TAMM: TriAdapter Multi-Modal Learning for 3D Shape Understanding

Supplementary Material

A. Implementation Details

For a fair comparison with previous methods, we adopt two representative 3D encoders for our study: Point-BERT [12] (Transformer-based) and SparseConv [3] (convolution-based), following the same architectural configurations as prior methods [7, 11]. We employ OpenCLIP-ViT-G/14 [2] as the pre-trained CLIP model. TAMM is pre-trained for 200 epochs using the AdamW optimizer [6, 8], and a cosine learning rate scheduler with and a two-epoch warm-up, and a base learning rate of 5×10^{-4}. Regarding the CLIP Image Adapter and Dual Adapters, we set α to 0.2 in Equation 3 and employ ReLU [1] and GELU [5] activation functions, respectively, following [4, 11].

B. Additional Results on Complex Scene Recognition

To further assess TAMM's capability in understanding 3D shapes from scene data, we conduct experiments using the Hypersim dataset [9], a photorealistic synthetic dataset designed for comprehensive indoor scene understanding. In this experiment, we extract the point clouds of object instances from segmentation annotations and focus on 17 classes to evaluate TAMM's zero-shot recognition ability. Some classes are excluded due to their amorphous shapes (*e.g.*, "floor," "ceiling") or because they are not well-defined for classification (*e.g.*, "otherfurniture," "otherstructure"). The results are detailed in Table 7, which demonstrate that TAMM surpasses OpenShape in terms of both overall accuracy and average of per-class accuracy, with respective improvements of $+5.9\%$ and $+1.8\%$. Significantly, TAMM also outperforms OpenShape in 11 out of the 17 evaluated classes. This evaluation on Hypersim underscores TAMM's robustness in recognizing and understanding 3D shapes derived from various scene contexts.

C. Additional Results on Instance Segmentation

To delve deeper into TAMM's proficiency in 3D scene understanding, we test whether the 3D backbone pre-trained by TAMM can further enhance SoftGroup++ [10], the current state-of-the-art 3D instance segmentation method. More specifically, we integrate the pre-trained Point-BERT model into the feature extractor module in the top-down refinement stage of SoftGroup++, and subsequently fine-tune the classification branch. The 3D instance segmentation results on ScanNet are illustrated in Table 6. These results reveal that TAMM can indeed improve the overall performance of SoftGroup++. Notably, TAMM attains an AP/AP_{50} score of $46.1\%/68.0\%$, marking an enhancement of $0.6\%/1.0\%$ over SoftGroup++. Furthermore, as an improved pre-training approach, TAMM exceeds OpenShape by 0.4% AP and 0.6% AP_{50}. The results on real-world instance segmentation underscores TAMM's significant potential in the tasks of scene-level 3D understanding.

Method	AP	AP_{50}	AP_{25}
SoftGroup++† [10]	45.5	67.0	78.7
SoftGroup++ & OpenShape [7]	45.7	67.4	78.7
SoftGroup++ & TAMM (Ours)	**46.1**	**68.0**	**79.0**

† Reproduced using the original implementation [10].

Table 6. **3D instance segmentation results on ScanNet v2.** Incorportating TAMM into SoftGroup++ improves the AP performance from 45.5% to 46.1%, achieving the best results.

D. Additional Qualitative Results

In this section, we provide additional qualitative results to supplement the visualizations presented in the main body of the paper. Figure 4 showcases examples of cross-modal retrieval from text to 3D point clouds. Figure 5 showcases examples of cross-modal retrieval from 2D images to 3D point clouds. More specifically, we extract the adapted features of the query text or image and employ the TAMM-learned 3D backbone to find the point clouds with the most similar features. The retrieved point clouds highly resemble objects in the query text or images, reflecting that the representations learned by TAMM are cross-modal and unified. Figure 6 demonstrates how our CLIP Image Adapter (CIA) effectively bridges the domain gap caused by rendered images, resulting in more accurate image-text matching. Additionally, Figure 7 illustrates the distinctive yet synergistic roles of Image Alignment Adapter (IAA) and Text Alignment Adapter (TAA). These adapters learn 3D representations with focuses on vision and semantics, respectively. Their integration yields more robust and comprehensive 3D representations, highlighting the effectiveness of our approach.

Method	OAvg.	Avg.	Cabi	Bed	Chair	Sofa	Tabl	Door	Bksh	Shlv	Curt	Pill	Clth	TV	Papr	Twl	Nght	Sink	Lamp
OpenShape† [7]	56.7	48.8	13.0	**40.0**	71.2	70.7	**73.3**	81.4	**20.5**	54.6	**66.7**	60.0	23.9	**43.1**	36.5	**24.0**	**38.1**	62.0	51.0
TAMM (Ours)†	**62.6**	**50.6**	**20.6**	38.2	**75.3**	**76.2**	72.6	**88.1**	8.2	**59.1**	61.9	**62.8**	**41.0**	26.2	**57.3**	**24.0**	31.0	**63.0**	**55.0**

† Results using Point-BERT [12] as 3D encoder, pre-trained on the Ensembled dataset.

Table 7. **Zero-shot classification results on the Hypersim dataset. OAvg.**: Overall Top-1 accuracy of all shapes. **Avg.**: Mean average Top-1 accuracy of all classes. TAMM achieves the best results under both metrics.

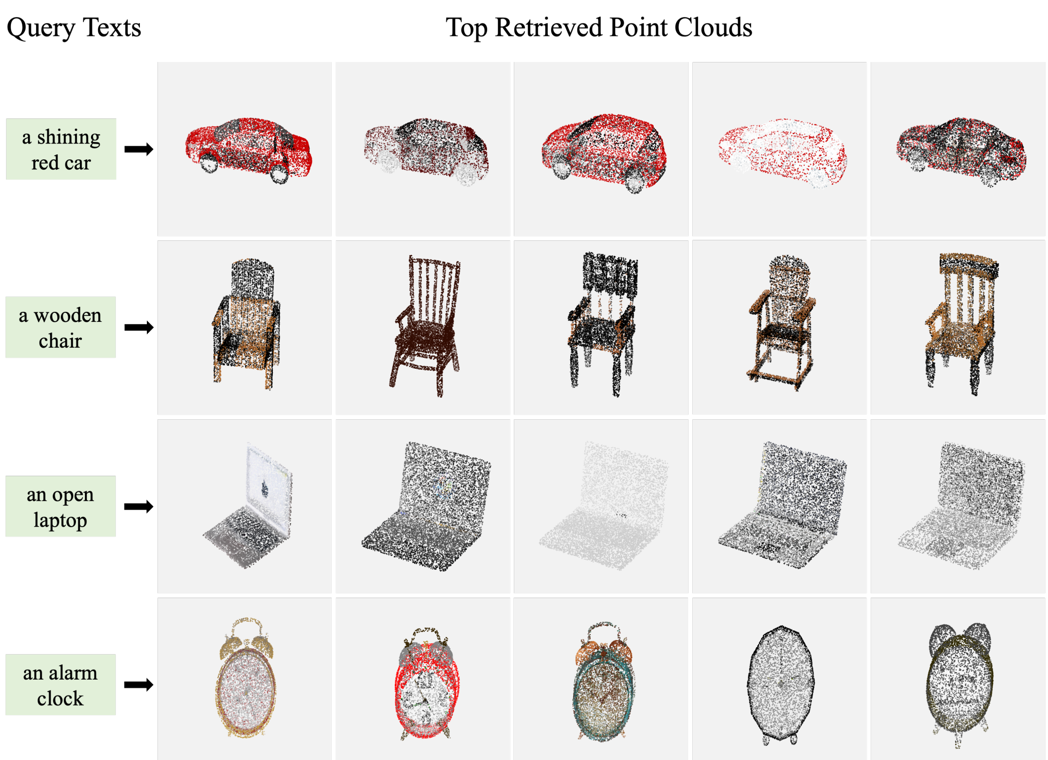

Figure 4. **Qualitative results of text-to-point-cloud retrieval.** We use TAMM to acquire the features of the given query text and retrieve the point clouds with the most similar features. The shown examples demonstrate TAMM's strong multi-modal comprehension.

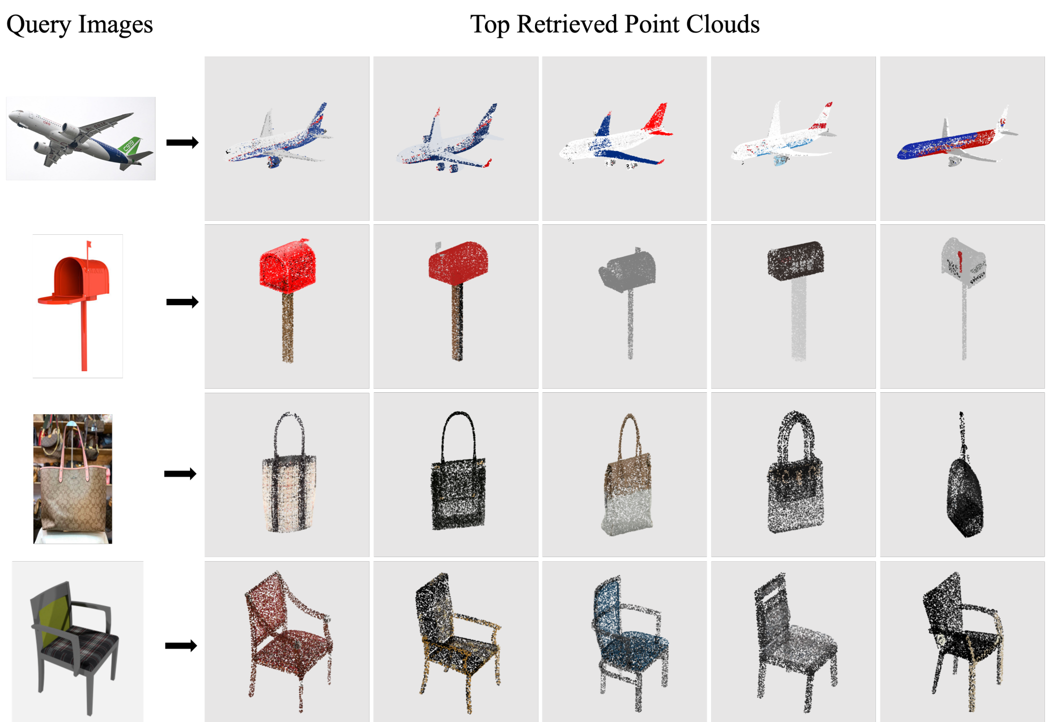

Figure 5. **Qualitative results of image-to-point-cloud retrieval.** We use TAMM to acquire the features of the given query images and retrieve the point clouds with the most similar features. The shown examples demonstrate TAMM's strong multi-modal comprehension.

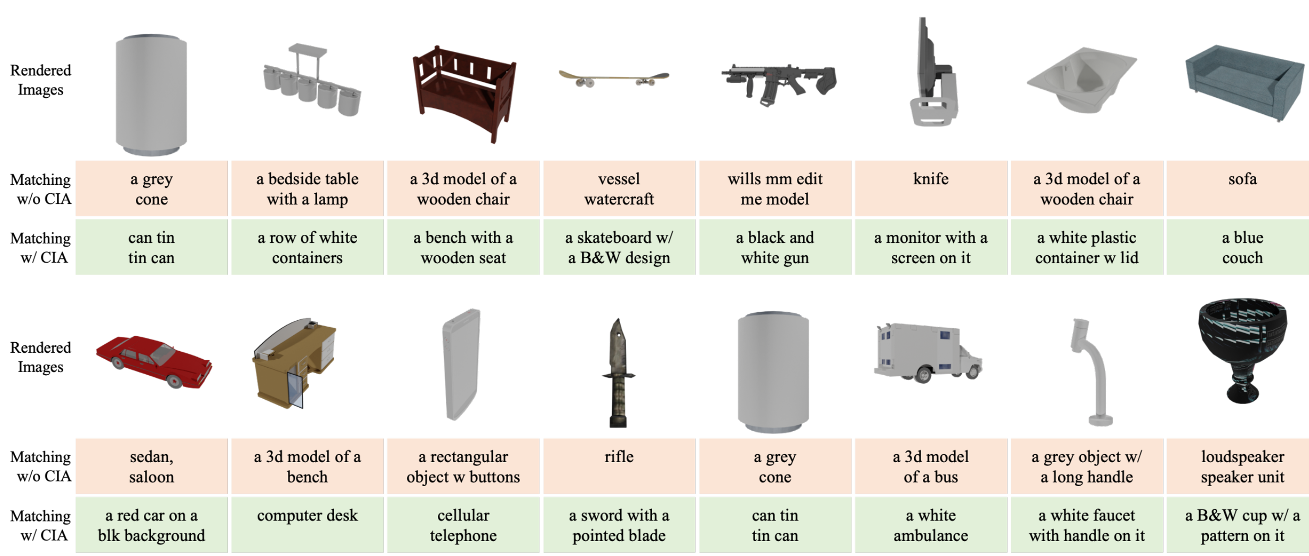

Figure 6. **Qualitative results of CLIP Image Adapter (CIA).** CIA re-aligns the images rendered from 3D shapes with the text descriptions. The rendered images are inaccurately matched with text when the image features are directly extracted by CLIP, and CIA can correct the matching.

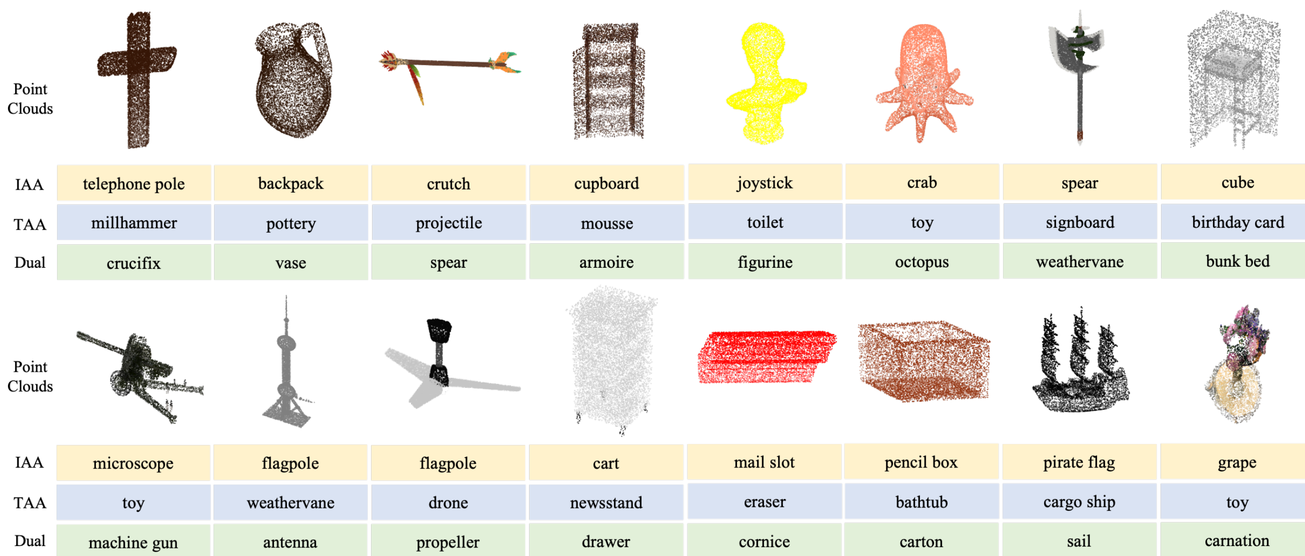

Figure 7. **Qualitative results of Image Alignment Adapter (IAA) and Text Alignment Adapter (TAA).** IAA and TAA decouple 3D features with complementary visual and semantic focuses. Features from one single adapter are matched with classes whose appearance or semantics resemble the true class; using both adapters leads to the correct class.

References

[1] Abien Fred Agarap. Deep learning using rectified linear units (ReLU). *arXiv preprint arXiv:1803.08375*, 2018. 1

[2] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, 2023. 1

[3] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4D spatio-temporal ConvNets: Minkowski convolutional neural networks. In *CVPR*, 2019. 1

[4] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. CLIP-Adapter: Better vision-language models with feature adapters. *IJCV*, 132(2):581–595, 2024. 1

[5] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016. 1

[6] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 1

[7] Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. OpenShape: Scaling up 3D shape representation towards open-world understanding. In *NeurIPS*, 2023. 1, 2

[8] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 1

[9] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *ICCV*, 2021. 1

[10] Thang Vu, Kookhoi Kim, Tung M. Luu, Xuan Thanh Nguyen, and Chang D. Yoo. SoftGroup for 3D instance segmentation on 3D point clouds. In *CVPR*, 2022. 1

[11] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. ULIP: Learning a unified representation of language, images, and point clouds for 3D understanding. In *CVPR*, 2023. 1

[12] 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 *CVPR*, 2022. 1, 2