MM-Mixing: Multi-Modal Mixing Alignment for 3D Understanding

Jiaze Wang[*2], **Yi Wang**[*1], **Ziyu Guo**[2], **Renrui Zhang**[2], **Donghao Zhou**[2], **Guangyong Chen**[3†], **Anfeng Liu**[1‡], **Pheng-Ann Heng**[2]

[1] Central South Unversity
[2] The Chinese University of Hong Kong
[3] Zhejiang Lab

Abstract

We introduce **MM-Mixing**, a multi-modal mixing alignment framework for 3D understanding. MM-Mixing applies mixing-based methods to multi-modal data, preserving and optimizing cross-modal connections while enhancing diversity and improving alignment across modalities. Our proposed two-stage training pipeline combines feature-level and input-level mixing to optimize the 3D encoder. The first stage employs feature-level mixing with contrastive learning to align 3D features with their corresponding modalities. The second stage incorporates both feature-level and input-level mixing, introducing mixed point cloud inputs to further refine 3D feature representations. MM-Mixing enhances intermodality relationships, promotes generalization, and ensures feature consistency while providing diverse and realistic training samples. We demonstrate that MM-Mixing significantly improves baseline performance across various learning scenarios, including zero-shot 3D classification, linear probing 3D classification, and cross-modal 3D shape retrieval. Notably, we improved the zero-shot classification accuracy on ScanObjectNN from 51.3% to 61.9%, and on Objaverse-LVIS from 46.8% to 51.4%. Our findings highlight the potential of multi-modal mixing-based alignment to significantly advance 3D object recognition and understanding while remaining straightforward to implement and integrate into existing frameworks.

Introduction

In the field of 3D vision, integrating multiple data modalities such as text, images, and point clouds has shown great potential for enhancing object recognition and scene understanding. This multi-modal approach is vital for applications in mixed reality (Dargan et al. 2023; Mendoza-Ramírez et al. 2023), autonomous navigation (Chen et al. 2020a; Tan, Robertson, and Czerwinski 2001) and 3D scene understanding (Armeni et al. 2016; Liu et al. 2021; Vu et al. 2022), where accurate 3D perception is crucial. Recent advancements in multi-modal learning have underscored their capability in this domain, with notable contributions from seminal works like PointCLIP (Zhang et al. 2022b; Zhu et al. 2023), CLIP^2 (Zeng et al. 2023), ULIP (Xue et al. 2023a,b), OpenShape (Liu et al. 2024), and TAMM (Zhang, Cao, and Wang 2024). These studies have demonstrated the effectiveness of leveraging text, images, and point clouds to improve 3D object recognition and understanding.

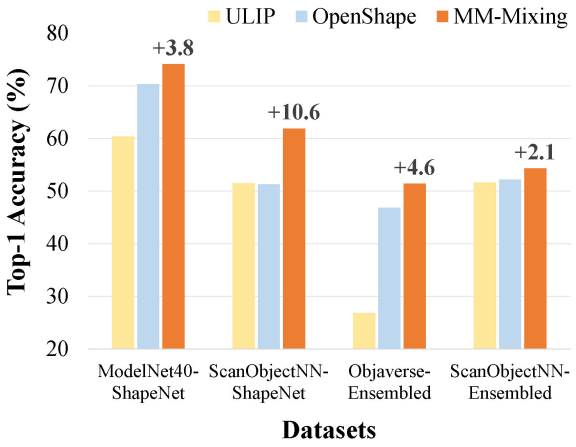

Figure 1: **Performance comparison with previous methods**. MM-Mixing achieves better performance than previous pre-training methods across various datasets with the same backbone Point-BERT. "ModelNet40-ShapeNet" represents the model is pretrained on ShapeNet and evaluated on ModelNet40, similarly for other dataset combinations.

However, a significant challenge remains in effectively aligning and utilizing these heterogeneous data sources to optimize model performance. With recent advancements in 3D vision, there's a growing emphasis on multi-modal learning approaches. These frameworks are becoming increasingly crucial, especially when it comes to processing and learning from multi-modal data, which integrates textual information, 2D images, and 3D point cloud data. Despite the success of these approaches, there is a notable gap in the literature regarding multi-modal data augmentation. The cohesive augmentation of triplets has the potential to unlock further performance improvements by enriching the diversity of data and promoting better alignment across modalities. This presents a promising avenue for research to explore comprehensively the benefits of multi-modal learning

[*]These authors contributed equally.
[†]Corresponding author: gychen@zhejianglab.com
[‡]Corresponding author: afengliu@csu.edu.cn

frameworks.

In previous studies, many mixing-based data augmentation methods have been proposed for point cloud (Kim et al. 2021; Rao et al. 2021; Lee et al. 2022). Mixing-based methods like PointCutMix (Zhang et al. 2022a) and PointMixup (Chen et al. 2020b) enhance training data diversity through techniques such as region splicing and feature interpolation. By introducing controlled perturbations and heterogeneity into the training process, these approaches enable models to learn invariant and discriminative features, thereby improving their robustness and generalization to diverse and unseen data distributions (Umam et al. 2022; Kim et al. 2021; Wang et al. 2024b).

However, the potential of mixing-based methods in multimodal scenarios remains largely unexplored. Integrating mixing-based techniques with multi-modal alignment could enhance multi-modal learning by generating diverse feature spaces, fostering robust cross-modal correspondences, and revealing invariant features across modalities. This leads to an important question: *Can we design a simple yet effective framework that improves alignment quality and stability while enhancing model generalization through augmented, coherent multi-modal representations?*

To address this issue, we introduce **MM-Mixing**, a multi-modal approach for 3D understanding that integrates mixing-based methods with multi-modal triplet data. Our two-stage training pipeline combines feature-level and input-level mixing to optimize the 3D encoder, enhancing intermodality relationships and promoting generalization. In the first stage, MM-Mixing leverages feature-level mixing and contrastive learning to align mixed 3D features with their corresponding modalities. This mixing-based alignment strategy fosters consistency across different modalities and significantly enhances the 3D encoder's cross-modal understanding. Specifically, by aligning point cloud mixed features with text mixed features, we capture semantic information that provides a contextual understanding of the 3D shapes. Additionally, aligning point cloud mixed features with image mixed features bolsters the capture of intricate visual details and spatial relationships. This dual alignment of mixed features not only ensures cross-modal consistency but also amplifies the 3D encoder's ability to understand and represent complex, multi-modal data effectively. The second stage incorporates feature-level and input-level mixing, introducing mixed point cloud inputs to refine 3D feature representations further. By aligning mixed point cloud features with feature-level mixed point cloud features, we enhance the network's ability to capture and represent variations and nuances within the data, resulting in more robust and discriminative feature representations. This stage generates diverse and realistic samples that enhance the 3D encoder's ability to generalize across different datasets.

By seamlessly integrating these methods, MM-Mixing significantly boosts the baseline model's performance across various settings, including zero-shot 3D classification, linear probing 3D classification, and cross-modal 3D shape retrieval, while remaining straightforward to implement and integrate into existing 3D understanding frameworks. Our main contributions can be summarized as follows:

- We introduce MM-Mixing, a novel multi-modal mixing alignment framework specifically designed for multi-modal data, addressing a previously unexplored issue in 3D understanding, which can be easily integrated with existing frameworks.
- An efficient two-stage framework is proposed that integrates feature-level and input-level augmentation to optimize the 3D encoder, enhance cross-modal relationships, and promote generalization.
- Our MM-Mixing not only strengthens the 3D understanding of models but also significantly enhances cross-dataset generalization, demonstrating exceptional performance in downstream tasks such as zero-shot 3D classification, linear probing 3D classification, and cross-modal retrieval.

Related Works

3D Understanding. Understanding 3D structures is a crucial aspect of computer vision (Peng et al. 2023; Qi et al. 2023). Mainly three representation learning methods have emerged: projecting-based methods where 3D point clouds are projected into various image planes (Goyal et al. 2021), voxel-based methods which transform the point clouds with 3D voxelization (Riegler, Osman Ulusoy, and Geiger 2017; Canfes et al. 2023), and direct modeling of 3D point clouds with point-centric architectures (Qian et al. 2022; Ma et al. 2022). These approaches highlight the use of specialized models like SparseConv (Choy, Gwak, and Savarese 2019) for efficiently handling sparse voxel data, and Transformer-based models (Zhang et al. 2023) such as Point-MAE (Pang et al. 2022) and Point-BERT (Yu et al. 2022) for leveraging self-supervised learning paradigms. Moreover, the integration of image-language models like CLIP (Radford et al. 2021) into 3D shape understanding represents a significant trend (Zhang, Cao, and Wang 2024; Zhu et al. 2024). Models are trained to align 3D shape embeddings with CLIP's language and/or image embeddings through multimodal contrastive learning (Huang et al. 2024). Notably, ULIP (Xue et al. 2023a,b), I2P-MAE (Zhang et al. 2023), and OpenShape (Liu et al. 2024) have refined the approach by optimizing the distillation of CLIP features into 3D representations and expanding training datasets for more generalizable learning outcomes.

3D Mixing-based Methods. Traditional techniques primarily involved simple transformations such as rotation, scaling, and jittering at the point level (Goyal et al. 2021). Recently, PointAugment (Li et al. 2020) optimizes both enhancer and classifier networks to generate complex samples, while techniques like Mixing-based augmentation (Chen et al. 2020b; Zhang et al. 2022a; Lee et al. 2021; Wang et al. 2024a) employ strategies from the 2D domain, such as optimal linear interpolation and rigid transformations, to mix multiple samples effectively. Furthermore, the advent of Transformer-based methods and attention mechanisms in point cloud processing has opened new possibilities for data augmentation. PointWOLF (Kim et al. 2021) introduces multiple weighted local transformations, and PointMixSwap (Umam et al. 2022) utilizes an attention-based method to swap divisions across point clouds, adding a layer of complexity and diversity.

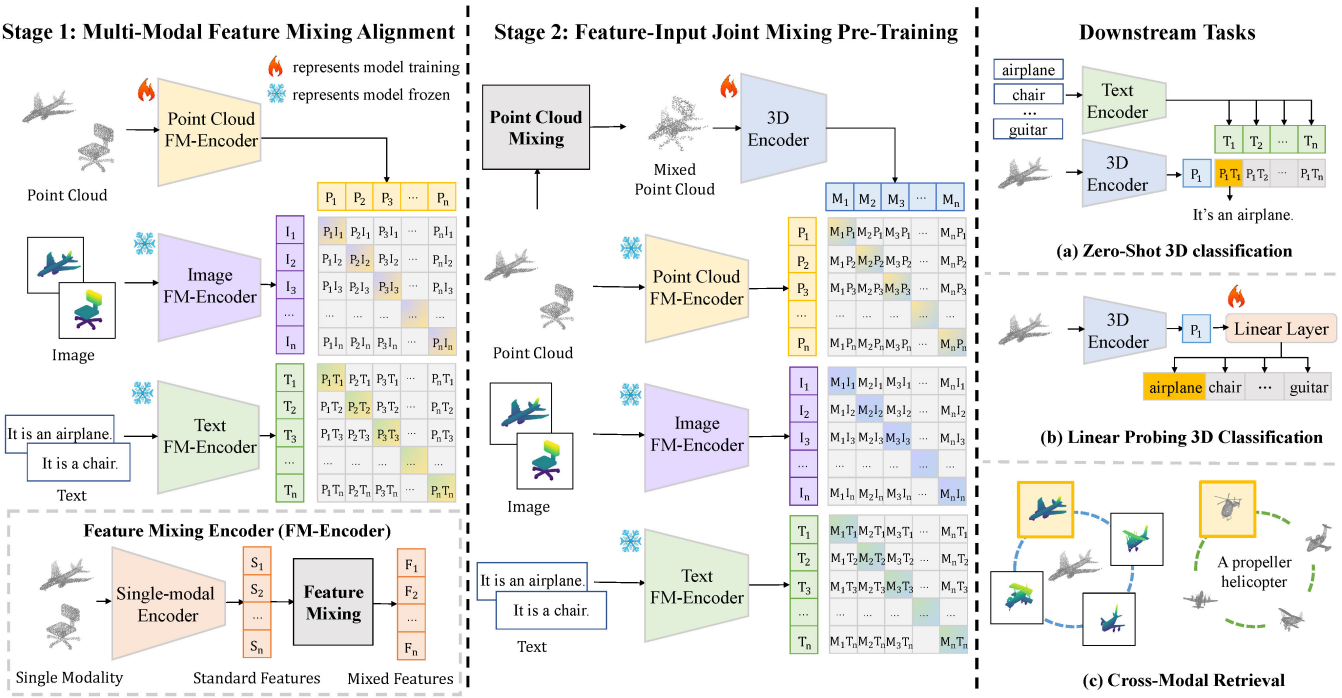

Figure 2: **The overall scheme of MM-Mixing.** MM-Mixing consists of two stages. In the first stage, the point cloud FM-Encoder is trainable, while the image and text FM-Encoders are pre-trained and frozen. Feature embeddings are extracted for contrastive learning with the 3D features. In the second stage, we initialize a new trainable 3D encoder. All FM-Encoders remain frozen. Two input point clouds are mixed using FPS and point-level mixing, and then fed into the 3D encoder. Then we adopt contrastive learning to align the features of mixed point clouds with mixed feature representations of all three modalities.

Method

The overall MM-Mixing pipeline is shown in Figure 2. We first review the problem definition to establish the context of our approach. Then, we introduce our mixing-based alignment strategy specifically designed for point clouds, images, and texts, which enhances the variability and robustness of the training data. Finally, we detail the MM-Mixing framework, demonstrating how our method integrates seamlessly into existing frameworks.

Problem Definition

Given a set of K triplets $\{(P_i, I_i, T_i)\}_{i=1}^{K}$, 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 (Li et al. 2022), the objective is to learn high-quality 3D representations from these triplets. Following ULIP (Xue et al. 2023a) and OpenShape (Liu et al. 2024) which leverage the CLIP (Radford et al. 2021) model, we enhance this framework by incorporating mixing-based methods. Specifically, the 3D features of the mixed point cloud $m_i^M = E_P(I_M(P_i, P_j))$ are obtained by passing two point clouds sequentially through the input-level mixing I_M and the 3D encoder E_P. The corresponding mixed features of the point cloud modality $m_i^P = F_M(E_P(P_i), E_P(P_j))$, the mixed features of the image modality $m_i^I = F_M(E_I(I_i), E_I(I_j))$, and the mixed features of the text modality $m_i^T = F_M(E_T(T_i), E_T(T_j))$ are generated by passing the features produced by the trained modality-specific encoders E_P, E_I and E_T through the feature-level mixing F_M, respectively. During the optimization of the 3D encoder E_P, contrastive learning is used to align the 3D features of the mixed point cloud m_i^M with the mixed features of the three modalities m_i^P, m_i^I, m_i^T.

Multi-Modal Mixing

We adopt two kinds of mixing methods for multi-model data, including feature-level mixing and input-level mixing.
Feature-level mixing. Feature-level mixing augments the features by combining features from two different inputs. This process involves first passing each input through the network independently to extract their respective features. Specifically, the first input is fed into the network, which processes it and extracts its feature vector f_i. Similarly, the second input is also passed through the network, resulting in the extraction of its feature vector f_j. Then the features are combined using a mixing operation to create a new, combined feature vector m_i, which can be expressed as:

$$m_i = \lambda f_i + (1 - \lambda) f_j. \tag{1}$$

Input-level mixing. For input-level mixing, we follow PointCutMix (Zhang et al. 2022a), which generates a new training point cloud $\tilde{p}$ from a pair of point clouds p_1 and p_2. The combination process of input-level augmentation is

defined as follows:

$$M = S \odot P_1 + (1 - S) \odot P_2, \tag{2}$$

$$\lambda = \sum S/N, \tag{3}$$

where M is the mixed point cloud, $S \in \{0, 1\}^N$ indicates which sample each point belongs to, $\odot$ represents element-wise multiplication, and λ is sampled from a beta distribution $Beta(\beta, \beta)$. This implies that $\lfloor \lambda N \rfloor$ points are selected from p_1, and $N - \lfloor \lambda N \rfloor$ points are selected from p_2.

Feature-level mixing operates on the encoded feature vectors, inducing implicit changes in the high-dimensional space. This allows for efficient data augmentation under cross-modal conditions, ensuring consistency of the augmented features across different modalities. In contrast, input-level augmentation directly manipulates the raw data, generating concrete and intuitive mixed samples. These realistic samples, which are both challenging and diverse, help the model better understand 3D shapes in downstream tasks. MM-Mixing combines these two augmentation strategies, achieving dual enhancement between raw data and latent features, thereby significantly improving the model's generalization ability.

MM-Mixing Framework

MM-Mixing refines feature representations through a combination of contrastive learning and mixing-based augmentation techniques, which improves the encoder's ability to generalize and discriminate between different classes through a two-stage training framework.

As shown in Figure 2, in the first stage, the point cloud Feature Mixing Encoder (FM-Encoder) is trainable, we freeze the image and text Feature Mixing Encoders (FM-Encoders), which are a combination of a single-modal encoder from CLIP (Radford et al. 2021) with a feature mixing module. Initially, point clouds are fed into the trainable point cloud Feature Mixing Encoder (FM-Encoder) to obtain 3D mixed feature embeddings. Concurrently, corresponding images and textual descriptions are processed through the frozen image and text Feature Mixing Encoders (FM-Encoders) to extract image and text mixed feature embeddings. These extracted 3D, image, and text features are then combined to mixed feature triplets. Employing a contrastive learning objective, the mixed 3D features are aligned with the image and text mixed features. This encourages the point cloud Feature Mixing Encoder (FM-Encoder) to learn a feature space that is consistent with the representations of the frozen encoders from other modalities, enhancing its ability to discriminate between different 3D objects. The Stage 1 corresponding contrastive loss L^{S1} is calculated as:

$$F(x, y) = \log \frac{\exp(x \cdot y/\tau)}{\sum_j \exp(x_j \cdot y_j/\tau)}, \tag{4}$$

$$\begin{aligned} L^{S1} = -\frac{1}{4n} \sum_i (F(m_i^P, m_i^I) + F(m_i^I, m_i^P) + \\ F(m_i^P, m_i^T) + F(m_i^T, m_i^P)), \end{aligned} \tag{5}$$

where n is the number of mixed features in a batch, τ is a learnable temperature, and m_j^P, m_j^I, m_j^T denote normalized projected features of the mixed features of point clouds, images, and text respectfully. Because the image encoder and text encoder are frozen, we extract and cache the features before training for acceleration.

In the second stage, We initialize a new trainable 3D encoder. All Feature Mixing Encoders (FM-Encoders) remain frozen in this stage. Then we introduce a mixed point cloud input to further refine the 3D feature representations. Two input point clouds are selected and processed using farthest point sampling (FPS) and point-level mixing to create a novel mixed point cloud. The mixed point cloud is input to the new trainable 3D encoder to obtain mixed 3D feature embeddings. Simultaneously, the frozen Feature Mixing Encoders (FM-Encoders), are used to extract mixed features from their respective inputs. Using a contrastive learning objective, the 3D features of the mixed point cloud are aligned with the mixed features from the frozen encoders, ensuring that the new 3D encoder learns robust and discriminative mixed feature representations from different modalities. The Stage 2 contrastive loss L^{S2} is calculated as:

$$\begin{aligned} L^{S2} = -\frac{1}{6n} \sum_i (F(m_i^M, m_i^I) + F(m_i^I, m_i^M) + \\ F(m_i^M, m_i^T) + F(m_i^T, m_i^M) + \\ F(m_i^M, m_i^P) + F(m_i^P, m_i^M)), \end{aligned} \tag{6}$$

where m_j^M denotes normalized projected features of the mixed point clouds M.

By leveraging these two stages, the MM-Mixing training pipeline fully exploits the complementary advantages of image and text encoders, integrating multi-modal information to develop a 3D encoder capable of producing highly discriminative features. In the first stage, the point cloud-image-text feature-level mixing ensures the consistency of augmented features across different modalities, facilitating the 3D encoder's cross-modal understanding. The second stage introduces input-level mixing, providing a vast array of complex and realistic samples that enhance the 3D encoder's generalization ability. Under the constraints of contrastive learning, MM-Mixing maintains the consistency between the features of the mixed point clouds and the mixed features of the point clouds.

Experiments

Experimental Setup

Pre-training datasets. In our experimental setup, we utilize datasets following the approach outlined by the state-of-the-art OpenShape (Liu et al. 2024). Our model is pre-trained using triplets generated from four key datasets: ShapeNetCore (Chang et al. 2015), 3D-FUTURE (Fu et al. 2021), ABO (Collins et al. 2022), and Objaverse (Deitke et al. 2023). Specifically, the "ShapeNet" training set is composed entirely of triplets from the ShapeNetCore dataset, which includes 52,470 3D shapes along with their associated images and text descriptions. The comprehensive "Ensembled" dataset includes a total of 875,665 triplets, encompassing

Pre-training Dataset	3D Backbone	Pre-training Method	ModelNet40			ScanObjectNN			Objaverse		
			Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5
Projected images	-	PointCLIP (Zhang et al. 2022b)	19.3	28.6	34.8	10.5	20.8	30.6	1.9	4.1	5.8
		PointCLIP v2 (Zhu et al. 2023)	63.6	77.9	85.0	42.2	63.3	74.5	4.7	9.5	12.9
ShapeNet	Transformer	ReCon (Qi et al. 2023)	61.2	73.9	78.1	42.3	62.5	75.6	1.1	2.7	3.7
		CG3D (Hegde, Valanarasu, and Patel 2023)	48.7	60.7	66.5	42.5	57.3	60.8	5.0	9.5	11.6
		CLIP2Point (Huang et al. 2023)	49.5	71.3	81.2	25.5	44.6	59.4	2.7	5.8	7.9
	SparseConv	OpenShape (Liu et al. 2024)	72.9	87.2	89.5	52.7	72.7	83.6	11.6	21.8	27.1
		MM-Mixing (Ours)	75.2	88.9	91.9	60.7	79.0	87.3	13.0	23.4	28.6
		↑ *Improve*	+2.3	+1.7	+2.4	+8.0	+6.3	+3.7	+1.4	+1.6	+1.5
	Point-BERT	ULIP (Xue et al. 2023a)	60.4	79.0	84.4	51.5	71.1	80.2	6.2	13.6	17.9
		OpenShape (Liu et al. 2024)	70.3	86.9	91.3	51.3	69.4	78.4	10.8	20.2	25.0
		MM-Mixing (Ours)	74.1	88.8	91.6	**61.9**	**83.0**	**91.8**	13.0	22.9	27.9
		↑ *Improve*	+3.8	+1.9	+0.3	+10.6	+13.6	+13.4	+2.2	+2.7	+2.9
Ensembled	SparseConv	OpenShape (Liu et al. 2024)	83.4	95.6	97.8	56.7	78.9	88.6	43.4	64.8	72.4
		MM-Mixing (Ours)	**86.7**	**97.7**	**98.7**	58.4	79.5	89.4	46.2	68.2	75.8
		↑ *Improve*	+3.3	+2.1	+0.9	+1.7	+0.6	+0.8	+2.8	+3.4	+3.4
	Point-BERT	ULIP (Xue et al. 2023a)	75.1	88.1	93.2	51.6	72.5	82.3	26.8	44.8	52.6
		OpenShape (Liu et al. 2024)	84.4	96.5	98.0	52.2	79.7	88.7	46.8	69.1	77.0
		MM-Mixing (Ours)	86.0	96.6	98.4	54.3	79.9	89.1	**51.4**	**73.1**	**80.1**
		↑ *Improve*	+1.6	+0.1	+0.4	+2.1	+0.2	+0.4	+4.6	+4.0	+3.1

Table 1: **Zero-shot 3D classification on ModelNet40, ScanObjectNN and Objaverse-LVIS.** We report the top-1, top-3 and top-5 classification accuracy (%) for different 3D backbones pre-trained on ShapeNet and Ensembled.

Pre-training Dataset	Pre-training Method	M-40	ScanObjectNN		
			OBJ-BG	OBJ-ONLY	PB-T50-RS
ShapeNet	ULIP	**90.6**	75.4	75.4	64.8
	OpenShape	88.5	77.8	78.5	64.1
	MM-Mixing	**90.6**	**83.3**	**85.0**	**73.2**
	↑ *Improve*	+2.1	+5.5	+6.5	+9.1
Ensembled	OpenShape	91.3	85.9	85.4	78.0
	MM-Mixing	**91.7**	**86.9**	**86.2**	**79.3**
	↑ *Improve*	+0.4	+1.0	+0.8	+1.3

Table 2: **Linear probing 3D classification results.** We report the classification accuracy (%) of Point-BERT on ModelNet40 and three splits of ScanObjectNN.

data from all four datasets, thereby providing a rich source of varied 3D shapes and their corresponding images and texts.

Evaluation datasets. For the evaluation of our model, we use a set of datasets that ensures a thorough assessment across different types of 3D data. The Objaverse-LVIS dataset (Deitke et al. 2023), which is part of our evaluation, contains an extensive variety of categories with 46,832 high-quality shapes distributed across 1,156 LVIS (Gupta, Dollar, and Girshick 2019) categories, offering a diverse and challenging environment for testing. Additionally, we include ModelNet40 (Wu et al. 2015) in our evaluation process, a well-known synthetic indoor 3D dataset consisting of 40 categories with a test split of 2,468 shapes. The ScanObjectNN (Uy et al. 2019) dataset, which includes scanned objects from 15 common categories, provides multiple variants such as OBJ-BG, OBJ-ONLY, and PB-T50-RS, each presenting unique challenges (Qi et al. 2023; Wu et al. 2022). Our experiments are conducted across several distinct tasks: zero-shot 3D classification, linear probing 3D classification, and cross-modal 3D shape retrieval to highlight the capabilities and versatility of our model. Further details regarding the implementation specifics for pre-training and evaluation are provided in the Appendix.

Zero-shot 3D Classification

Zero-shot classification refers to the process where a pre-trained model is directly employed to classify a target dataset without any supervision or prior knowledge from that specific dataset. This task presents a considerable challenge for the model, requiring it to exhibit robust knowledge generalization, deep understanding of 3D shapes, and efficient cross-modal alignment. We conduct extensive experiments to validate the effectiveness and robustness of our proposed MM-Mixing on three benchmark datasets: ModelNet40, ScanObjectNN, and Objaverse.

As shown in Table 1, MM-Mixing consistently outperforms state-of-the-art methods under the same configurations (e.g., pre-trained datasets, training epochs, 3D backbones) and enhances the performance of various 3D models across all datasets. For instance, when pre-trained on ShapeNet, MM-Mixing boosts the accuracy of Point-BERT from 51.3% to 61.9% on the real-world dataset ScanObjectNN, even surpassing the 52.2% achieved by OpenShape pre-training on the Ensembled dataset. It indicates that MM-Mixing makes full use of limited multi-modal data to improve the model's understanding of 3D shapes and shows strong performance in handling complex noise interference.

Moreover, on the challenging long-tail dataset, Objaverse, Point-BERT pre-trained with MM-Mixing achieves the accuracy of 51.4%, outperforming OpenShape's 46.8%. Another 3D backbone, SparseConv, also showed a 2.8% improvement in accuracy with our pre-training method. It indicates that existing 3D encoders can be easily incorporated into MM-Mixing framework, leading to a significant en-

Mixing level	ModelNet40 Top1	ModelNet40 Top5	ScanObjectNN Top1	ScanObjectNN Top5	Objaverse Top1	Objaverse Top5
Baseline	72.9	89.5	52.7	83.6	11.6	27.1
FM	74.1	90.1	56.4	84.7	12.2	27.3
IM	73.8	90.4	58.9	85.2	12.4	27.5
FM+IM	**75.2**	**91.9**	**60.7**	**87.3**	**13.0**	**28.6**

Table 3: **Ablation studies on Mixing level in alignment.** "FM" represents feature-level mixing. "IM" represents input-level mixing.

Stage	ModelNet40 Top1	ModelNet40 Top5	ScanObjectNN Top1	ScanObjectNN Top5	Objaverse Top1	Objaverse Top5
One stage	73.6	90.2	59.5	85.8	12.3	27.7
Two stages	**75.2**	**91.9**	**60.7**	**87.3**	**13.0**	**28.6**

Table 4: **Ablation studies on Alignment stage.** "One stage" represents all learnable networks are trained simultaneously.

hancement in 3D shape understanding.

When the pre-training data is expanded from ShapeNet to a larger Ensembled dataset, the performance gains from MM-Mixing are slightly diminished. However, it still provides consistent accuracy gains to the models, underscoring the effectiveness of MM-Mixing on large-scale datasets.

Linear Probing 3D Classification

To better adapt the model to the specific classification of downstream tasks, we train a dataset-dependent learnable linear layer to process the 3D features generated by the pre-trained model. Since only the linear layer is activated in this process, the training is lightweight.

The linear probing results are illustrated in Table 2. When pre-trained on ShapeNet, MM-Mixing achieves 90.6% accuracy on ModelNet40, outperforming OpenShape by 2.1%. On ScanObjectNN, MM-Mixing shows significant improvements, surpassing OpenShape (Liu et al. 2024) by 5.5%, 6.5% and 9.1% on OBJ-BG, OBJ-ONLY, and PB-T50-RS, respectively. When using the Ensembled dataset for pre-training, MM-Mixing maintains its lead with 91.7% accuracy on ModelNet40 and consistent superiority on ScanObjectNN three subsets, with accuracies of 86.9%, 86.2%, and 79.3% respectively. These findings emphasize that MM-Mixing has learned robust and discriminative 3D feature representations during pre-training, which can be efficiently applied to downstream specific classification tasks through a simple linear layer.

Ablation Study

We systematically study the impact of different components in MM-Mixing on the model's performance, including the mixing level, alignment stage, modality loss function, and training costs analysis. All results are the classification accuracy (%) of SparseConv pre-trained on ShapeNet.

Mixing levels in alignment. We investigate the impact of different mixing levels, including Feature-level Mixing (FM), Input-level Mixing (IM), and their combination (FM+IM). Compared to the baseline without mixing, all three strategies consistently improve the performance across all datasets. In Table 3, Feature-level Mixing (FM) and Input-level Mixing (IM) individually contribute to the performance gains, and their combination (FM+IM) further improves the results. It confirms that the two mixing levels complement each other: Feature-level Mixing (FM) ensures cross-modal consistency in the feature latent space, while Input-level Mixing (IM) refines the realistic point cloud representation with challenging samples. Together, they enhance the model's ability of 3D understanding.

$\mathcal{L}_\mathcal{T}$	$\mathcal{L}_\mathcal{I}$	$\mathcal{L}_\mathcal{P}$	ModelNet40 Top1	ModelNet40 Top5	ScanObjectNN Top1	ScanObjectNN Top5	Objaverse Top1	Objaverse Top5
✓			72.6	89.2	58.9	85.4	11.4	24.7
✓	✓		73.8	90.8	**60.7**	85.4	12.5	27.9
✓		✓	73.9	89.7	60.4	85.6	11.7	25.7
✓	✓	✓	**75.2**	**91.9**	**60.7**	**87.3**	**13.0**	**28.6**

Table 5: **Ablation studies on Modality loss function.** $\mathcal{L}_\mathcal{T}$ represents the text loss. $\mathcal{L}_\mathcal{I}$ represents the image loss. $\mathcal{L}_\mathcal{P}$ represents the point cloud loss.

Alignment stages. As shown in Table 4, we evaluate the effectiveness of our two-stage alignment design. Compared to one-stage alignment, the two-stage alignment method can better utilize diverse mixed samples to enhance cross-modal consistency.

Modality loss functions. Our ablation studies on different modality loss functions are shown in Table 5. The text loss $\mathcal{L}_\mathcal{T}$ provides a strong foundation for learning 3D representations with semantic information, while the image loss $\mathcal{L}_\mathcal{I}$ and point cloud loss $\mathcal{L}_\mathcal{P}$ offer complementary visual and shape information, enhancing the model's performance. The combination of all three modality loss functions consistently achieves the best results across all datasets, demonstrating the effectiveness of our framework.

Training costs analysis. Notably, the epochs of one-stage methods are the same as the two-stage training epochs of MM-Mixing for a fair comparison. Both 3D encoders are trained independently for the duration of one stage without shared weights. The experimental results demonstrate that the performance gains of MM-Mixing primarily stem from our mixing-based alignment framework, and the two-stage training framework further enhances the effectiveness of dual-level mixing. Moreover, for previous methods like OpenShape, adding additional training costs (e.g. training time and training parameters) does not significantly improve the performance of the 3D backbone (See Appendix for more details).

Qualitative Analysis

Hard sample recognition. In real-world scenarios, numerous objects exhibit similar morphological or visual characteristics despite belonging to distinct categories. We designate these challenging instances as "hard samples." There are some such category pairs in ModelNet40, such as: "vase & cup", "table & desk", "TV stand & dresser", and "plant & flower pot". As illustrated in Figure 3, MM-Mixing demon-

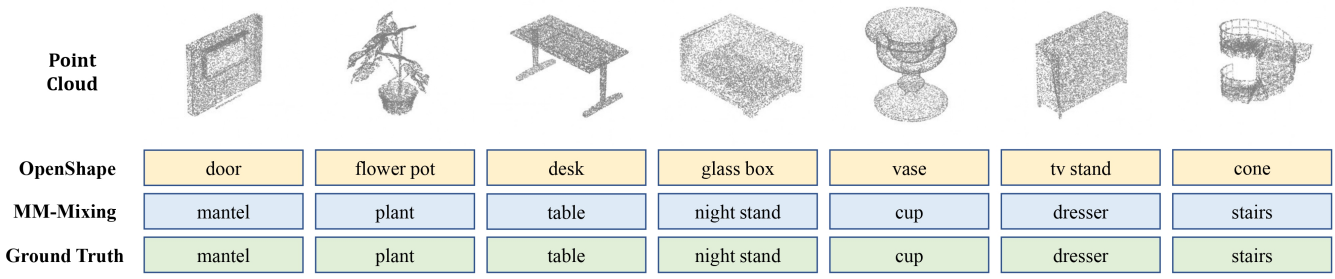

Figure 3: **Hard sample recognition on ModelNet40.** Compared to OpenShape, MM-Mixing enables the model to better capture typical features across different categories and the ability to distinguish hard samples.

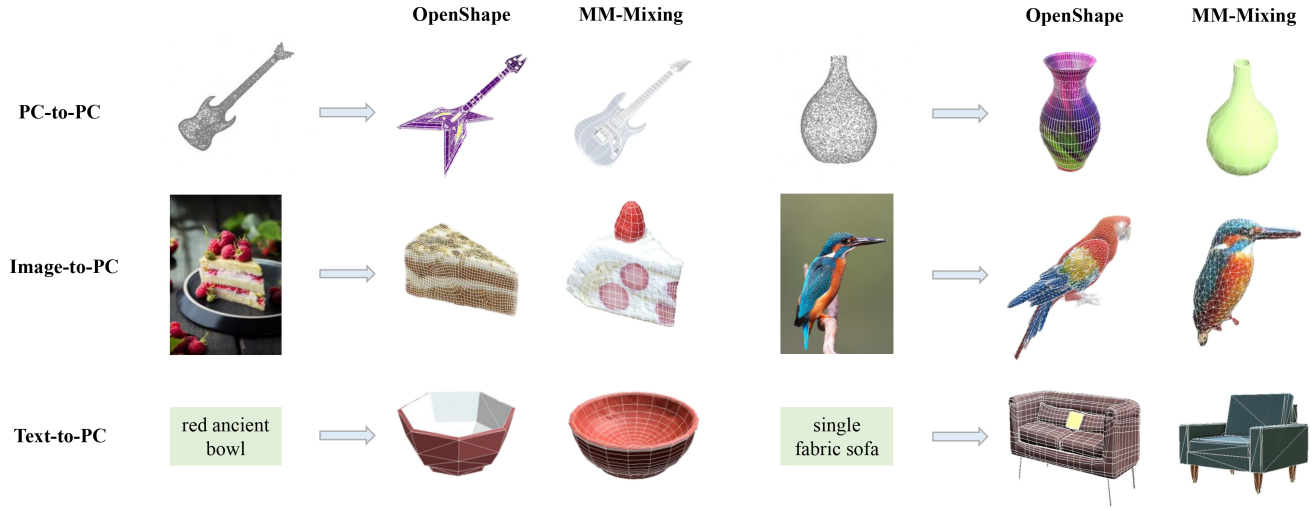

Figure 4: **Cross-modal 3D shape retrieval on Objaverse.** Compared to OpenShape, MM-Mixing enhances the model's understanding of point cloud shapes, image colors, and textual descriptions, effectively improving cross-modal 3D shape retrieval capabilities. PC represents Point Cloud.

strates the capability to capture subtle differences between objects that may appear similar but have different categories. Additionally, it can leverage detailed features to prevent misidentifying a table as a desk. It can be confirmed that MM-Mixing enhances model performance in 3D object recognition.

Cross-modal 3D shape retrieval. The visualization in Figure 4 illustrates the superior performance of our method, MM-Mixing, compared to OpenShape in various cross-modal retrieval tasks. For PC-to-PC retrieval, MM-Mixing demonstrates a finer capture of shape details, as seen with the more accurate symmetrical guitar shape. In Image-to-PC retrieval, our method excels in preserving color details, which can retrieve more rational and approximate point clouds, such as the cake example. Additionally, in text-to-PC retrieval, MM-Mixing shows enhanced compatibility with complex textual descriptions, accurately reflecting shape, color, and material details, as evidenced by the "single fabric sofa" example. These results highlight MM-Mixing's effectiveness in improving shape fidelity, color accuracy, and textual comprehension in cross-modal retrieval.

Conclusion

In this paper, we propose **MM-Mixing**, a multimodal mixing alignment approach that addresses the challenges of multi-modal alignment and enhances model generation for 3D understanding. By integrating the mixing-based method with multimodal data through a two-stage training pipeline, MM-Mixing enhances the performance and generalization capabilities of the models, which ensures a cohesive enhancement of features from different modalities. Extensive experiments demonstrate the effectiveness of MM-Mixing, significantly boosting baseline performance across various settings, including zero-shot 3D classification, linear probing 3D classification, and cross-modal 3D shape retrieval. Moreover, MM-Mixing addresses the previously unexplored issue of multimodal mixing alignment, offering a simple yet effective solution that can be easily integrated into existing frameworks. As 3D vision continues to evolve and find applications in various domains, MM-Mixing represents a significant step forward in meeting the challenges of robust and generalizable models.

Acknowledgments

This work was supported partially by the National Key R&D Program of China (2023YFC3604204), a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project Number: 14201321), the following grants from the Hong Kong Innovation and Technology Fund (Project Number: MHP/086/21 and MHP/092/22), and the Joint Funds of the National Natural Science Foundation of China under Grant U24A20248.

References

Armeni, I.; Sener, O.; Zamir, A. R.; Jiang, H.; Brilakis, I.; Fischer, M.; and Savarese, S. 2016. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1534–1543.

Canfes, Z.; Atasoy, M. F.; Dirik, A.; and Yanardag, P. 2023. Text and image guided 3d avatar generation and manipulation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 4421–4431.

Chang, A. X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*.

Chen, S.; Liu, B.; Feng, C.; Vallespi-Gonzalez, C.; and Wellington, C. 2020a. 3d point cloud processing and learning for autonomous driving: Impacting map creation, localization, and perception. *IEEE Signal Processing Magazine*, 38(1): 68–86.

Chen, Y.; Hu, V. T.; Gavves, E.; Mensink, T.; Mettes, P.; Yang, P.; and Snoek, C. G. 2020b. Pointmixup: Augmentation for point clouds. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, 330–345. Springer.

Choy, C.; Gwak, J.; and Savarese, S. 2019. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3075–3084.

Collins, J.; Goel, S.; Deng, K.; Luthra, A.; Xu, L.; Gundogdu, E.; Zhang, X.; Vicente, T. F. Y.; Dideriksen, T.; Arora, H.; et al. 2022. Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 21126–21136.

Dargan, S.; Bansal, S.; Kumar, M.; Mittal, A.; and Kumar, K. 2023. Augmented reality: A comprehensive review. *Archives of Computational Methods in Engineering*, 30(2): 1057–1080.

Deitke, M.; Schwenk, D.; Salvador, J.; Weihs, L.; Michel, O.; VanderBilt, E.; Schmidt, L.; Ehsani, K.; Kembhavi, A.; and Farhadi, A. 2023. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13142–13153.

Fu, H.; Jia, R.; Gao, L.; Gong, M.; Zhao, B.; Maybank, S.; and Tao, D. 2021. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision*, 129: 3313–3337.

Goyal, A.; Law, H.; Liu, B.; Newell, A.; and Deng, J. 2021. Revisiting point cloud shape classification with a simple and effective baseline. In *International Conference on Machine Learning*, 3809–3820. PMLR.

Gupta, A.; Dollar, P.; and Girshick, R. 2019. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5356–5364.

Hegde, D.; Valanarasu, J. M. J.; and Patel, V. 2023. Clip goes 3d: Leveraging prompt tuning for language grounded 3d recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2028–2038.

Huang, R.; Pan, X.; Zheng, H.; Jiang, H.; Xie, Z.; Wu, C.; Song, S.; and Huang, G. 2024. Joint representation learning for text and 3D point cloud. *Pattern Recognition*, 147: 110086.

Huang, T.; Dong, B.; Yang, Y.; Huang, X.; Lau, R. W.; Ouyang, W.; and Zuo, W. 2023. Clip2point: Transfer clip to point cloud classification with image-depth pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22157–22167.

Kim, S.; Lee, S.; Hwang, D.; Lee, J.; Hwang, S. J.; and Kim, H. J. 2021. Point cloud augmentation with weighted local transformations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 548–557.

Lee, D.; Lee, J.; Lee, J.; Lee, H.; Lee, M.; Woo, S.; and Lee, S. 2021. Regularization strategy for point cloud via rigidly mixed sample. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15900–15909.

Lee, S.; Jeon, M.; Kim, I.; Xiong, Y.; and Kim, H. J. 2022. Sagemix: Saliency-guided mixup for point clouds. *Advances in Neural Information Processing Systems*, 35: 23580–23592.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Li, R.; Li, X.; Heng, P.-A.; and Fu, C.-W. 2020. Pointaugment: an auto-augmentation framework for point cloud classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6378–6387.

Liu, M.; Shi, R.; Kuang, K.; Zhu, Y.; Li, X.; Han, S.; Cai, H.; Porikli, F.; and Su, H. 2024. Openshape: Scaling up 3d shape representation towards open-world understanding. *Advances in Neural Information Processing Systems*, 36.

Liu, Z.; Zhang, Z.; Cao, Y.; Hu, H.; and Tong, X. 2021. Group-free 3d object detection via transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2949–2958.

Ma, X.; Qin, C.; You, H.; Ran, H.; and Fu, Y. 2022. Rethinking network design and local geometry in point cloud: A simple residual MLP framework. *arXiv preprint arXiv:2202.07123*.

Mendoza-Ramírez, C. E.; Tudon-Martinez, J. C.; Félix-Herrán, L. C.; Lozoya-Santos, J. d. J.; and Vargas-Martínez, A. 2023. Augmented reality: survey. *Applied Sciences*, 13(18): 10491.

Pang, Y.; Wang, W.; Tay, F. E.; Liu, W.; Tian, Y.; and Yuan, L. 2022. Masked autoencoders for point cloud self-supervised learning. In *European conference on computer vision*, 604–621. Springer.

Peng, S.; Genova, K.; Jiang, C.; Tagliasacchi, A.; Pollefeys, M.; Funkhouser, T.; et al. 2023. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 815–824.

Qi, Z.; Dong, R.; Fan, G.; Ge, Z.; Zhang, X.; Ma, K.; and Yi, L. 2023. Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. In *International Conference on Machine Learning*, 28223–28243. PMLR.

Qian, G.; Li, Y.; Peng, H.; Mai, J.; Hammoud, H.; Elhoseiny, M.; and Ghanem, B. 2022. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. *Advances in Neural Information Processing Systems*, 35: 23192–23204.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rao, Y.; Liu, B.; Wei, Y.; Lu, J.; Hsieh, C.-J.; and Zhou, J. 2021. Randomrooms: Unsupervised pre-training from synthetic shapes and randomized layouts for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3283–3292.

Riegler, G.; Osman Ulusoy, A.; and Geiger, A. 2017. Octnet: Learning deep 3d representations at high resolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3577–3586.

Tan, D. S.; Robertson, G. G.; and Czerwinski, M. 2001. Exploring 3D navigation: combining speed-coupled flying with orbiting. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 418–425.

Umam, A.; Yang, C.-K.; Chuang, Y.-Y.; Chuang, J.-H.; and Lin, Y.-Y. 2022. Point mixswap: Attentional point cloud mixing via swapping matched structural divisions. In *European Conference on Computer Vision*, 596–611. Springer.

Uy, M. A.; Pham, Q.-H.; Hua, B.-S.; Nguyen, T.; and Yeung, S.-K. 2019. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1588–1597.

Vu, T.; Kim, K.; Luu, T. M.; Nguyen, T.; and Yoo, C. D. 2022. Softgroup for 3d instance segmentation on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2708–2717.

Wang, Y.; Wang, J.; Guo, Z.; Zhang, R.; Zhou, D.; Chen, G.; Liu, A.; and Heng, P.-A. 2024a. Point Cloud Understanding via Attention-Driven Contrastive Learning. *arXiv preprint arXiv:2411.14744*.

Wang, Y.; Wang, J.; Li, J.; Zhao, Z.; Chen, G.; Liu, A.; and Heng, P. A. 2024b. Pointpatchmix: Point cloud mixing with patch scoring. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Wu, X.; Lao, Y.; Jiang, L.; Liu, X.; and Zhao, H. 2022. Point transformer v2: Grouped vector attention and partition-based pooling. *Advances in Neural Information Processing Systems*, 35: 33330–33342.

Wu, Z.; Song, S.; Khosla, A.; Yu, F.; Zhang, L.; Tang, X.; and Xiao, J. 2015. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1912–1920.

Xue, L.; Gao, M.; Xing, C.; Martín-Martín, R.; Wu, J.; Xiong, C.; Xu, R.; Niebles, J. C.; and Savarese, S. 2023a. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1179–1189.

Xue, L.; Yu, N.; Zhang, S.; Li, J.; Martín-Martín, R.; Wu, J.; Xiong, C.; Xu, R.; Niebles, J. C.; and Savarese, S. 2023b. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. *arXiv preprint arXiv:2305.08275*.

Yu, X.; Tang, L.; Rao, Y.; Huang, T.; Zhou, J.; and Lu, J. 2022. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19313–19322.

Zeng, Y.; Jiang, C.; Mao, J.; Han, J.; Ye, C.; Huang, Q.; Yeung, D.-Y.; Yang, Z.; Liang, X.; and Xu, H. 2023. Clip2: Contrastive language-image-point pretraining from real-world point cloud data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15244–15253.

Zhang, J.; Chen, L.; Ouyang, B.; Liu, B.; Zhu, J.; Chen, Y.; Meng, Y.; and Wu, D. 2022a. Pointcutmix: Regularization strategy for point cloud classification. *Neurocomputing*, 505: 58–67.

Zhang, R.; Guo, Z.; Zhang, W.; Li, K.; Miao, X.; Cui, B.; Qiao, Y.; Gao, P.; and Li, H. 2022b. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8552–8562.

Zhang, R.; Wang, L.; Qiao, Y.; Gao, P.; and Li, H. 2023. Learning 3d representations from 2d pre-trained models via image-to-point masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21769–21780.

Zhang, Z.; Cao, S.; and Wang, Y.-X. 2024. TAMM: TriAdapter Multi-Modal Learning for 3D Shape Understanding. *arXiv preprint arXiv:2402.18490*.

Zhu, X.; Zhang, R.; He, B.; Guo, Z.; Liu, J.; Xiao, H.; Fu, C.; Dong, H.; and Gao, P. 2024. No Time to Train: Empowering Non-Parametric Networks for Few-shot 3D Scene Segmentation. *CVPR 2024 Highlight*.

Zhu, X.; Zhang, R.; He, B.; Guo, Z.; Zeng, Z.; Qin, Z.; Zhang, S.; and Gao, P. 2023. Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2639–2650.