Sculpting Holistic 3D Representation in Contrastive Language-Image-3D Pre-training

Yipeng Gao[1] Zeyu Wang[2] Wei-Shi Zheng[1] Cihang Xie[2] Yuyin Zhou[2]
[1]Sun Yat-sen University [2]University of California, Santa Cruz

Abstract

*Contrastive learning has emerged as a promising paradigm for 3D open-world understanding, i.e., aligning point cloud representation to image and text embedding space individually. In this paper, we introduce MixCon3D, a simple yet effective method aiming to sculpt holistic 3D representation in contrastive language-image-3D pre-training. In contrast to point cloud only, we develop the 3D object-level representation from complementary perspectives, e.g., multi-view rendered images with the point cloud. Then, MixCon3D performs language-3D contrastive learning, comprehensively depicting real-world 3D objects and bolstering text alignment. Additionally, we pioneer the first thorough investigation of various training recipes for the 3D contrastive learning paradigm, building a solid baseline with improved performance. Extensive experiments conducted on three representative benchmarks reveal that our method significantly improves over the baseline, surpassing the previous state-of-the-art performance on the challenging 1,156-category Objaverse-LVIS dataset by **5.7%**. The versatility of MixCon3D is showcased in applications such as text-to-3D retrieval and point cloud captioning, further evidencing its efficacy in diverse scenarios. The code is available at https://github.com/UCSC-VLAA/MixCon3D.*

1. Introduction

The ability to perceive and comprehend 3D environments is crucial in applications like augmented and virtual reality, autonomous driving, and embodied AI. Despite significant progress achieved in closed-set 3D recognition [29, 50, 51, 53, 54, 68, 69, 80, 88], there is still a distinct gap between the advanced development of 2D and 3D vision methods. This phenomenon primarily stems from the limited diversity and complexity of existing 3D datasets caused by high data acquisition costs.

Recent research endeavors have turned to well-trained 2D foundation models to unlock the full potential of 3D

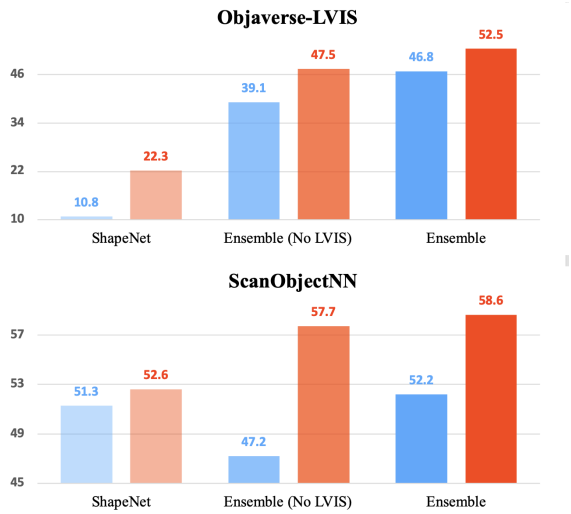

Figure 1. Comparison of zero-shot point cloud recognition between the OpenShape (blue) and our MixCon3D (red) under different pre-training datasets (ShapeNet, Ensemble (No LVIS) and Ensemble). Our model obtains consistent improvements on different training datasets on various downstream benchmarks.

open-world recognition. A line of such works is built upon CLIP [56], a pioneering foundation model known for its extraordinary zero-shot recognition capability [17, 34, 82, 85, 89] by training on web-scale data [60, 61]. The knowledge learned from millions or even billions of image-text pairs proves to be invaluable in assisting the model to learn 3D shapes. In this context, ULIP [77] and CLIP^2 [81] first propose to keep the image and text encoder frozen while training the 3D encoder on the (image, text, point cloud) triplets, which leads to substantially increased zero-shot 3D recognition performance. While existing methods have demonstrated great promise, they predominantly center on a vanilla correspondence between point-text and point-image to form contrastive pairs, typically overlooking the intricate relationships across various modalities and perspectives. For instance, multi-view RGB images and 3D point clouds are known to capture distinct yet complementary attributes of a 3D object [4, 7, 19, 26, 38, 62, 71] — point

clouds emphasize depth and geometry, whereas multi-view images excel at representing semantic information from diverse parts. However, distinct characteristics from each modality of the same 3D object are isolated in the previous contrastive pre-training scheme.

To bridge these gaps, in this paper, we propose to sculpt a comprehensive 3D object-level representation in contrastive language-image-3D pre-training, termed as **MixCon3D**, a simple yet effective method tailored to maximize the efficacy and potential of contrastive learning across images, texts, and 3D objects. Central to our approach is *utilizing the complementary information between multi-view 2D images and 3D point clouds to jointly represent a 3D object and align the 3D object-level representation to the text embedding space*. Specifically, we construct holistic 3D representations by simply concatenating the point cloud features and fused multi-view projected image features before contrastive learning, thus fortifying cross-modal alignment. We also establish an advanced training guideline by carefully examining the training recipe (*e.g.*, batch size, temperature parameters, and learning rate schedules). This not only stabilizes the training process but also drives enhanced performance.

As illustrated in Figure 1, our MixCon3D consistently shows remarkable improvements over multiple popular 3D understanding benchmarks. For example, on the well-established ScanObjectNN dataset, our approach substantially outperforms the prior art by 6.4%, demonstrating the strong generalization ability of MixCon3D. Moreover, on the challenging 1,156-category Objaverse-LVIS dataset with long-tailed distribution, our MixCon3D attains an accuracy of 52.5%, surpassing the competing models by a significant margin of 5.7%. Lastly, by following OpenShape [36] to employ the learned 3D features in the tasks of text to 3D shape retrieval and point cloud caption generation, we showcase our newly learned 3D embedding space is well aligned with CLIP image and text embedding space.

2. Related Works

3D Representation Learning. Point-based methods, a prominent category of 3D representation learning, have garnered much attention for their simplicity, effectiveness, and efficiency. The pioneering work, PointNet [50], models the inherent permutation invariance of points with point-wise feature extraction and max-pooling, enabling direct processing of unstructured point sets. PointNet++ [51] enhances PointNet with a hierarchical network architecture to capture local and global geometric cues effectively. Building upon this foundation, the 3D community has witnessed the emergence of a plethora of point-based methods, with a particular focus on the design of effective local modules [37, 53, 64, 65, 68, 69, 76, 88]. PointNext [54] explores an orthogonal direction, underscoring the pivotal role of training and scaling strategies in effective 3D representation learning.

Another line of work focuses on designing self-supervised learning techniques tailored for point cloud understanding. Early endeavors along this direction centered around the proposition of various low-level pretext tasks, including self-reconstruction [1, 12], distortion reconstruction [42, 59], and normal estimation [57]. Recently, the remarkable success of self-supervised learning in the language and vision domain has prompted researchers in the 3D domain to adopt analogous self-supervised learning paradigms [23, 28]. PointContrast [74], for instance, leverages the concept of contrasting two views of the same point cloud to facilitate high-level scene understanding. PointBERT [80] and PointMAE [46], based on the idea of masked modeling, train an autoencoder to recover the masked portion of data with the unmasked part of the input. Recent works [2, 14, 33, 55, 70, 86] employ 2D models as an auxiliary tool to learn 3D point cloud representations.

Unlike designing 3D backbones or self-supervised learning pretext tasks, this paper focuses on multimodal contrastive learning for 3D open-world understanding.

CLIP for 3D open-world understanding. By training on web-scale image-text pairs, CLIP [56] has revolutionized the area of visual representation learning via language supervision. The extraordinary zero-shot recognition performance of CLIP has found applications in a lot of domains, including zero-shot text-to-3D generation [3, 5, 22, 25, 43, 58, 75], zero-shot 3D segmentation or detection [13, 27, 41, 48, 79, 83], and 3D shape understanding [21, 52, 77, 78, 84, 91]. The early exploration in leveraging CLIP for 3D shape understanding typically involves the projection of the original point cloud into depth maps, followed by the direct application of 2D CLIP on them [84, 91]. However, this approach suffers from information loss during projection while introducing extra latency. Additionally, the domain gap between the synthetically rendered depth maps and natural images could significantly hurt CLIP performance.

More recently, multiple works [30, 36, 77, 78, 87, 90] propose to learn a unified embedding space for image, text, and point cloud, through training a 3D encoder aligned with CLIP image/text encoder. JM3D [67] aligns the point cloud representation to a joint image-text embedding space, which ignores the completeness of 3D object representation. In contrast to general modality fusion methods [35, 49], our work follows this line of work but takes one step ahead to reveal the power of constructing a more holistic 3D object representation. We explore the importance of the complementarity between view difference and shape information for a comprehensive 3D object-level representation.

3. MixCon3D

3.1. Preliminaries

Optimization Objectives of Cross-modal Contrastive Learning. By exploiting a massive amount of image-text pairs crawled from the web, the CLIP model [56] has demonstrated exceptional open-world image understanding capability. Typically, given batched image-text pairs $\{(\boldsymbol{x}_i^I, \boldsymbol{x}_i^T)\}_{i=1}^N$ and the image, text encoders f^I, f^T, the CLIP is trained to bring the representations of paired image and text data $(\boldsymbol{x}_i^I, \boldsymbol{x}_i^T)$ closer by the contrastive loss $\mathcal{L}^{I \leftrightarrow T}$ as follows:

$$l^{I \to T} = \sum_i^N \log \frac{\exp(\boldsymbol{z}_i^I \cdot \boldsymbol{z}_i^T / \tau)}{\sum_j \exp(\boldsymbol{z}_i^I \cdot \boldsymbol{z}_j^T / \tau)} \tag{1}$$

$$l^{T \to I} = \sum_i^N \log \frac{\exp(\boldsymbol{z}_i^T \cdot \boldsymbol{z}_i^I / \tau)}{\sum_j \exp(\boldsymbol{z}_i^T \cdot \boldsymbol{z}_j^I / \tau)} \tag{2}$$

$$\mathcal{L}^{I \leftrightarrow T}(\boldsymbol{x}_i^I, \boldsymbol{x}_i^T) = -\frac{1}{2}(l^{I \to T} + l^{T \to I}) \tag{3}$$

where $\boldsymbol{z}_i^I = g^I \circ f^I(\boldsymbol{x}_i^I)/||g^I \circ f^I(\boldsymbol{x}_i^I)||$ and $\boldsymbol{z}_i^T = g^T \circ f^T(\boldsymbol{x}_i^T)/||g^T \circ f^T(\boldsymbol{x}_i^T)||$ are the l_2 normalized image and text features output by projection heads. g^I and g^T are image and text projection heads and τ is a learnable temperature.

As the scale of 3D datasets is relatively smaller, previous works [36, 77, 78, 81] have resorted to the pre-trained CLIP image and text embedding space for training a vanilla 3D model $g^P \circ f^P$ (including 3D encoder f^P and projection head g^P) with open-world recognition ability. Since CLIP is pre-trained on a much larger data scale and is well aligned, its image model $g^I \circ f^I$ and text model $g^T \circ f^T$ are frozen during training. Specifically, given batched N input image $\boldsymbol{x}_i^I$, text $\boldsymbol{x}_i^T$, and point cloud $\boldsymbol{x}_i^P$ triplets $\{(\boldsymbol{x}_i^I, \boldsymbol{x}_i^T, \boldsymbol{x}_i^P)\}_{i=1}^N$ (hence the name image-text-3D), the 3D model $g^P \circ f^P$ is trained to align the point cloud representation $\boldsymbol{z}_i^P = g^P \circ f^P(\boldsymbol{x}_i^P)/||g^P \circ f^P(\boldsymbol{x}_i^P)||$ to the CLIP embedding space by $\mathcal{L}^{P \leftrightarrow I}$ and $\mathcal{L}^{P \leftrightarrow T}$ (each has the similar formulation of Equation 3). In this case, the optimization objective becomes:

$$\mathcal{L}^{P \leftrightarrow I}(\boldsymbol{x}_i^P, \boldsymbol{x}_i^I) + \mathcal{L}^{P \leftrightarrow T}(\boldsymbol{x}_i^P, \boldsymbol{x}_i^T) \tag{4}$$

Revisiting Training Recipe. It is known to the 3D community that a well-tuned training recipe can lead to a dramatic performance boost [54]. Yet, despite its impressively promising performance, the training recipe of the image-text-3D contrastive learning paradigm is underexplored. Thus, before diving deep into our method, we first revisit the training recipe of ULIP [77] and OpenShape [36], identifying useful changes, as listed in Table 1:

Method	Temperature Parameter	Batchsize	Learning Rate Schedule	Warm Up	EMA
ULIP	Share	64	Cosine Decay	✓	✗
OpenShape	Share	200	Step Decay	✗	✗
Improved Recipe	Separate	∼2k	Cosine Decay	✓	✓

Table 1. The summary and comparisons between the baseline and our improved training recipe.

- **Batchsize.** Contrastive learning benefits significantly from a large batch size [8, 56]. Nevertheless, the state-of-the-art model [77] still adopts a small batchsize of 64. We scale the batchsize and note a medium 2k strikes a good trade-off between different datasets.
- **Learning rate schedule.** Unlike ULIP, OpenShape adopts the step learning rate decay schedule without warmup. We adopt the cosine learning rate schedule as CLIP and find it leads to clear improvement.
- **Exponential moving average.** During training, we observe the model performance steadily increase on the synthetic Objaverse-LVIS dataset, while fluctuating drastically on the real-scanned ScanObjectNN dataset, presumably due to the domain gap. We employ Exponential Moving Average (EMA) [63] to alleviate the fluctuation issue to stabilize training.
- **Separate temparature.** Features from different modalities may have different distributions. Prior works [36, 77] use a shared temperature parameter τ [73] to control the concentration level of multi-modal features. We hereby use separate temperature parameters for each modality.

Together, as shown in Table 3, our enhanced recipe substantially boosts the top-1 accuracy of OpenShape baseline by 3.3%, 3.6%, and 1.9% on Objaverse-LVIS, ScanObjectNN, and ModelNet40, respectively. Next, we introduce our proposed MixCon3D for 3D contrastive learning.

3.2. 3D-Text Alignment

Point cloud and 2D images are known to encode different yet complementary cues: point cloud better captures depth and geometry information, while images focus on catching dense semantic information [4, 7, 38, 71]. Meanwhile, in the 3D world, a single-view image only contains partial information captured from a specific camera pose with an angle. Instead, multi-view, a prominent property in 3D representation, has demonstrated promising effectiveness in 3D understanding tasks [19, 20, 26, 62]. Though previous works [36, 77, 78] render images from multiple viewpoints of the same point cloud when creating the data triplets, they merely sample one image from the rendered multi-view images when extracting the image features, which inherently encode only partial facets of the 3D object. Overall, both point cloud and multi-view images are two different per-

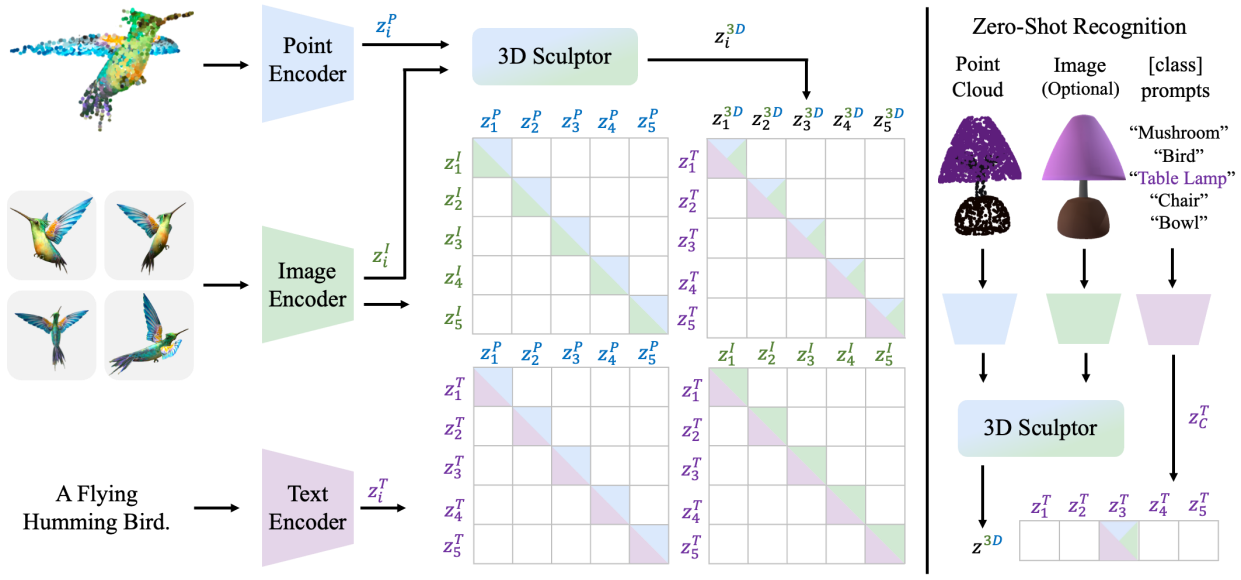

Figure 2. Summary of our MixCon3D framework. MixCon3D first extracts the representation of input triplets (images, text, point cloud) from a pre-trained vision-language model (*e.g.*, CLIP) and a 3D encoder (*e.g.*, Point-BERT) with corresponding projection heads. Then, the image and point cloud features go through a 3D sculptor to obtain the 3D object-level features, serving as complementary representations. The contrastive losses are applied to align features among three modalities (image-text-3D) and 3D representation to text.

spectives from which a 3D object is treated. To this end, we introduce a simple yet effective 3D-text alignment approach for contrastive language-image-3D pre-training, which constructs a new 3D object-level representation by aggregating the respective features extracted from multi-view RGB images and point cloud modalities. On top of the contrast between the conventional tri-modal features with each other, the joint representation will also be aligned with text features via a 3D-text contrastive loss.

Sculpting Holistic 3D Representation. Given batched data triplets $\{\boldsymbol{x}_i = (\boldsymbol{x}_i^I, \boldsymbol{x}_i^T, \boldsymbol{x}_i^P)\}_{i=1}^N$ and image-text-3D models (f^I, f^T, f^P), the corresponding features are denoted as $\mathbb{R}^D$ vectors ($\boldsymbol{z}_i^I$, $\boldsymbol{z}_i^T$, $\boldsymbol{z}_i^P$), respectively. With M multi-view images $\boldsymbol{x}_i^I = \{\boldsymbol{x}_{(i,j)}^I\}_{j=1}^M$, which corresponds to the text description $\boldsymbol{x}_i^T$ and point cloud $\boldsymbol{x}_i^P$, we replace the single-view image feature with the fusion of individual image features $\boldsymbol{z}_{(i,j)}^I$ extracted from images $\boldsymbol{x}_{(i,j)}^I$. We first construct the multi-view image feature $\boldsymbol{z}_{(i,j)}^I$ by fusing the features $\boldsymbol{z}_{(i,j)}^I$ of every single view $\boldsymbol{x}_{(i,j)}^I$:

$$\boldsymbol{z}_i^I = g^{MV}(\{\boldsymbol{z}_{(i,j)}^I\}_{j=1}^M) \tag{5}$$

where g^{MV} is the multi-view fusion function (*e.g.*, view-pooling, maxpooling, or MLP) for comprehensive RGB representation modeling.

To model a holistic 3D object-level representation, we concatenate the image features and point cloud features (*i.e.*, $\text{concat}(\boldsymbol{z}_i^I, \boldsymbol{z}_i^P) \in \mathbb{R}^{2\times D}$), and use a 3D sculptor g^{3D} (a fully connected layer) to project the joint representation as follows:

$$\boldsymbol{z}_i^{3D} = g^{3D}(\text{concat}(\boldsymbol{z}_i^I, \boldsymbol{z}_i^P)) \tag{6}$$

Training Objectives. With a holistic 3D representation $\boldsymbol{z}_i^{3D}$ and text features $\boldsymbol{z}_i^I$, the 3D-text contrastive term is:

$$\mathcal{L}^{3D\leftrightarrow T}(\boldsymbol{x}_i^I, \boldsymbol{x}_i^P, \boldsymbol{x}_i^T) = -\frac{1}{2}(l^{3D\to T} + l^{T\to 3D}) \tag{7}$$

where $l^{3D\to T}$ and $l^{T\to 3D}$ have similar formulation as shown in Equation 1 and Equation 2 with input $\boldsymbol{z}_i^{3D}$ and $\boldsymbol{z}_i^I$. In this case, the overall objective becomes:

$$\mathcal{L}^{\text{All}}(\boldsymbol{x}_i) = \mathcal{L}^{3D\leftrightarrow T} + \mathcal{L}^{P\leftrightarrow I} + \mathcal{L}^{P\leftrightarrow T} + \mathcal{L}^{I\leftrightarrow T} \tag{8}$$

where $\{\boldsymbol{x}_i = (\boldsymbol{x}_i^I, \boldsymbol{x}_i^T, \boldsymbol{x}_i^P)\}_{i=1}^N$ is the input image-text-point cloud data triplet. Note that we keep the conventional point cloud to text loss $\mathcal{L}^{P\leftrightarrow T}$, enabling the model to make predictions solely based on 3D input even when corresponding images are unavailable [66, 72]. Additionally, different from ULIP and OpenShape, we retain the CLIP loss $\mathcal{L}^{I\leftrightarrow T}$ with an additional learnable projection head upon the frozen CLIP encoder.

Zero-Shot Inference. The texts of class labels in the downstream tasks are used to connect to the learned 3D representation from the point cloud encoder, enabling the ability of zero-shot recognition. Specifically, the C class text

Method	Encoder	Training data	Objaverse-LVIS				ScanObjectNN				ModelNet40			
			Top1	Top1-C	Top3	Top5	Top1	Top1-C	Top3	Top5	Top1	Top1-C	Top3	Top5
PointCLIP [84]	-	Depth inference	1.9	-	4.1	5.8	10.5	-	20.8	30.6	19.3	-	28.6	34.8
PointCLIP v2 [91]	-		4.7	-	9.5	12.9	42.2	-	63.3	74.5	63.6	-	77.9	85.0
ReCon [52]	-	ShapeNet	1.1	-	2.7	3.7	42.3	-	62.5	75.6	61.2	-	73.9	78.1
CG3D [21]	-		5.0	-	9.5	11.6	42.5	-	57.3	60.8	48.7	-	60.7	66.5
CLIP2Point [24]	-		2.7	-	5.8	7.9	25.5	-	44.6	59.4	49.5	-	71.3	81.2
ULIP [77]	PointBERT		6.2	-	13.6	17.9	51.5	-	71.1	80.2	60.4	-	79.0	84.4
OpenShape [36]	SparseConv		11.6	-	21.8	27.1	52.7	-	72.7	**83.6**	72.9	-	87.2	93.0
MixCon3D (**Ours**)	SparseConv		**23.5**	**17.5**	**40.2**	**47.1**	**54.4**	**56.1**	**73.9**	83.3	**73.9**	**70.2**	**88.2**	**94.0**
OpenShape [36]	PointBERT		10.8	-	20.2	25.0	51.3	-	69.4	78.4	70.3	-	86.9	91.3
MixCon3D (**Ours**)	PointBERT		22.3	16.2	37.5	44.3	52.6	52.1	69.9	78.7	72.6	68.2	87.1	91.3
ULIP [77]	PointBERT	Ensemble (No LVIS)	21.4	-	38.1	46.0	46.0	-	66.1	76.4	71.4	-	84.4	89.2
OpenShape [36]	SparseConv		37.0	-	58.4	66.9	54.9	-	76.8	87.0	82.6	-	95.0	97.5
MixCon3D (**Ours**)	SparseConv		45.7	33.5	67.0	73.2	56.5	60.5	77.8	87.5	83.3	82.4	95.6	97.6
OpenShape [36]	PointBERT		39.1	-	60.8	68.9	47.2	-	72.4	84.7	85.3	-	96.2	97.4
MixCon3D (**Ours**)	PointBERT		**47.5**	**34.6**	**69.0**	**76.2**	**57.7**	**61.5**	**80.7**	**89.8**	**87.3**	**86.7**	**96.8**	**98.1**
ULIP [77]	PointBERT	Ensemble	26.8	-	44.8	52.6	51.6	-	72.5	82.3	75.1	-	88.1	93.2
OpenShape [36]	SparseConv		43.4	-	64.8	72.4	56.7	-	78.9	88.6	83.4	-	95.6	97.8
MixCon3D (**Ours**)	SparseConv		47.3	35.0	68.7	76.1	57.1	61.2	79.2	88.9	83.9	83.2	95.9	98.0
OpenShape [36]	PointBERT		46.8	34.0	69.1	77.0	52.2	53.2	79.7	88.7	84.4	84.9	96.5	98.0
MixCon3D (**Ours**)	PointBERT		**52.5**	**38.8**	**74.5**	**81.2**	**58.6**	**62.3**	**80.3**	**89.2**	**86.8**	**86.8**	**96.9**	**98.3**

Table 2. Comparison with state-of-the-art methods on three representative zero-shot 3D reognition benchmarks. "Top1-C" means the top-1 class average accuracy. "Encoder" denotes the point cloud encoder used in the framework. "*" denotes the results we reproduce in public OpenShape datasets by corresponding methods.

features $z_C^T \in \mathbb{R}^{C\times D}$ are obtained by inputting the class label to the text encoder with prompt engineering. Then, for single and mixture modality inference, given the trained 3D sculptor g^{3D} and extracted image $\boldsymbol{z}_i^I$, point cloud $\boldsymbol{z}_i^P$ features, the logits y_i^{3D}, y_i^P, y_i^I between the 3D object and texts are calculated in different ways as follows:

$$\begin{aligned} y_i^{3D} &= g^{3D}(\text{concat}(\boldsymbol{z}_i^I, \boldsymbol{z}_i^P)) \cdot z_C^T \,, \\ y_i^P &= \boldsymbol{z}_i^P \cdot z_C^T \,, y_i^I = \boldsymbol{z}_i^I \cdot z_C^T \end{aligned} \tag{9}$$

Note that our MixCon3D also flexibly supports single-modality zero-shot inference, *i.e.*, utilizing y_i^P for point cloud-to-text [36, 77, 78]) or y_i^I for image-to-text [84, 91].

4. Experiments

We first introduce our experimental setup in Section 4.1. Then, we compare previous state-of-the-art methods in Section 4.2. We also conduct a series of analyses on the key components (Section 4.3), including the improved training strategies, contrastive loss, multi-view, and effect of inference ways. Additionally, our MixCon3D can benefit cross-modal applications such as text to 3D object retrieval and point cloud captioning (Section 4.4).

4.1. Experimental Setup

Pre-training datasets. Following OpenShape [36], the full pre-training dataset (denoted as "Ensemble") contains four pieces: ShapeNet [6], 3D-FUTURE [15], ABO [10]) and Objaverse [11]. The point cloud is obtained by sampling 10,000 points from the mesh surface and the color is interpolated based on the mesh textures. The images are rendered from 12 preset camera poses that cover the whole object uniformly. Then, the paired texts are generated by BLIP [31, 32] and Azure cognition services with GPT4 [45] to filter out noisy text. In addition, we verify the 3D open-world understanding ability of our method trained by the ShapeNet dataset only and the ensembled dataset except for the LVIS [18] categories (denoted as "Ensemble (No LVIS)"), which have fewer categories in training data and constitute a more challenging scenario.

Down-stream datasets. Three datasets are used for evaluating zero-shot point cloud recognition:

- ModelNet40 [72] is a synthetic dataset comprising 3D CAD models, including 9,843 training samples and 2,468 testing samples, distributed across 40 categories.
- ScanObjectNN [66] is a dataset composed of 3D objects acquired through real-world scanning techniques,

Improvements	Objaverse-LVIS				ScanObjectNN				ModelNet40			
	Top1	Top1-C	Top3	Top5	Top1	Top1-C	Top3	Top5	Top1	Top1-C	Top3	Top5
Baseline	46.5	34.0	69.0	76.8	52.0	53.2	77.5	87.5	84.2	84.9	95.9	97.4
+ Separate Temperature	46.8	34.4	69.2	77.1	52.8	54.0	77.6	87.4	84.4	84.6	96.1	97.4
+ Large Batchsize	48.0	35.3	70.1	77.4	53.5	55.5	78.0	87.7	84.8	85.3	96.4	97.7
+ LR Schedule	48.5	36.0	70.6	77.7	54.1	56.3	78.2	87.9	85.0	85.0	96.4	97.9
+ EMA	49.8	36.9	71.7	78.7	55.6	58.9	79.3	88.6	86.1	86.2	96.8	98.3

Table 3. Ablation studies for sequentially applying the improved training strategies for constructing a strong baseline on downstream zero-shot tasks. The baseline denotes only using vanilla $\mathcal{L}^{P\leftrightarrow I}$ and $\mathcal{L}^{P\leftrightarrow T}$ with Point-BERT and OpenShape training recipe.

$\mathcal{L}^{I\leftrightarrow T}$	$\mathcal{L}^{3D\leftrightarrow T}$	Multi-View	Objaverse-LVIS				ScanObjectNN				ModelNet40			
			Top1	Top1-C	Top3	Top5	Top1	Top1-C	Top3	Top5	Top1	Top1-C	Top3	Top5
✗	✗	✗	**49.8**	**36.9**	**71.7**	**78.7**	**55.6**	58.9	**79.3**	**88.6**	**86.1**	**86.2**	**96.8**	**98.3**
✓	✗	✗	48.7	36.2	70.4	77.7	55.4	**59.7**	75.8	85.6	84.7	84.8	96.6	97.9
✓	✗	✓	49.6	36.7	71.0	78.4	55.0	59.3	75.6	85.2	84.7	84.7	96.5	97.8
✗	✓	✗	51.0	37.8	73.2	79.5	57.9	61.4	79.8	**89.3**	86.5	86.4	96.6	98.0
✗	✓	✓	51.5	38.2	73.8	80.5	58.2	61.5	79.8	89.2	86.6	86.5	96.6	98.0
✓	✓	✗	51.6	38.2	73.7	80.6	58.1	61.9	**80.3**	89.2	86.6	86.6	96.4	98.1
✓	✓	✓	**52.5**	**38.8**	**74.5**	**81.2**	**58.6**	**62.3**	**80.3**	89.2	**86.8**	**86.8**	**96.9**	**98.3**

Table 4. The ablation studies of different optimization objectives in the proposed MixCon3D.

encompassing a total of 2,902 objects that are systematically categorized into 15 distinct categories. We follow [36, 77, 78] and use the variants provided by [80] in our experiments.

- Objaverse-LVIS, an annotated subset of the Objaverse [11], incorporates a corpus of 46,832 shapes originating from 1,156 categories in LVIS dataset [18].

Implementation details. We implement our approach in PyTorch [47] and train the models on a server with 8 NVIDIA A5000 GPUs with a batch size of 2048. We train the model for 200 epochs with the AdamW [40] optimizer, a warmup epoch of 10, and a cosine learning rate decay schedule [39]. The base learning rate is set to 1e-3, based on the linear learning rate scaling rule [16]: $lr = base\_lr \times$ batchsize / 256. The EMA factor is set to 0.9995. Following Liu et al. [36], OpenCLIP ViT-bigG-14 [8] is adopted as the pretrained CLIP model.

4.2. Main Results

In Table 2, we compare the performance of our MixCon3D with state-of-the-arts across two representative encoders, SparseConv [9] and PointBERT [80]; three different training set, "ShapeNet", "Ensemble (No LVIS)", and "Ensemble"; and three popular 3D recognition benchmarks, Objaverse-LVIS, ScanObjectNN, and ModelNet40.

We observe that our MixCon3D consistently exhibits superior performance on different scales of the dataset (From "ShapeNet" to "Ensemble" and types of 3D encoders (SparseConv and PointBERT)). Specifically, on the challenging long-tailed benchmark Objaverse-LVIS, MixCon3D greatly improves the zero-shot Top1 accuracy from 46.8% of OpenShape to **52.5%** with PointBERT encoder and "Ensemble" training data. In addition, when tested on the ScanObjectNN dataset that comprises scanned points of real objects and thus a bigger domain gap [66], our MixCon3D also achieves a significant performance boost of **6.4%** (58.6% *vs.*52.2%). These results altogether validate the effectiveness of our proposed MixCon3D, demonstrating a more powerful open-world 3D understanding ability.

4.3. Ablation Studies

Improved training recipe. We show the effect of improved training strategies in Table 3. The separate temperatures obtain a notable performance improvement (+ 0.8% Top1 on ScanObjectNN), indicating the necessity of using separate dynamic scales of logits. A larger batchsize benefits image-text-3D contrastive pre-training, significantly increasing 1.2%/0.7% Top1 accuracy on Objaverse-LVIS and ScanObjectNN. We observe a similar effect of the cosine learning rate schedule with warmup, achieving 48.5% Top1 and 36.0% Top1-C on Objaverse-LVIS without any additional training cost. Lastly, the exponential moving average update brings consistent improvement, especially on the ScanObjectNN (+ 1.5%) with a larger domain gap.

MixCon3D component. In Table 4, we analyze the effect of each critical component in MixCon3D. Interestingly, we find that the image-text alignment alone even leads to

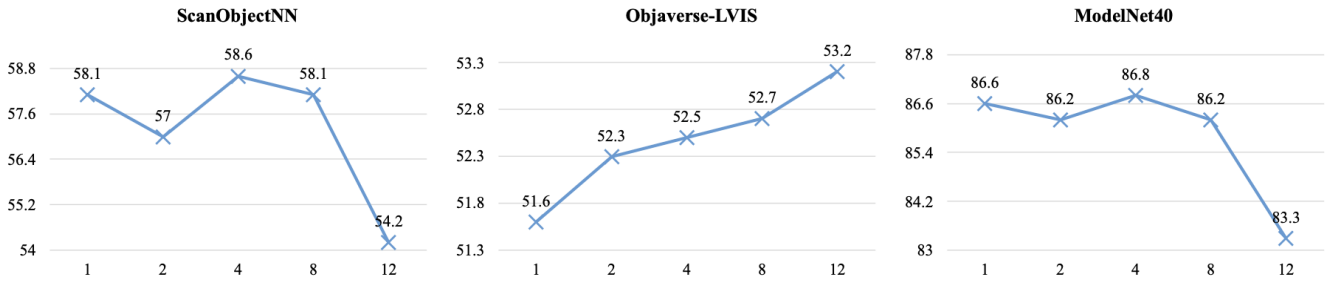

Figure 3. Analysis of the number in the multi-view mechanism. We report the Top 1 Accuracy results from **1** to **12** views.

Function g^{MV}	O-LVIS		S-Object	
	Top1	Top1-C	Top1	Top1-C
-	51.6	38.2	58.1	61.9
View-pooling	52.5	38.8	**58.6**	**62.3**
View-pooling + FC	**52.7**	**39.1**	52.4	54.1
Max pooling	52.1	38.4	56.7	60.0
Max pooling + FC	51.6	38.0	55.8	58.7

Table 5. Ablation studies of the design of fusion function g^{MV}. We report results on Objaverse-LVIS (O-LVIS) and ScanObjectNN (S-Object).

Point Cloud	Image	Multi-View	Objaverse-LVIS			
			Top1	Top1-C	Top3	Top5
✓	✗	-	50.4	37.4	72.2	79.1
✗	✓	✗	44.5	34.5	64.2	70.6
✗	✓	✓	51.9	38.5	73.1	79.4
✓	✓	✗	51.6	37.6	73.4	80.1
✓	✓	✓	**52.5**	**38.8**	**74.5**	**81.2**

Table 6. Analysis of different types of 3D object-level representaton. We report results on Objaverse-LVIS.

worse performance compared to the baseline (decreasing from 49.8% to 48.7% on Objaverse-LVIS and 86.1% to 84.7% on ModelNet40), which potentially hurts the alignment effectiveness on $\mathcal{L}^{P\leftrightarrow I}$ and $\mathcal{L}^{P\leftrightarrow T}$. By contrast, our proposed image-3D to text joint alignment loss $\mathcal{L}^{3D\leftrightarrow T}$ itself brings a considerable performance boost of at least 1.8% on all three datasets, and combining $\mathcal{L}^{I\leftrightarrow T}$ leads to further improvement. This clearly shows the paramount importance of aggregating complimentary useful cues in contrastive learning with image, point cloud, and text. Moreover, we adopt multi-view images to construct a more comprehensive representation of the 3D object on image modality and result in a further improvement of 0.9% and 0.5% Top1 on Objaverse-LVIS and ScanObjectNN, suggesting the importance of considering the holism of 3D objects on cross-modal alignment.

Multi-view component. We next analyze the effect of fusion function g^{MV} (Table 5) and the number of views (Figure 3) used during the pre-training. For g^{MV}, compared with the max poling operation, we observe that simply adopting view-pooling achieves promising improvement (52.5% v.s. 52.1% Top1 on Objaverse-LVIS). Adding an additional fully connected layer (FC) after the pooling operation may boost the performance on in-distribution Objaverse-LVIS (+ 0.2% Top1) while severely lowering the generalization ability on ScanObjectNN (- 6.2% Top1). Since the image modality is only accessible when testing on Objaverse-LVIS, increasing the number of views during training obtains a consistent improvement (from 51.6% to 53.2% Top1 and 38.2% to 39.5% Top1-C) but may slightly hurt the performance on ScanObjectNN (decreasing 0.5% Top1 when increasing the number of views from 4 to 8). To keep a trade-off between datasets, we choose the view-pooling as g^{MV} and view amount $M = 4$ by default.

Multi-modal Inference. The introduction of joint alignment and multi-view images leads to a lot of inference options. For instance, whether we should use point cloud input alone, or combine point cloud with image input. Also, it is necessary to decide whether to apply single-view or multi-view images for complete coverage. We ablate a series of inference ways that aggregate different representations and show the results in Table 6. Even with multi-view images, simply using the point cloud (50.4% Top1) or image modality (51.6% Top1) obtains sub-optimal solutions (compared to 52.5% Top1 that uses modality fusion) since both only cover partial information of a 3D instance. As can be seen, the way of point cloud and image representation fusion, plus multi-view image feature extraction (achieving 52.5% Top1 and 38.8% Top1-C), surpasses all other options by a clear margin, underpinning the significance of knowledge aggregation from different representations.

4.4. Cross-modal Applications

To test how well the point cloud representation of our Mix-Con3D is aligned with CLIP pre-trained representations,

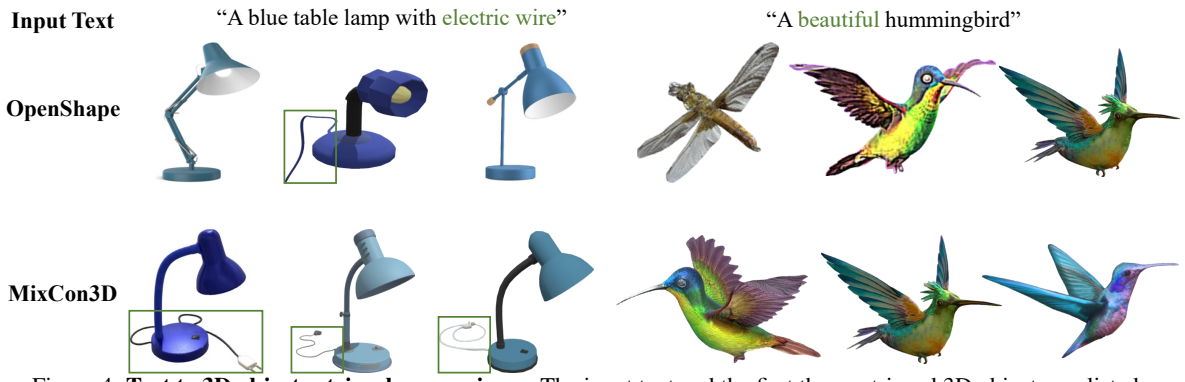

Figure 4. **Text to 3D object retrieval comparisons.** The input text and the first three retrieved 3D objects are listed.

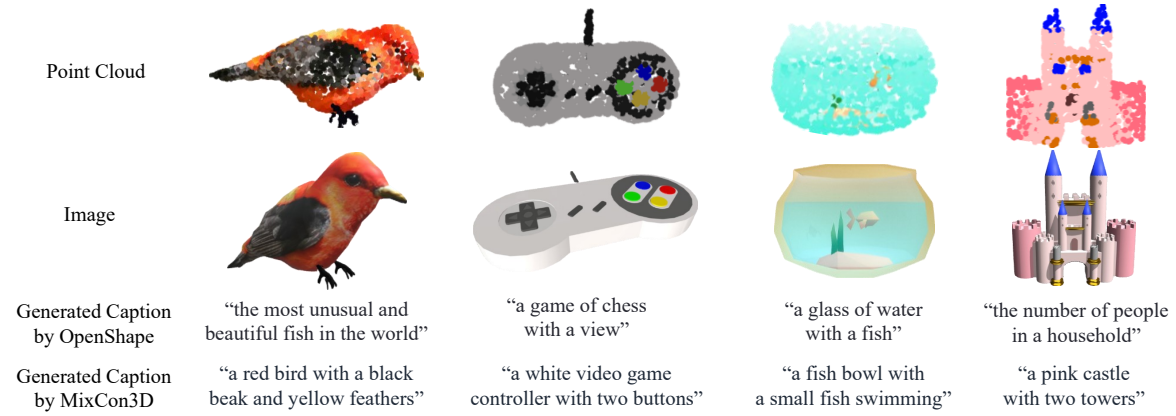

Figure 5. **Point cloud captioning comparisons.** In each row, we list the input point cloud, corresponding images, and generated captions.

we conduct qualitative studies on the following cross-modal tasks, following the practice in Liu et al. [36].

Text to 3D object retrieval. We use cosine similarity between text embeddings of a specific input and 3D shape embeddings from the ensembled dataset as the ranking metric. We compare the retrieval result of our MixCon3D with that of OpenShape. As shown in Figure 4, our MixCon3D can capture more comprehensive feature representation, *e.g.*, allowing for more accurate indexing such as the hummingbird and fine-grained retrieval in situations where the "lamp" is required to have an "electric wire".

Point cloud captioning. We feed the 3D shape embeddings of our MixCon3D into an off-the-shelf image captioning model ClipCap [44] and compare the results with that of OpenShape. As can be observed in Figure 5, our MixCon3D facilitates off-shelf models to generate more accurate and comprehensive captions, indicating that our method can better map the point cloud feature to the pre-aligned image-text feature space.

5. Conclusion

In this paper, we present MixCon3D, a simple yet effective image-text-3D contrastive learning approach, which synergizes multi-modal joint alignment and multi-view representations for better open-world 3D understanding capability. Specifically, we propose constructing a simple yet effective 3D-text alignment training scheme and capitalizing on the features accumulated from multi-view images for a holistic 3D object-level representation. In addition, we provide the first detailed training guideline in the field of contrastive language-image-3D pre-training. Together with the improved training pipeline, MixCon3D achieves superior performance on a wide range of 3D recognition benchmarks and facilitates downstream cross-modal applications such as text-to-3D object retrieval and point cloud captioning. We hope our work could encourage more research endeavors to build the next-generation open-world 3D model.

Acknowledgement

This work is partially supported by TPU Research Cloud (TRC) program and Google Cloud Research Credits program.

References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *ICML*. PMLR, 2018. 2

[2] Mohamed Afham, Isuru Dissanayake, Dinithi Dissanayake, Amaya Dharmasiri, Kanchana Thilakarathna, and Ranga Rodrigo. Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In *CVPR*, 2022. 2

[3] Shivangi Aneja, Justus Thies, Angela Dai, and Matthias Nießner. Clipface: Text-guided editing of textured 3d morphable models. In *SIGGRAPH*, 2023. 2

[4] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *CVPR*, 2022. 1, 3

[5] Zehranaz Canfes, M Furkan Atasoy, Alara Dirik, and Pinar Yanardag. Text and image guided 3d avatar generation and manipulation. In *WACV*, 2023. 2

[6] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 5

[7] Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinhong Jiang, and Feng Zhao. BEVDistill: Cross-modal BEV distillation for multi-view 3d object detection. In *ICLR*, 2023. 1, 3

[8] 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. 3, 6

[9] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*, 2019. 6

[10] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *CVPR*, 2022. 5

[11] 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. 5, 6

[12] Haowen Deng, Tolga Birdal, and Slobodan Ilic. Ppf-foldnet: Unsupervised learning of rotation invariant 3d local descriptors. In *ECCV*, 2018. 2

[13] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Pla: Language-driven open-vocabulary 3d scene understanding. In *CVPR*, 2023. 2

[14] Runpei Dong, Zekun Qi, Linfeng Zhang, Junbo Zhang, Jianjian Sun, Zheng Ge, Li Yi, and Kaisheng Ma. Autoencoders as cross-modal teachers: Can pretrained 2d image transformers help 3d representation learning? *arXiv preprint arXiv:2212.08320*, 2022. 2

[15] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *IJCV*, 2021. 5

[16] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 6

[17] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022. 1

[18] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 5, 6

[19] Abdullah Hamdi, Silvio Giancola, and Bernard Ghanem. Mvtn: Multi-view transformation network for 3d shape recognition. In *ICCV*, 2021. 1, 3

[20] Abdullah Hamdi, Silvio Giancola, and Bernard Ghanem. Voint cloud: Multi-view point cloud representation for 3d understanding. In *ICLR*, 2023. 3

[21] Deepti Hegde, Jeya Maria Jose Valanarasu, and Vishal M Patel. Clip goes 3d: Leveraging prompt tuning for language grounded 3d recognition. *arXiv preprint arXiv:2303.11313*, 2023. 2, 5

[22] 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 TOG*, 2022. 2

[23] Di Huang, Sida Peng, Tong He, Honghui Yang, Xiaowei Zhou, and Wanli Ouyang. Ponder: Point cloud pre-training via neural rendering. In *ICCV*, 2023. 2

[24] 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. 5

[25] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *CVPR*, 2022. 2

[26] Maximilian Jaritz, Jiayuan Gu, and Hao Su. Multi-view pointnet for 3d scene understanding. In *ICCVW*, 2019. 1, 3

[27] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Alaa Maalouf, Shuang Li, Ganesh Subramanian Iyer, Soroush Saryazdi, Nikhil Varma Keetha, et al. Conceptfusion: Open-set multimodal 3d mapping. In *ICRAW*, 2023. 2

[28] Zhao Jin, Munawar Hayat, Yuwei Yang, Yulan Guo, and Yinjie Lei. Context-aware alignment and mutual masking for 3d-language pre-training. In *CVPR*, 2023. 2

[29] Xin Lai, Jianhui Liu, Li Jiang, Liwei Wang, Hengshuang Zhao, Shu Liu, Xiaojuan Qi, and Jiaya Jia. Stratified transformer for 3d point cloud segmentation. In *CVPR*, 2022. 1

[30] Weixian Lei, Yixiao Ge, Jianfeng Zhang, Dylan Sun, Kun Yi, Ying Shan, and Mike Zheng Shou. Vit-lens: Towards omni-modal representations. *arXiv preprint arXiv:2308.10185*, 2023. 2

[31] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*. PMLR, 2022. 5

[32] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 5

[33] Zhenyu Li, Zehui Chen, Ang Li, Liangji Fang, Qinhong Jiang, Xianming Liu, Junjun Jiang, Bolei Zhou, and Hang Zhao. Simipu: Simple 2d image and 3d point cloud unsupervised pre-training for spatial-aware visual representations. In *AAAI*, 2022. 2

[34] Kun-Yu Lin, Henghui Ding, Jiaming Zhou, Yi-Xing Peng, Zhilin Zhao, Chen Change Loy, and Wei-Shi Zheng. Rethinking clip-based video learners in cross-domain open-vocabulary action recognition. *arXiv preprint arXiv:2403.01560*, 2024. 1

[35] Zhenxi Lin, Ziheng Zhang, Meng Wang, Yinghui Shi, Xian Wu, and Yefeng Zheng. Multi-modal contrastive representation learning for entity alignment. In *COLING*, 2022. 2

[36] 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. *NeurIPS*, 2024. 2, 3, 5, 6, 8

[37] Yongcheng Liu, Bin Fan, Shiming Xiang, and Chunhong Pan. Relation-shape convolutional neural network for point cloud analysis. In *CVPR*, 2019. 2

[38] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *ICRA*, 2023. 1, 3

[39] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2016. 6

[40] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2018. 6

[41] Yuheng Lu, Chenfeng Xu, Xiaobao Wei, Xiaodong Xie, Masayoshi Tomizuka, Kurt Keutzer, and Shanghang Zhang. Open-vocabulary point-cloud object detection without 3d annotation. In *CVPR*, 2023. 2

[42] Benedikt Mersch, Xieyuanli Chen, Jens Behley, and Cyrill Stachniss. Self-supervised point cloud prediction using 3d spatio-temporal convolutional networks. In *CoRL*, 2022. 2

[43] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *CVPR*, 2022. 2

[44] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 8

[45] OpenAI. Gpt-4 technical report. *arXiv:2303.08774*, 2023. 5

[46] 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

[47] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 2019. 6

[48] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *CVPR*, 2023. 2

[49] Petra Poklukar, Miguel Vasco, Hang Yin, Francisco S Melo, Ana Paiva, and Danica Kragic. Geometric multimodal contrastive representation learning. In *ICML*. PMLR, 2022. 2

[50] 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. 1, 2

[51] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *NeurIPS*, 2017. 1, 2

[52] 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. 2, 5

[53] Guocheng Qian, Abdulellah Abualshour, Guohao Li, Ali Thabet, and Bernard Ghanem. Pu-gcn: Point cloud upsampling using graph convolutional networks. In *CVPR*, 2021. 1, 2

[54] Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. *NeurIPS*, 2022. 1, 2, 3

[55] Guocheng Qian, Xingdi Zhang, Abdullah Hamdi, and Bernard Ghanem. Pix4point: Image pretrained transformers for 3d point cloud understanding. *3DV*, 2022. 2

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 2021. 1, 2, 3

[57] Yongming Rao, Jiwen Lu, and Jie Zhou. Global-local bidirectional reasoning for unsupervised representation learning of 3d point clouds. In *CVPR*, 2020. 2

[58] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. Clip-forge: Towards zero-shot text-to-shape generation. In *CVPR*, 2022. 2

[59] Jonathan Sauder and Bjarne Sievers. Self-supervised deep learning on point clouds by reconstructing space. *NeurIPS*, 2019. 2

[60] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 1

[61] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 2022. 1

[62] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *ICCV*, 2015. 1, 3

[63] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *NeurIPS*, 2017. 3

[64] Maxim Tatarchenko, Jaesik Park, Vladlen Koltun, and Qian-Yi Zhou. Tangent convolutions for dense prediction in 3d. In *CVPR*, 2018. 2

[65] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *ICCV*, 2019. 2

[66] 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. 4, 5, 6

[67] Haowei Wang, Jiji Tang, Jiayi Ji, Xiaoshuai Sun, Rongsheng Zhang, Yiwei Ma, Minda Zhao, Lincheng Li, Zeng Zhao, Tangjie Lv, et al. Beyond first impressions: Integrating joint multi-modal cues for comprehensive 3d representation. In *ACM MM*, 2023. 2

[68] Lei Wang, Yuchun Huang, Yaolin Hou, Shenman Zhang, and Jie Shan. Graph attention convolution for point cloud semantic segmentation. In *CVPR*, 2019. 1, 2

[69] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM TOG*, 2019. 1, 2

[70] Ziyi Wang, Xumin Yu, Yongming Rao, Jie Zhou, and Jiwen Lu. P2p: Tuning pre-trained image models for point cloud analysis with point-to-pixel prompting. *NeurIPS*, 2022. 2

[71] Zeyu Wang, Dingwen Li, Chenxu Luo, Cihang Xie, and Xiaodong Yang. Distillbev: Boosting multi-camera 3d object detection with cross-modal knowledge distillation. In *ICCV*, 2023. 1, 3

[72] 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. 4, 5

[73] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 3

[74] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *ECCV*, 2020. 2

[75] 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. 2

[76] Yifan Xu, Tianqi Fan, Mingye Xu, Long Zeng, and Yu Qiao. Spidercnn: Deep learning on point sets with parameterized convolutional filters. In *ECCV*, 2018. 2

[77] 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

[78] Le Xue, Ning Yu, Shu Zhang, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. *arXiv preprint arXiv:2305.08275*, 2023. 2, 3, 5, 6

[79] 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. 2

[80] 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, 6

[81] Yihan Zeng, Chenhan Jiang, Jiageng Mao, Jianhua Han, Chaoqiang Ye, Qingqiu Huang, Dit-Yan Yeung, Zhen Yang, Xiaodan Liang, and Hang Xu. Clip2: Contrastive language-image-point pretraining from real-world point cloud data. In *CVPR*, 2023. 1, 3

[82] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. In *ICCV*, 2023. 1

[83] Junbo Zhang, Runpei Dong, and Kaisheng Ma. Clip-fo3d: Learning free open-world 3d scene representations from 2d dense clip. In *ICCVW*, 2023. 2

[84] 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. 2, 5

[85] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *ECCV*, 2022. 1

[86] Renrui Zhang, Liuhui Wang, Yu Qiao, Peng Gao, and Hongsheng Li. Learning 3d representations from 2d pre-trained models via image-to-point masked autoencoders. In *CVPR*, 2023. 2

[87] Zhihao Zhang, Shengcao Cao, and Yu-Xiong Wang. Tamm: Triadapter multi-modal learning for 3d shape understanding. *arXiv preprint arXiv:2402.18490*, 2024. 2

[88] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *ICCV*, 2021. 1, 2

[89] Jiaming Zhou, Junwei Liang, Kun-Yu Lin, Jinrui Yang, and Wei-Shi Zheng. Actionhub: A large-scale action video description dataset for zero-shot action recognition. *arXiv preprint arXiv:2401.11654*, 2024. 1

[90] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. In *ICLR*, 2024. 2

[91] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyao Zeng, Shanghang Zhang, and Peng Gao. Pointclip v2: Adapting clip for powerful 3d open-world learning. *arXiv preprint arXiv:2211.11682*, 2022. 2, 5