6

After pretraining on these datasets for 32 epochs, we initialized downstream tasks using ScanNet 10% as the training set. We used the two pretrained models as parameter initializations and trained for 384 epochs. The performance on the downstream tasks is outlined in Table 4. The point cloud acquisition method using virtual scanning has gained a slight advantage over the sampling method observed on the SceneNetScan dataset. The increase in performance isn't substantial. This might be because learning in the feature space is not highly sensitive to the completeness of the point cloud; both acquisition methods can learn relatively good feature representations.

6. Dataset visualization

We visualized three large-scale synthetic datasets, InteriorNetScan, procTHORScan, and SceneNetScan. For better visualization, we obscured the ceiling. The visual results are shown in Fig. 2, Fig. 3, and Fig. 4.

References

[1] L. Jiang, H. Zhao, S. Shi, S. Liu, C. Fu, J. Jia, Pointgroup: Dual-set point grouping for 3d instance segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4866–4875.

[2] J. Wald, A. Avetisyan, N. Navab, F. Tombari, M. Nießner, Rio: 3d object instance re-localization in changing indoor environments, in: IEEE International Conference on Computer Vision (ICCV), 2019, pp. 7657–7666.

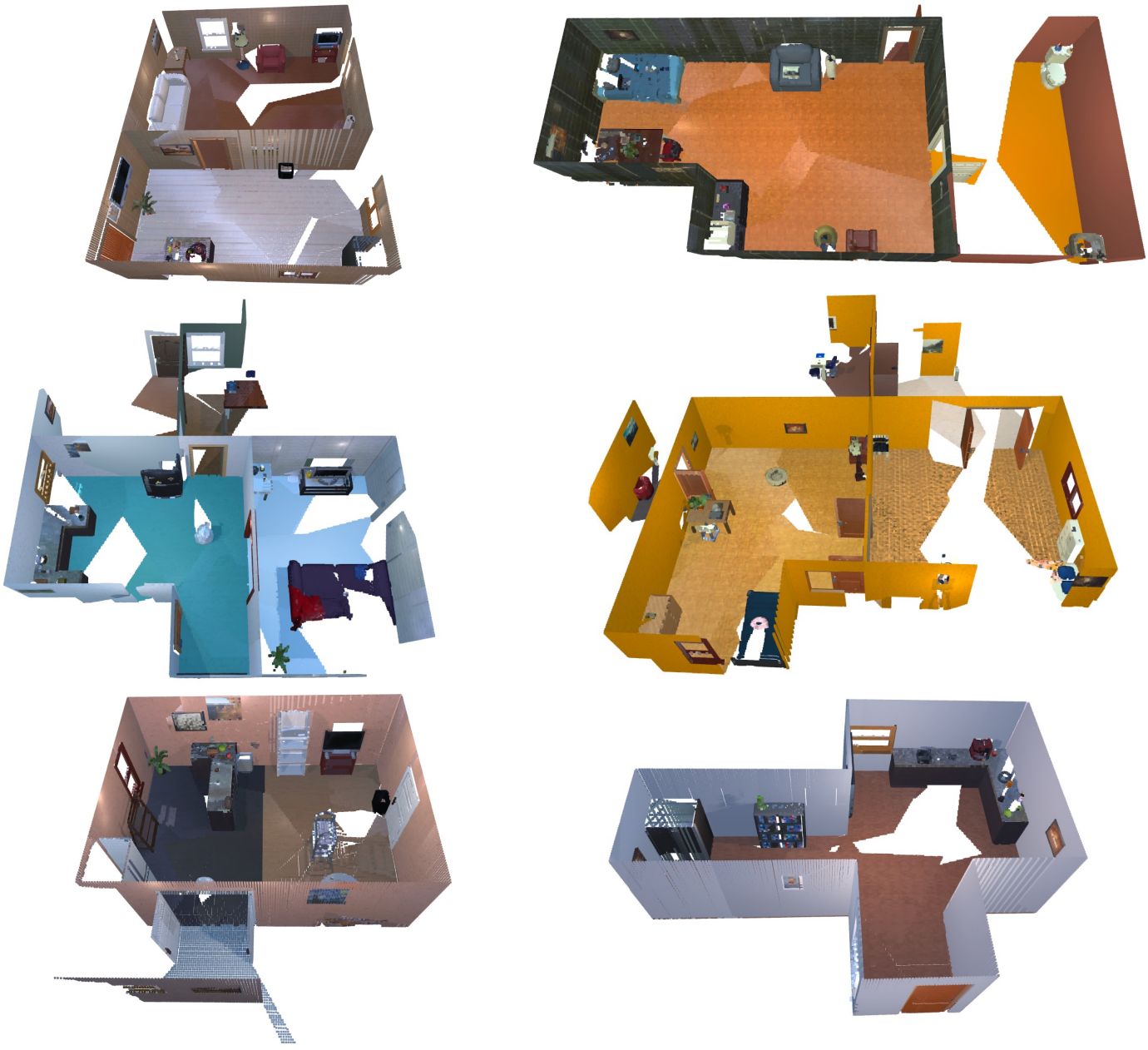

Fig. 3. Visualization of the synthetic point clouds in procTHORScan.

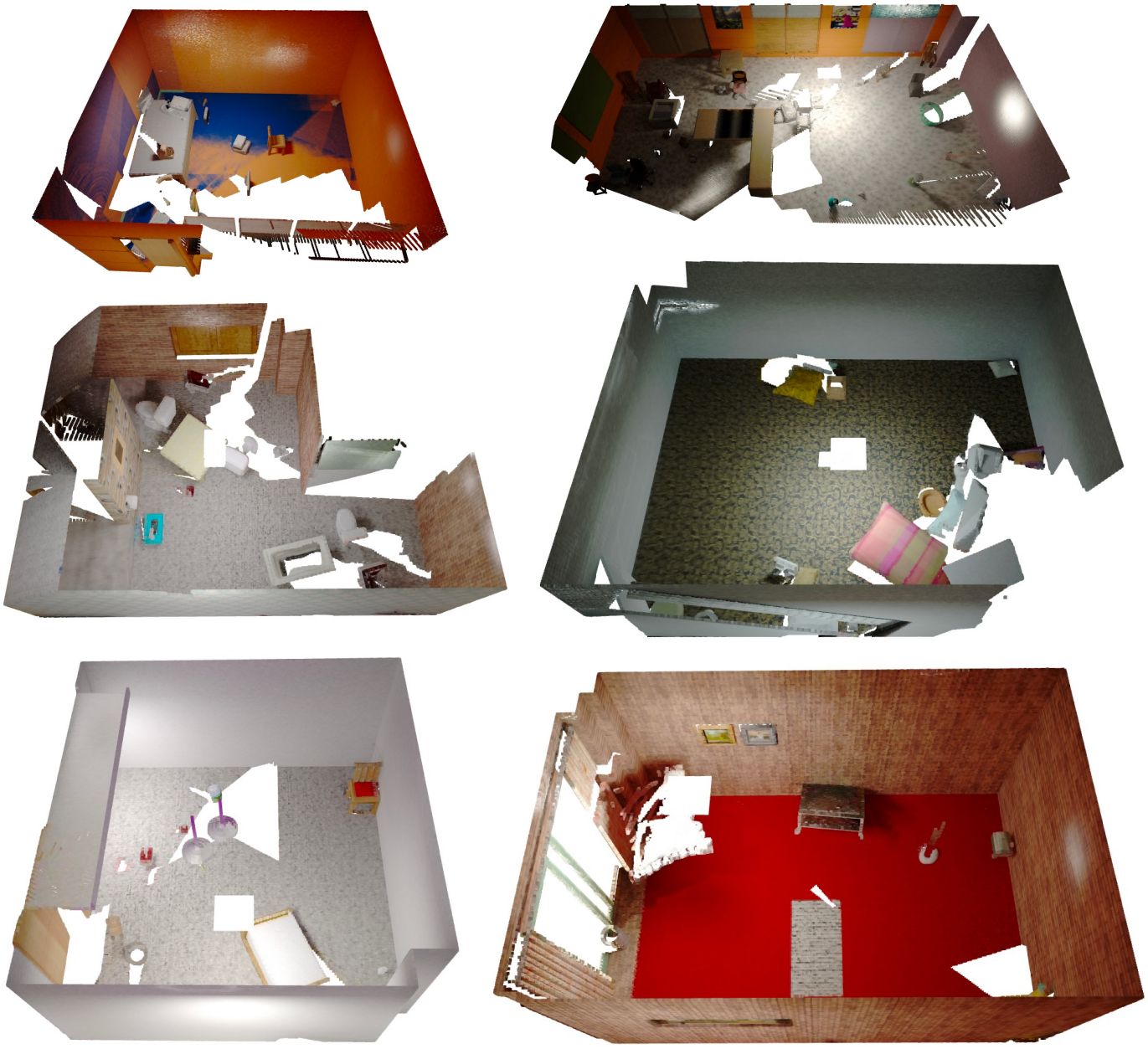

Fig. 4. Visualization of the synthetic point clouds in SceneNetScan.