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

Zhihao Zhang[1*] Shengcao Cao[2*] Yu-Xiong Wang[2]
[1]Xi'an Jiaotong University [2]University of Illinois Urbana-Champaign
[1]zh1142@stu.xjtu.edu.cn [2]{cao44,yxw}@illinois.edu

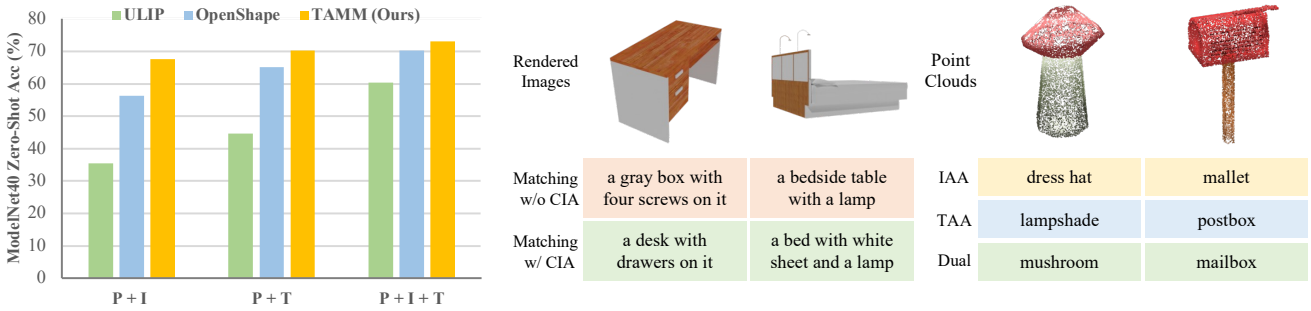

Figure 1. **Our TriAdapter Multi-Modal Learning (TAMM) significantly enhances 3D shape understanding. Left:** When aligning features of 3D point clouds (P) with 2D images (I) and/or text (T), prior methods (*e.g.*, ULIP [45] and OpenShape [23]) *under-utilize the image modality*, due to the overlooked or unsolved image domain gap. TAMM better exploits the image modality and brings more gains when learning from both image and text data. The results are produced by pre-training Point-BERT [47] on ShapeNet [3]. **Middle:** Our CLIP Image Adapter (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. **Right:** Our Image Alignment Adapter (IAA) and Text Alignment Adapter (TAA) *decouple 3D features* with complementary visual and semantic focuses. In the visualized examples, 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.

Abstract

The limited scale of current 3D shape datasets hinders the advancements in 3D shape understanding, and motivates multi-modal learning approaches which transfer learned knowledge from data-abundant 2D image and language modalities to 3D shapes. However, even though the image and language representations have been aligned by cross-modal models like CLIP, we find that the image modality fails to contribute as much as the language in existing multi-modal 3D representation learning methods. This is attributed to the domain shift in the 2D images and the distinct focus of each modality. To more effectively leverage both modalities in the pre-training, we introduce TriAdapter Multi-Modal Learning (TAMM) – a novel two-stage learning approach based on three synergistic adapters. First, our CLIP Image Adapter mitigates the domain gap between 3D-rendered images and natural images, by adapting the visual representations of CLIP for synthetic image-text pairs. Subsequently, our Dual Adapters decouple the 3D shape representation space into two complementary sub-spaces: one focusing on visual attributes and the other for semantic understanding, which ensure a more comprehensive and effective multi-modal pre-training. Extensive experiments demonstrate that TAMM consistently enhances 3D representations for a wide range of 3D encoder architectures, pre-training datasets, and downstream tasks. Notably, we boost the zero-shot classification accuracy on Objaverse-LVIS from 46.8% to 50.7%, and improve the 5-way 10-shot linear probing classification accuracy on ModelNet40 from 96.1% to 99.0%. Project page: https://alanzhangcs.github.io/tamm-page.

1. Introduction

Despite the recent success of 3D shape understanding [16, 19, 24, 42, 47, 51, 56], the limited scale of existing 3D shape datasets hinders the learning of more robust and generalizable 3D representations. Due to the significant human labor and expertise required in collecting and annotating 3D

*Equal contribution.

shape data, building large-scale 3D shape datasets is a prohibitive endeavor [3, 9, 26].

To mitigate the challenge of the limited data scale, recent research in 3D shape representation learning shows a promising direction of multi-modal learning [23, 45]. By establishing connections among data of 3D shapes, 2D images, and text, it becomes possible to pre-train 3D shape representations by *distilling learned knowledge from image and language modalities*. There exists a comprehensive body of literature of massive datasets [36], high-quality models pre-trained via self-supervised learning for 2D image [2, 4, 15] and language [1, 10, 39] modalities, and furthermore, cross-modal vision-language models [5, 34]. For instance, ULIP [45] creates triplets of 3D point clouds, 2D images, and text by rendering images from 3D shapes and generating text descriptions from their metadata. Then, ULIP adopts contrastive learning to align the 3D shape features with both the image features and text features extracted by CLIP [34].

While the image and language modalities are pre-aligned by CLIP pre-training and share the same feature space, we observe that directly aligning 3D features with image features leads to *considerably worse representation quality* as compared with aligning 3D features with text features, as shown in Figure 1-left. This phenomenon suggests that existing methods (*e.g.*, ULIP [45] and OpenShape [23]) are not optimally leveraging the 2D image modality in 3D representation pre-training.

There are two reasons for this *counter-intuitive* observation: 1) The 2D images in the generated triplets follow a data distribution *different from natural images*, on which CLIP is pre-trained. The images are rendered or projected from 3D to 2D, usually lacking a realistic background and texture. Since the image domain is shifted, the image features are no longer well-aligned with the text features, as exemplified in Figure 1-middle. 2) Intuitively, the *image* features represent more *visual* attributes including the shape, texture, or color, while the *text* features have a focus on *semantics* such as the function of the object. As shown in Figure 1-right, if the 3D features are specifically aligned with one single modality of images or text, they pay attention to different aspects of the 3D shape. Therefore, enforcing the 3D shape to simultaneously align with two modalities that convey subtly distinct information could be challenging. Such an approach may not fully leverage the potential of multi-modal learning signals.

Aiming at addressing these two issues hindering multi-modal pre-training for 3D shape understanding, we propose **T**ri**A**dapter **M**ulti-**M**odal Learning (**TAMM**), a two-stage pre-training approach based on three synergistic adapters. In the first stage, to mitigate the domain gap between the 2D images rendered from 3D shapes and the natural images on which CLIP is pre-trained, we adapt the visual representations of CLIP based on the synthetic image-text pairs. More specifically, we fine-tune a lightweight CLIP Image Adapter on top of CLIP visual encoder through contrastive learning and re-align the adapted image features with the text features in the new domain. This CLIP Image Adapter allows us to establish more accurate relations between 3D shapes, 2D images, and language in an updated feature space, and avoids learning 3D representations from mismatched image features and text features.

In the following second stage, to prevent the vision-semantic feature disparity from impairing our 3D representation pre-training, we choose to embrace this disparity and decouple the 3D representations into two sub-spaces. To comprehensively encode a 3D shape, the 3D encoder needs to capture both the visual and semantic aspects of its representation, which are centric to the corresponding image and text features, respectively. Therefore, we decouple the 3D feature space for the two focuses on visual and semantic representations. In particular, we attach two independent feature adapters to the 3D backbone and transform the 3D features into two sub-spaces. One sub-space focuses more on the visual representations, and is aligned with the 2D image feature space; the other sub-space focuses more on the semantic representations, and is aligned with the language feature space. This approach of decoupled feature spaces makes the learned 3D representations more comprehensive and expressive.

To summarize, our main contribution includes:

- We identify the under-utilization of the 2D image modality in existing multi-modal methods. The image domain gap and feature disparity in image-text pairs hinder representation learning in 3D shape understanding.
- We propose a novel multi-modal learning framework with two learning stages and three unified adapter modules. Our proposed TAMM better exploits both image and language modalities and improves 3D shape representations.
- Our TAMM consistently enhances 3D representations for a variety of 3D encoder architectures (*e.g.*, Point-BERT [47], SparseConv [6]), pre-training datasets (*e.g.*, ShapeNet [3], an ensembled dataset [23]), and downstream tasks (*e.g.*, zero-shot and linear probing shape classification on Objaverse-LVIS [9], ModelNet40 [43], and ScanObjectNN [40]).

2. Related Work

3D Shape Understanding. There are two mainstreams for 3D shape representation learning: 1) Projecting 3D shapes into voxel or grid-based formats [37] and then using 2D/3D convolutions [6] for feature extraction. 2) Directly modeling 3D point clouds with point-centric architectures [24, 28, 30, 31, 33, 42, 50, 52]. In this work, to ensure a fair and comprehensive comparison with previous methods [23, 45] on the pre-training scheme, we follow

their selection and utilize two representative 3D encoders from these two mainstreams: SparseConv [6] and Point-BERT [47]. SparseConv is designed for efficiently processing sparse voxels using specialized convolutions that focus computations on non-zero data points. Point-BERT [42] utilizes a Transformer-based architecture [41] and can be self-supervised by masked modeling [10].

Multi-Modal Representation Learning. Contrastive Language-Image Pre-training (CLIP) [34] has enabled various downstream applications including object detection [48, 53, 55] and language grounding [22]. Recently, CLIP has been extended to 3D-based tasks, such as zero-shot text-to-3D generation [17, 20, 25, 27, 38, 44] and scene-level 3D segmentation [14, 29, 35, 46]. Meanwhile, developing general and robust representations for 3D shape understanding with the foundation of CLIP [19, 23, 45, 51, 54, 56] becomes a major focus. Among these methods, ULIP [45], as a pioneering work, utilizes contrastive learning to distill CLIP features into 3D representations. OpenShape [23] follows this learning paradigm with a focus on building a larger pre-training dataset with enriched and filtered text data. Unlike OpenShape [23], we focus on improving the multi-modal learning paradigm by more effectively leveraging both the image and text modalities via a two-stage pre-training approach based on three synergistic adapters.

3. Method

Figure 2 shows our proposed two-stage approach, **T**ri**A**dapter **M**ulti-**M**odal Learning (**TAMM**), for pre-training robust and generalizable 3D representations by leveraging 2D images and text. We first revisit the triplet data generation and problem formulation in multi-modal 3D representation learning. Section 3.1 delves into our fine-tuning of the original CLIP model for better fitting 3D understanding tasks. Section 3.2 details our strategy to enhance alignment across the 3D, 2D, and text feature spaces.

Problem Formulation. Given n triplets $\{(P_i, I_i, T_i)\}_{i=1}^{n}$, where P_i is a 3D point cloud, I_i represents the corresponding image produced by projecting the 3D point cloud P_i into 2D from an arbitrary perspective, and T_i denotes the associated text generated using advanced vision-language models such as BLIP [21], the objective is to learn high-quality 3D representations from the triplets. The basic framework to achieve this objective is proposed by ULIP [45] (and followed by OpenShape [23]) with the help of CLIP [5, 34].

Formally, the 3D feature $f_i^P = E_P(P_i)$ is produced by a learnable 3D encoder E_P, and the corresponding image feature $f_i^I = E_I(I_i)$ and text feature $f_i^T = E_T(T_i)$ are generated by frozen CLIP encoders E_I, E_T. Then the 3D encoder E_P is optimized by aligning the 3D feature space f^P simultaneously to the pre-aligned CLIP image space f^I and text space f^T through contrastive learning. The corresponding contrastive loss $L_{\text{contrast}}(f^{M_1}, f^{M_2})$ between two modalities M_1, M_2 (3D-2D or 3D-text) is formulated as:

$$-\frac{1}{2n}\sum_{i=1}^{n}\left(\log\frac{\exp(f_i^{M_1}\cdot f_i^{M_2}/\tau)}{\sum_{j=1}^{n}\exp(f_i^{M_1}\cdot f_j^{M_2}/\tau)}+\log\frac{\exp(f_i^{M_2}\cdot f_i^{M_1}/\tau)}{\sum_{j=1}^{n}\exp(f_i^{M_2}\cdot f_j^{M_1}/\tau)}\right), \tag{1}$$

where τ denotes the temperature hyperparameter.

3.1. Image-Text Re-Alignment

Unlike ULIP [45] or OpenShape [23], in which the 3D feature space is aligned with the image-text feature spaces *directly produced by CLIP*, we introduce an image feature space tuning strategy, aiming to *foster better alignment* across 3D, 2D, and language modalities (Figure 2-left). We argue that the image feature space produced directly by the CLIP visual encoder does not perfectly align with the text feature space and is thus sub-optimal. The reason is that the 2D images in the triplets, which originate from 3D point cloud projections, lack backgrounds and are visually different from natural images on which CLIP is pre-trained.

Therefore, when using such 2D images from a shifted data domain to perform multi-modal pre-training, it becomes necessary to further fine-tune CLIP and re-align its image and language feature spaces. For the first time, we redesign domain adapters [18] for multi-modal contrastive 3D representation learning, and propose a **C**LIP **I**mage **A**dapter (**CIA**), to adapt the image feature space for the rendered 2D images in our data triplets. We append a lightweight, learnable multi-layer perceptron (MLP) to CLIP image encoder, with a residual connection which seamlessly integrates the new knowledge acquired from fine-tuning with the existing knowledge from the pre-trained CLIP backbone. To avoid heavy computation and over-fitting, we only fine-tune the additional parameters in CIA, instead of the whole CLIP backbone. Formally, given the image feature f_i^I and text feature f_i^T extracted from the triplet (P_i, T_i, I_i), we use a learnable, two-layer CIA $A_C(\cdot)$ to adapt the image feature, formulated as:

$$A_C(f_i^I) = \sigma(f_i^I W_1)\cdot W_2, \tag{2}$$

where W_1 and W_2 are the parameters associated with the linear transformation layers, and σ is the non-linear activation function. The refined image feature $\widetilde{f}_i^I$ can be computed with a residual connection:

$$\widetilde{f}_i^I = \alpha A_C(f_i^I) + (1-\alpha) f_i^I, \tag{3}$$

where α is a hyperparameter. Finally, CIA $A_C(\cdot)$ is optimized by minimizing the contrastive loss function (Equation 1), instantiated as:

$$\mathcal{L}_{\text{realign}} = L_{\text{contrast}}(\widetilde{f}^I, f^T). \tag{4}$$

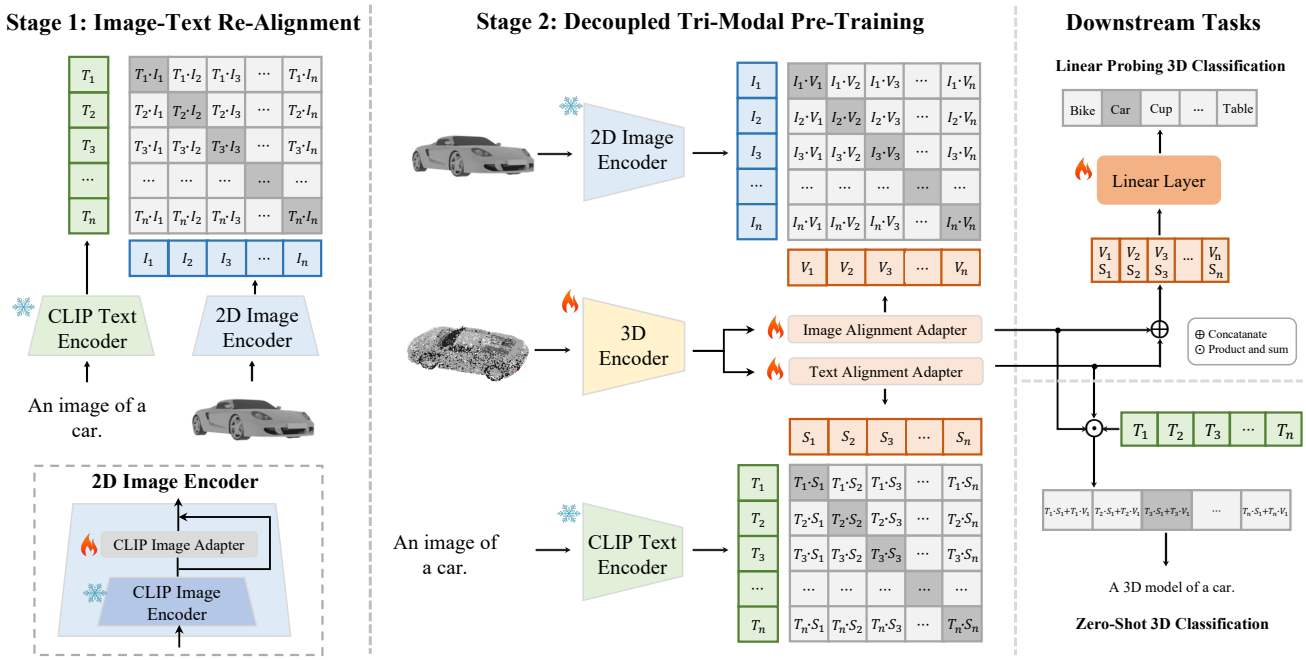

Figure 2. **Overview of TAMM. Left:** In Stage 1, TAMM fine-tunes a lightweight CLIP Image Adapter (CIA) through contrastive learning and re-aligns the image features with the text features to alleviate the domain shift originated from rendered images. Contrastive learning maximizes inner products between features from corresponding text-image pairs, and reduces similarities of mismatched pairs. **Middle:** In Stage 2, TAMM introduces Image Alignment Adapter (IAA) and Text Alignment Adapter (TAA) to decouple 3D representations into two sub-spaces: one focusing more on visual attributes and the other for semantic understanding, ensuring a more comprehensive and effective multi-modal pre-training strategy. **Right:** TAMM adaptively utilizes decoupled 3D features for various downstream tasks including linear probing classification (top) and zero-shot classification (bottom), achieving more robust classification results.

3.2. Decoupled Tri-Modal Pre-Training

Although the image and language feature spaces have been aligned by CLIP and our CLIP Image Adapter, they still encode *subtly different information*. For instance, the color of a 3D shape may be inferred from the image feature, but it cannot possibly be included in the text feature if the text description does not contain color information. Similarly, the image feature may not include semantic information such as the function or name of an object. In such cases, enforcing the 3D feature to align with both the image and text features *simultaneously* is challenging.

In order to overcome the obstacle introduced by aligning the 3D shape feature space with two distinct modalities, we propose a novel decoupled tri-modal pre-training framework, which aligns two decoupled 3D shape feature spaces with the refined image feature space and the text feature space, respectively (Figure 2-middle). We encourage the 3D encoder to cover both the *visual* representation and the *semantic* information inherent to the 3D shape, avoiding the dilemma of aligning with disparate image and text features at the same time. Formally, given the triplets $\{(P_i, I_i, T_i)\}_{i=1}^{n}$, we first generate the image feature $\widetilde{f}_i^I$ using the frozen, adapted 2D image encoder (Equation 3) and the text feature f_i^T using the original CLIP text encoder. We introduce a pair of lightweight Dual Adapters: **I**mage **A**lignment **A**dapter (**IAA**) $A_V(\cdot)$ and **T**ext **A**lignment **A**dapter (**TAA**) $A_S(\cdot)$. They split the 3D feature f_i^P originated from the 3D encoder E_P into a **v**ision-focusing feature f_i^{VP} and a **s**emantic-focusing feature f_i^{SP}, respectively. Dual Adapters $A_V(\cdot)$ and $A_S(\cdot)$ share the same architecture as the CLIP Image Adapter in Section 3.1 and can be formulated as:

$$\begin{aligned} f_i^{VP} &= A_V(f_i^P) = \sigma(f_i^P W_1^V) \cdot W_2^V, \\ f_i^{SP} &= A_S(f_i^P) = \sigma(f_i^P W_1^S) \cdot W_2^S, \end{aligned} \tag{5}$$

where W_1^V, W_2^V, W_1^S, and W_2^S are the parameters associated with the linear layers, and σ represents the activation function. By decoupling 3D features into these sub-spaces, the 3D encoder E_P interprets the 3D shape with a more comprehensive visual and semantic understanding, improving the expressivity of the learned representations.

Finally, instead of enforcing the 3D shape feature space to directly mimic the pre-aligned image-text feature space, we use the decoupled 3D features f^{VP}, f^{SP} and align them with the adapted image feature space $\widetilde{f}^I$ and the text feature space f^T, respectively. Moreover, since one single image can only capture the 3D shape from one perspective, we align the 3D feature with the adapted features of *multi-view images*, to fully exploit the image modality and achieve a

better alignment between 3D and 2D. The overall loss function is defined as:

$$\mathcal{L}_{\text{trimodal}} = L_{\text{contrast}}(f^{SP}, f^{T}) + \frac{1}{m}\sum_{k}^{m} L_{\text{contrast}}(f^{VP}, \widetilde{f}^{I,k}), \quad (6)$$

where m represents the number of rendered images and $\widetilde{f}^{I,k}$ is the adapted image feature from the k-th view.

Application in Downstream Tasks. The learned 3D feature sub-spaces, f^{VP} and f^{SP} can be adaptively applied to a variety of downstream tasks (Figure 2-right). Specifically, in zero-shot 3D classification, we leverage both the 3D vision-focusing feature f^{VP} and the 3D semantic-focusing feature f^{SP}. We calculate the similarity between these features and category embeddings generated by the CLIP text encoder, respectively. After summing up the per-category similarity scores, the category with the highest similarity is chosen as the predicted class, yielding more robust and enhanced classification results compared with using a single sub-space alone. In the linear probing classification task, we concatenate the 3D vision-focusing feature f^{VP} and 3D semantic-focusing feature f^{SP} as input to the learnable linear classification layer, providing a more comprehensive and robust representation for 3D understanding.

4. Experiments

Pre-Training Datasets. Following the prior state-of-the-art method, OpenShape [23], our TAMM is pre-trained on the triplets generated from four datasets: ShapeNetCore [3], 3D-FUTURE [12], ABO [7], and Objaverse [9]. Our training sets are defined as follows: "ShapeNet" is a triplet set derived exclusively from the ShapeNetCore dataset, containing $52,470$ 3D shapes with corresponding images and text; "Ensembled (no LVIS)" is a set of $829,460$ triplets from the above datasets excluding Objaverse-LVIS; "Ensembled" denotes the triplet set comprising data from all four datasets, containing $875,665$ 3D shapes and their associated images and text.

Evaluation Datasets. Our TAMM is evaluated on the following datasets: Objaverse-LVIS [9], ModelNet40 [43], ScanObjectNN [40], and ScanNet [8]. Objaverse-LVIS encompasses a broad range of categories, featuring $46,832$ high-quality shapes distributed across $1,156$ LVIS [13] categories. ModelNet40 is a synthetic indoor 3D dataset comprising 40 categories. We use the test split of $2,468$ shapes in our experiments. ScanObjectNN consists of scanned objects from 15 common categories. It has three main variants: OBJ-BG, OBJ-ONLY, and PB-T50-RS. ScanNet, characterized by its real-world scans, includes 1,513 indoor scenes containing 36,213 objects. We conduct experiments on four tasks including zero-shot 3D classification, linear probing 3D classification, few-shot linear probing 3D classification, and real-world recognition, to demonstrate the advantages of our TAMM. Other implementation details regarding pre-training and evaluation are introduced in the supplementary material.

4.1. Zero-Shot 3D Classification

Zero-shot ability is a key metric for reflecting the quality of learned 3D representations, which requires the representations to be *directly applicable to datasets where the model has never been explicitly supervised*. Without any further tuning, the 3D representations are compared with text embeddings of categories to predict the classes of 3D shapes. To make a fair comparison and keep consistency with prior work, we adopt the same settings as OpenShape [23] on three benchmarks: Objaverse-LVIS [9], ModelNet40 [43], and OBJ-ONLY (ScanObjectNN) [40].

The results are summarized in Table 1. First, we observe that TAMM, benefited from multi-modal pre-training, outperforms PointCLIP [51] and PointCLIP v2 [56] by a large margin. Furthermore, it is evident that our pre-trained models consistently outperform those pre-trained by ULIP and OpenShape, *irrespective of* whether they are pre-trained on smaller datasets like ShapeNet or more expansive ones such as the Ensembled dataset. For instance, Point-BERT [47] pre-trained by our TAMM on the Ensembled (no LVIS) dataset surpasses ULIP and OpenShape by margins of $+20.6\%$ and $+2.9\%$ in Top-1 accuracy on the long-tailed Objaverse-LVIS benchmark, which validates the effectiveness of our multi-modal pre-training scheme.

4.2. Linear Probing 3D Classification

To further evaluate the 3D representation quality of our pre-trained models, we design and conduct linear probing classification experiments. Here, we first freeze our pre-trained Point-BERT model along with the Image Alignment Adapter (IAA) and Text Alignment Adapter (TAA), and then append a single learnable linear layer to the model. Given a batch of point clouds, as illustrated in Figure 2, we generate features from both IAA and TAA and concatenate them, and learn the appended linear classification layer on the concatenated features. We evaluate the accuracy of the linear classifier on three benchmarks: Objaverse-LVIS [9], ModelNet40 [43], and ScanObjectNN [40]. Objaverse-LVIS is a challenging long-tailed dataset with $1,156$ categories, which has not been evaluated by prior methods [23, 45]. To ensure evaluation with statistically meaningful number of samples per class, we exclude categories with fewer than 10 instances, leaving us $1,046$ categories. Samples in each class are divided into training and testing sets at an $8:2$ ratio. For ModelNet40 and ScanObjectNN, we use the standard splits, following [32, 42].

The results are shown in Table 2. TAMM consistently outperforms all previous methods by a large margin. For example, pre-trained on the Ensembled dataset, TAMM improves over OpenShape by $+11.2\%$ and $+2.2\%$ overall ac-

Pre-Training Dataset	Model	Pre-Training Method	Objaverse-LVIS [9]			ModelNet40 [43]			ScanObjectNN [40]		
			Top-1	Top-3	Top-5	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5
-	PointCLIP [51]	-	1.9	4.1	5.8	19.3	28.6	34.8	10.5	20.8	30.6
-	PointCLIP v2 [56]	-	4.7	9.5	12.9	63.6	77.9	85.0	42.2	63.3	74.5
ShapeNet	ViT-B/32 [11]	CLIP2Point [19]	2.7	5.8	7.9	49.5	71.3	81.2	25.5	44.6	59.4
	Transformer [41]	ReCon [32]	1.1	2.7	3.7	61.2	73.9	78.1	42.3	62.5	75.6
	SparseConv [6]	OpenShape [23]	11.6	21.8	27.1	72.9	87.2	93.0	52.7	72.7	83.6
		TAMM (Ours)	13.6	24.2	29.3	74.6	88.2	94.0	57.9	75.3	83.1
	Point-BERT [47]	ULIP [45]	6.2	13.6	17.9	60.4	79.0	84.4	51.5	71.1	80.2
		OpenShape [23]	10.8	20.2	25.0	70.3	86.9	91.3	51.3	69.4	78.4
		TAMM (Ours)	13.7	24.2	29.2	73.1	88.5	91.9	54.8	74.5	83.3
Ensembled (no LVIS)	SparseConv [6]	OpenShape [23]	37.0	58.4	66.9	82.6	95.0	97.5	54.9	76.8	87.0
		TAMM (Ours)	39.8	62.0	70.4	85.7	**96.8**	**98.3**	57.5	81.3	90.0
	Point-BERT [47]	ULIP [45]	21.4	38.1	46.0	71.4	84.4	89.2	46.0	66.1	76.4
		OpenShape [23]	39.1	60.8	68.9	85.3	96.2	97.4	47.2	72.4	84.7
		TAMM (Ours)	42.0	63.6	71.7	**86.3**	96.6	98.1	56.7	78.3	86.1
Ensembled	SparseConv [6]	OpenShape [23]	43.4	64.8	72.4	83.4	95.6	97.8	56.7	78.9	88.6
		TAMM (Ours)	43.8	66.2	74.1	85.4	96.4	98.1	**58.5**	**81.3**	**89.5**
	Point-BERT [47]	ULIP [45]	26.8	44.8	52.6	75.1	88.1	93.2	51.6	72.5	82.3
		OpenShape [23]	46.8	69.1	77.0	84.4	96.5	98.0	52.2	79.7	88.7
		TAMM (Ours)	**50.7**	**73.2**	**80.6**	85.0	96.6	98.1	55.7	80.7	88.9

Table 1. **Zero-shot 3D classification results.** TAMM sets new state of the art in zero-shot classification accuracy across Objaverse-LVIS, ModelNet-40, and ScanObjectNN benchmarks, outperforming existing methods in diverse settings of pre-training datasets and 3D model architectures. The performance gain is more significant on the most challenging long-tailed Objaverse-LVIS dataset.

Pre-Training Dataset	Method	O-LVIS [9]	M-40 [43]	ScanObjectNN [40]		
				OBJ-BG	OBJ-ONLY	PB-T50-RS
ShapeNet	ULIP [45]	34.6	90.6	75.4	75.4	64.8
	OpenShape [23]	29.3	88.5	77.8	78.5	64.1
	TAMM (Ours)	39.1	91.0	80.6	81.1	68.5
	Rel. Improv.	+9.8	+2.5	+2.8	+2.6	+4.4
Ensembled	OpenShape [23]	48.3	91.3	85.9	85.4	78.0
	TAMM (Ours)	**59.5**	**93.5**	**88.5**	**88.0**	**80.3**
	Rel. Improv.	+11.2	+2.2	+2.6	+2.6	+1.7

Table 2. **Linear probing 3D classification results.** TAMM outperforms previous methods by a large margin, *e.g.*, $+11.2\%$ accuracy gain on the challenging Objaverse-LVIS dataset.

Pre-Training Dataset	Method	5-way		10-way	
		10-shot	20-shot	10-shot	20-shot
ShapeNet	ULIP [45]	$94.4_{\pm 3.7}$	$93.2_{\pm 4.2}$	$86.6_{\pm 5.3}$	$90.6_{\pm 5.2}$
	OpenShape [23]	$95.3_{\pm 2.6}$	$97.9_{\pm 3.9}$	$89.2_{\pm 5.1}$	$92.9_{\pm 3.9}$
	TAMM (Ours)	$97.8_{\pm 1.9}$	$98.1_{\pm 1.4}$	$95.3_{\pm 3.9}$	$95.9_{\pm 2.6}$
Ensembled	OpenShape [23]	$96.1_{\pm 2.7}$	$95.7_{\pm 2.5}$	$89.1_{\pm 4.6}$	$91.8_{\pm 3.7}$
	TAMM (Ours)	$\mathbf{99.0_{\pm 1.3}}$	$\mathbf{99.4_{\pm 0.7}}$	$\mathbf{96.8_{\pm 2.9}}$	$\mathbf{97.4_{\pm 2.2}}$

Table 3. **Few-shot linear probing classification results on ModelNet40.** We report the average accuracy and standard deviation of 10 independent experiments. Our TAMM consistently achieves both the best average accuracy and the lowest variance in various few-shot settings.

curacy on Objaverse-LVIS and ModelNet40, respectively. This demonstrates that TAMM effectively exploits knowledge from CLIP and learns generalizable 3D representations that are *directly applicable to novel linear classification tasks*. Notably, TAMM even surpasses the Point-BERT model pre-trained without multi-modality but allowed to be *fully fine-tuned* on OBJ-BG (87.4%) [47] by $+1.1\%$ accuracy. These strong results indicate that TAMM has learned excellent transferable 3D representations, showing a great potential in real-world applications.

4.3. Few-Shot Linear Probing 3D Classification

Following the previous evaluation, we perform few-shot classification experiments on ModelNet40 [43] to assess TAMM in low-data scenarios. Similar to the linear probing experiments described in Section 4.2, we extend our model with an additional linear classification layer and only train this linear layer instead of fine-tuning the entire model. Following previous work [47, 50], we adopt the "K-way N-shot" configuration, wherein we select K classes at random and sample $(N + 20)$ instances per class. Training is conducted on a *support set* of $K \times N$ samples, while evaluation is based on a *query set* comprised of the remaining 20 instances per class. We assess our model under four distinct scenarios: "5-way 10-shot," "5-way 20-shot," "10-way 10-shot," and "10-way 20-shot." For each scenario, we carry out 10 separate trials, and report both the mean performance and the standard deviation across these trials.

As illustrated in Table 3, our TAMM achieves the best results and sets a new state of the art in the few-shot clas-

Method	Avg.	Bed	Cab	Chair	Sofa	Tabl	Door	Wind	Bksf	Pic	Cntr	Desk	Curt	Fridg	Bath	Showr	Toil	Sink
CLIP2Point [19]	24.9	20.8	0.0	**85.1**	43.3	26.5	**69.9**	0.0	20.9	1.7	31.7	27.0	0.0	1.6	46.5	0.0	22.4	25.6
PointCLIP w/ TP. [56]	26.1	0.0	55.7	72.8	5.0	5.1	1.7	0.0	77.2	0.0	0.0	51.7	0.3	0.0	0.0	40.3	85.3	**49.2**
CLIP2Point w/ TP. [19]	35.2	11.8	3.0	45.1	27.6	10.5	61.5	2.6	71.9	0.3	**33.6**	29.9	4.7	**11.5**	**72.2**	**92.4**	**86.1**	34.0
CLIP^2 [49]	38.5	32.6	**67.2**	69.3	42.3	18.3	19.1	4.0	62.6	1.4	12.7	52.8	40.1	9.1	59.7	41.0	71.0	45.5
OpenShape† [23]	45.6	**66.7**	3.2	75.8	83.5	37.7	49.2	47.5	64.9	48.2	1.9	**66.1**	70.2	1.8	50.0	57.1	45.2	7.1
TAMM (Ours)†	**49.4**	**66.7**	4.8	83.6	**84.5**	**48.9**	57.9	**48.2**	**80.5**	**61.3**	1.9	60.6	**83.6**	7.0	41.4	56.1	48.4	3.6

w/ TP. denotes training with the real-world data provided by CLIP^2. † Results using Point-BERT [47] as 3D encoder, pre-trained on the Ensembled dataset.

Table 4. **Zero-shot classification results on the real-world ScanNet dataset. Avg.**: the mean average Top-1 accuracy of all classes. TAMM achieves the best results.

CIA	IAA	TAA	Img-Txt Acc	Objaverse-LVIS Top-1	Objaverse-LVIS Top-3	Objaverse-LVIS Top-5	ModelNet-40 Top-1	ModelNet-40 Top-3	ModelNet-40 Top-5
	✓	✓	40.1	13.5	23.9	29.1	72.8	88.2	91.7
✓			60.9	12.5	22.4	27.3	71.4	86.0	89.8
✓		✓	60.9	12.9	22.9	27.7	**73.8**	88.2	**92.5**
✓	✓		60.9	13.0	23.0	28.2	72.9	87.8	90.4
✓	✓	✓	60.9	**13.7**	**24.2**	**29.2**	73.1	**88.5**	91.9

(a) **Adapters in pre-training.** Pre-training with CLIP Image Adapter (CIA), Image Alignment Adapter (IAA), and Text Alignment Adapter (TAA) is the most effective.

Stage	Img-Txt Acc	Objaverse-LVIS Top-1	Objaverse-LVIS Top-3	Objaverse-LVIS Top-5	ModelNet-40 Top-1	ModelNet-40 Top-3	ModelNet-40 Top-5
One stage	55.3	12.2	21.4	27.0	71.4	86.4	90.7
Two stages	60.9	**13.7**	**24.2**	**29.2**	**73.1**	**88.5**	**91.9**

(b) **Pre-training stages.** Learning the CLIP Image Adapter first and then pre-train the 3D encoder with Dual Adapters is better than training modules all together.

Image	Text	Objaverse-LVIS Top-1	Objaverse-LVIS Top-3	Objaverse-LVIS Top-5	ModelNet-40 Top-1	ModelNet-40 Top-3	ModelNet-40 Top-5
✓		11.9	20.8	25.9	67.6	87.3	**92.5**
	✓	10.5	19.2	23.7	70.2	85.0	88.0
✓	✓	**13.7**	**24.2**	**29.2**	**73.1**	**88.5**	91.9

(c) **Pre-training modalities.** Exploiting both image and text data provides the most performance gain. They contribute almost equally in TAMM.

IAA	TAA	Objaverse-LVIS Top-1	Objaverse-LVIS Top-3	Objaverse-LVIS Top-5	ModelNet-40 Top-1	ModelNet-40 Top-3	ModelNet-40 Top-5
✓		13.0	23.1	28.3	68.1	86.5	91.4
	✓	12.9	22.5	27.5	72.4	86.2	90.4
✓	✓	**13.7**	**24.2**	**29.2**	**73.1**	**88.5**	**91.9**

(d) **Adapters for inference.** Dual Adapters are complementary, and their combination more accurately classifies 3D shapes at test time.

Images	Objaverse-LVIS Top-1	Objaverse-LVIS Top-3	Objaverse-LVIS Top-5	ModelNet-40 Top-1	ModelNet-40 Top-3	ModelNet-40 Top-5
1	13.6	23.6	28.9	72.5	88.0	92.6
2	13.5	23.5	28.8	73.1	87.6	**92.8**
4	**13.7**	**24.2**	**29.2**	**73.1**	**88.5**	91.9
8	13.5	23.1	28.4	73.0	89.0	92.8

(e) **Number of images.** Aligning a 3D shape with 4 images of it is the best.

Table 5. **Ablation study of components in TAMM.** We pre-train Point-BERT models on ShapeNet in various settings and evaluate their zero-shot classification performance on Objaverse-LVIS and ModelNet40. The baseline method OpenShape [23] achieves 10.8% and 70.3% Top-1 accuracy on these two benchmarks (Table 1), respectively. The setting adopted by TAMM is marked.

sification task. TAMM shows *both higher overall accuracy and smaller deviations* than other methods, showcasing the generalizability and robustness of TAMM-learned 3D representations. For example, when pre-trained on the Ensembled dataset, our TAMM surpasses OpenShape [23] by 2.9%, 3.7%, 7.7%, 5.6%, respectively for all four settings. These results indicate that our TAMM is able to learn 3D representations that are more generalizable and can be readily adapted to downstream tasks under *low-data regimes*.

4.4. Real-World Recognition

To evaluate the capability of TAMM in understanding *3D shapes from the real world*, we follow the previous work CLIP^2 [49] and test TAMM on a real-world recognition task, in which the model aims at correctly classifying each instance from a complex scene in a *zero-shot* manner. Specifically, we select the real-world scene-level dataset ScanNet [8] and adopt same data splits as [49], containing 17 classes. We perform zero-shot classification (Section 4.1) on the point cloud of each object instance extracted from scenes, and report both the Top-1 accuracy of each class and the mean accuracy of all 17 classes. As shown in Table 4, TAMM significantly outperforms all prior methods, showing improvements of 3.8% and 10.9% in average Top-1 accuracy compared with OpenShape and CLIP^2, respectively. These results underscore TAMM's ability in recognizing and understanding 3D shapes captured from real-world scenarios. Further results on complex scene recognition and instance segmentation are included in the supplementary material.

4.5. Ablation Study

In this section, we provide additional experimental results to further test the performance gain from each design of TAMM. Due to limited computation, we *pre-train* Point-BERT [47] on the *small-scale* ShapeNet [3], and use two *more challenging* zero-shot classification benchmarks ModelNet40 [43] and Objaverse-LVIS [9] for *evaluation*. We alter only one component in each set of experiments.

Adapters in Pre-Training. As described in Section 3.1, our CLIP Image Adapter (CIA) mitigates the domain gap between rendered images and natural images. We first measure the contrastive accuracy, computed as the ratio of image-text pairs where the text feature is more similar to the image from the same triplet than any other text. The contrastive accuracy of CLIP without CIA is only 40.1%,

indicating that the original image-text feature space is compromised by the domain gap. CIA brings the accuracy up to 60.9%, showing the effectiveness of our CIA in resolving the shifted data domain. Furthermore, as shown in Table 5a, CIA increases the zero-shot accuracy on Objaverse-LVIS dataset from 12.2% to 13.7%, validating its effectiveness on improving the pre-trained 3D representations.

TAMM adopts Image Alignment Adapter (IAA) and Text Alignment Adapter (TAA) to decouple the feature space and enrich the learned 3D representations. To further explore the performance gains introduced by our Dual Adapters (IAA and TAA), we experiment tri-modal pre-training with or without IAA and TAA. As shown in Table 5a, both adapters bring performance gains, and their combination achieves the best result, demonstrating the effectiveness of Dual Adapters.

Pre-Training Stages. We investigate the significance of our two-stage pre-training design, which initially learns CLIP Image Adapter (CIA) to re-align image-text pairs and subsequently employs Dual Adapters in decoupled tri-modal pre-training. An alternative could be learning these modules all together in one stage. As shown in Table 5b, remarkably, the two-stage pre-training approach achieves a better performance compared with one-stage pre-training.

Pre-Training Modalities. We further explore the effectiveness of image and language modalities in pre-training. Table 5c shows that integrating the 3D modality with both the image and language modalities consistently yields superior performance across various benchmarks, compared with aligning it with either modality alone. Moreover, unlike prior methods which ineffectively learn from images (Figure 1-left), the image and language modalities contribute almost equally in TAMM, which also signifies that TAMM better exploits the image modality and brings more gains by learning from both image and text data.

Adapters for Inference. Our IAA and TAA decouple 3D features with visual and semantic focuses, respectively, offering a more accurate and comprehensive understanding for 3D shapes. At inference time, we combine the outputs from Dual Adapters for classification tasks. To investigate this design, we separately utilize features produced solely by either IAA or TAA in zero-shot classification evaluation. As shown in Table 5d, the combined features achieve more accurate classification, demonstrating that the learned features from IAA and TAA are complementary to each other.

Number of Images. In order to gain more comprehensive knowledge from the image modality, TAMM simultaneously aligns the 3D feature with multi-view 2D images, each projected from a different perspective. We evaluate this design by aligning the 3D feature with a varying number of corresponding 2D images, ranging from 1 to 8 images. As shown in Table 5e, multi-view image features offer a performance gain by fully exploiting the image modality.

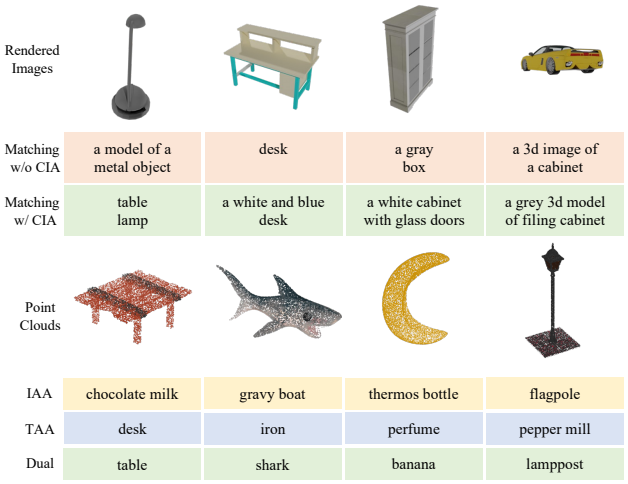

Figure 3. **Qualitative results. Top**: CIA re-aligns the images rendered from 3D shapes with the text descriptions. The rendered images are inaccurately matched with text when the image features are directly extracted by CLIP, and CIA can correct the matching. **Bottom:** IAA and TAA decouple 3D features 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.

Aligning a 3D shape with 4 views achieves the best result.

Qualitative Results. Finally, we provide some visualizations to intuitively demonstrate the benefit of our proposed TAMM. As illustrated in Figure 3, CIA successfully mitigates the domain gap introduced by rendered images, and leads to the more accurate matching between images and text. IAA and TAA learn 3D representations with subtly different but complementary focuses on vision and semantics, respectively, and their combination brings more robust and comprehensive 3D representations. More visualized results are included in the supplementary material.

5. Conclusion

In this work, we examine the sub-optimal utilization of 2D images in existing multi-modal pre-training methods for 3D shape understanding, and propose TriAdapter Multi-Modal Learning (TAMM), a novel two-stage representation learning approach built on three synergistic adapter modules. Extensive experiments verify that TAMM consistently learns improved 3D features in various settings.

Limitations. Due to the limited computation resources, we are not able to perform pre-training on very large-scale 3D backbones with billions of parameters. The quality of the learned 3D representation could be further improved if the 3D backbones are scaled to a larger size.

Acknowledgement. This work was supported in part by NSF Grant 2106825, NIFA Award 2020-67021-32799, the Jump ARCHES endowment, and the IBM-Illinois Discovery Accelerator Institute. This work used NVIDIA GPUs at NCSA Delta through allocations CIS220014, CIS230012, and CIS230013 from the ACCESS program.

References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020. 2

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 2

[3] Angel X. Chang, Thomas A. Funkhouser, Leonidas J. Guibas, Pat Hanrahan, Qi-Xing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An information-rich 3D model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1, 2, 5, 7

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 2

[5] 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. 2, 3

[6] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4D spatio-temporal ConvNets: Minkowski convolutional neural networks. In *CVPR*, 2019. 2, 3, 6

[7] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F. Yago Vicente, Thomas Dideriksen, Himanshu Arora, Matthieu Guillaumin, and Jitendra Malik. ABO: Dataset and benchmarks for real-world 3D object understanding. In *CVPR*, 2022. 5

[8] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *CVPR*, 2017. 5, 7

[9] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3D objects. In *CVPR*, 2023. 2, 5, 6, 7

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 2, 3

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 6

[12] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3D-FUTURE: 3D furniture shape with texture. *IJCV*, 129:3313–3337, 2021. 5

[13] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 5

[14] Huy Ha and Shuran Song. Semantic abstraction: Open-world 3D scene understanding from 2D vision-language models. In *CoRL*, 2022. 3

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 2

[16] Deepti Hegde, Jeya Maria Jose Valanarasu, and Vishal Patel. CLIP goes 3D: Leveraging prompt tuning for language grounded 3D recognition. In *ICCV*, 2023. 1

[17] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. AvatarCLIP: Zero-shot text-driven generation and animation of 3D avatars. *ACM Transactions on Graphics*, 41(4):1–19, 2022. 3

[18] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *ICML*, 2019. 3

[19] Tianyu Huang, Bowen Dong, Yunhan Yang, Xiaoshui Huang, Rynson WH Lau, Wanli Ouyang, and Wangmeng Zuo. CLIP2Point: Transfer CLIP to point cloud classification with image-depth pre-training. In *ICCV*, 2023. 1, 3, 6, 7

[20] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *CVPR*, 2022. 3

[21] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 3

[22] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022. 3

[23] 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, 3, 5, 6, 7

[24] Xu Ma, Can Qin, Haoxuan You, Haoxi Ran, and Yun Fu. Rethinking network design and local geometry in point cloud: A simple residual MLP framework. In *ICLR*, 2022. 1, 2

[25] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2Mesh: Text-driven neural stylization for meshes. In *CVPR*, 2022. 3

[26] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su. PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In *CVPR*, 2019. 2

[27] Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa. CLIP-Mesh: Generating textured meshes from text using pretrained image-text models. In *SIGGRAPH Asia*, 2022. 3

[28] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *ECCV*, 2022. 2

[29] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. OpenScene: 3D scene understanding with open vocabularies. In *CVPR*, 2023. 3

[30] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *CVPR*, 2017. 2

[31] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*, 2017. 2

[32] Zekun Qi, Runpei Dong, Guofan Fan, Zheng Ge, Xiangyu Zhang, Kaisheng Ma, and Li Yi. Contrast with reconstruct: Contrastive 3D representation learning guided by generative pretraining. In *ICML*, 2023. 5, 6

[33] Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. PointNeXt: Revisiting PointNet++ with improved training and scaling strategies. In *NeurIPS*, 2022. 2

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 3

[35] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3D semantic segmentation in the wild. In *ECCV*, 2022. 3

[36] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS Datasets and Benchmarks Track*, 2022. 2

[37] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. PV-RCNN: Point-voxel feature set abstraction for 3D object detection. In *CVPR*, 2020. 2

[38] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. MotionCLIP: Exposing human motion generation to CLIP space. In *ECCV*, 2022. 3

[39] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[40] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *ICCV*, 2019. 2, 5, 6

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3, 6

[42] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point Transformer V2: Grouped vector attention and partition-based pooling. In *NeurIPS*, 2022. 1, 2, 3, 5

[43] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3D ShapeNets: A deep representation for volumetric shapes. In *CVPR*, 2015. 2, 5, 6, 7

[44] Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaohu Qie, and Shenghua Gao. Dream3D: Zero-shot text-to-3D synthesis using 3D shape prior and text-to-image diffusion models. In *CVPR*, 2023. 3

[45] 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, 2, 3, 5, 6

[46] Jihan Yang, Runyu Ding, Zhe Wang, and Xiaojuan Qi. RegionPLC: Regional point-language contrastive learning for open-world 3D scene understanding. *arXiv preprint arXiv:2304.00962*, 2023. 3

[47] 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, 3, 5, 6, 7

[48] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *CVPR*, 2021. 3

[49] Yihan Zeng, Chenhan Jiang, Jiageng Mao, Jianhua Han, Chaoqiang Ye, Qingqiu Huang, Dit-Yan Yeung, Zhen Yang, Xiaodan Liang, and Hang Xu. CLIP^2: Contrastive language-image-point pretraining from real-world point cloud data. In *CVPR*, 2023. 7

[50] Renrui Zhang, Ziyu Guo, Peng Gao, Rongyao Fang, Bin Zhao, Dong Wang, Yu Qiao, and Hongsheng Li. Point-M2AE: Multi-scale masked autoencoders for hierarchical point cloud pre-training. In *NeurIPS*, 2022. 2, 6

[51] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. PointCLIP: Point cloud understanding by CLIP. In *CVPR*, 2022. 1, 3, 5, 6

[52] Zaiwei Zhang, Rohit Girdhar, Armand Joulin, and Ishan Misra. Self-supervised pretraining of 3D features on any point-cloud. In *ICCV*, 2021. 2

[53] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. RegionCLIP: Region-based language-image pretraining. In *CVPR*, 2022. 3

[54] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3D: Exploring unified 3D representation at scale. In *ICLR*, 2024. 3

[55] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 3

[56] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, and Peng Gao. Point-

CLIP V2: Prompting CLIP and GPT for powerful 3D open-world learning. In *ICCV*, 2023. 1, 3, 5, 6, 7