

Open-Pose 3D Zero-Shot Learning: Benchmark and Challenges

Weiguang Zhao^{a,b,1}, Guanyu Yang^{f,1}, Rui Zhang^{b,*}, Chenru Jiang^f,
Chaolong Yang^{a,c}, Yuyao Yan^d, Amir Hussain^e, Kaizhu Huang^{f,*}

^a*Department of Computer Science, University of Liverpool, Liverpool L69 7ZX, UK.*

^b*Department of Foundational Mathematics, Xi'an Jiaotong-Liverpool University, Suzhou, 215123, China.*

^c*Department of Mechatronics and Robotics, Xi'an Jiaotong-Liverpool University, Suzhou, 215123, China.*

^d*School of Robotic, Xi'an Jiaotong-Liverpool University, Suzhou, 215123, China.*

^e*School of Computing, Edinburgh Napier University, Edinburgh, EH11 4BN, UK.*

^f*Data Science Research Center, Duke Kunshan University, Kunshan, 215316, China.*

Abstract

With the explosive 3D data growth, the urgency of utilizing zero-shot learning to facilitate data labeling becomes evident. Recently, methods transferring language or language-image pre-training models like Contrastive Language-Image Pre-training (CLIP) to 3D vision have made significant progress in the 3D zero-shot classification task. These methods primarily focus on 3D object classification with an aligned pose; such a setting is, however, rather restrictive, which overlooks the recognition of 3D objects with open poses typically encountered in real-world scenarios, such as an

*Corresponding authors

Email addresses: Weiguang.Zhao@liverpool.ac.uk (Weiguang Zhao),
Guanyu.Yang@dukekunshan.edu.cn (Guanyu Yang), Rui.Zhang02@xjtlu.edu.cn (Rui Zhang),
Chenru.Jiang@dukekunshan.edu.cn (Chenru Jiang),
Chaolong.Yang@liverpool.ac.uk (Chaolong Yang), yuyao.yan@xjtlu.edu.cn (Yuyao Yan),
A.Hussain@napier.ac.uk (Amir Hussain), Kaizhu.Huang@dukekunshan.edu.cn (Kaizhu Huang)

¹Equal contribution

overturned chair or a lying teddy bear. To this end, we propose a more realistic and challenging scenario named open-pose 3D zero-shot classification, focusing on the recognition of 3D objects regardless of their orientation. First, we revisit the current research on 3D zero-shot classification, and propose two benchmark datasets specifically designed for the open-pose setting. We empirically validate many of the most popular methods in the proposed open-pose benchmark. Our investigations reveal that most current 3D zero-shot classification models suffer from poor performance, indicating a substantial exploration room towards the new direction. Furthermore, we study a concise pipeline with an iterative angle refinement mechanism that automatically optimizes one ideal angle to classify these open-pose 3D objects. In particular, to make validation more compelling and not just limited to existing CLIP-based methods, we also pioneer the exploration of knowledge transfer based on Diffusion models. While the proposed solutions can serve as a new benchmark for open-pose 3D zero-shot classification, we discuss the complexities and challenges of this scenario that remain for further research development. The code is available publicly at <https://github.com/weiguangzhao/Diff-OP3D>.

Keywords: Zero-Shot, 3D Classification, Open-Pose, Text-Image Matching

1. Introduction

Deep learning models have achieved remarkable advancements in computer vision tasks (Xue et al., 2023a; Chen et al., 2023a; Zhao et al., 2023; Jiang et al., 2023). However, their outstanding performance relies heavily on large amounts of labeled data. In this regard, zero-shot learning, where

classes are learned without corresponding samples, has drawn substantial scholarly focus. While 2D zero-shot classification research (Radford et al., 2021; Wang et al., 2019; Naeem et al., 2023; Ye et al., 2023, 2021) is thriving, research on 3D zero-shot classification (Zhang et al., 2022; Zhu et al., 2023; Naeem et al., 2022) is still in nascent stages. Considering the inherent irregularity and sparsity of 3D data, the extracted features exhibit significant differences compared to 2D data. As a result, most existing 2D zero-shot learning methods fail to produce effective results when applied directly to the 3D domain (Cheraghian et al., 2019b,a, 2020, 2022).

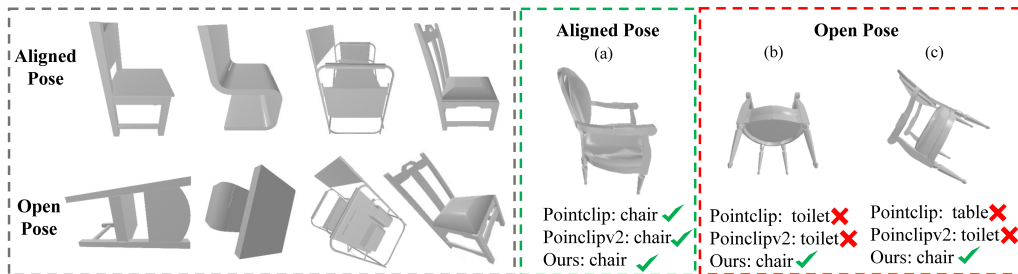


Figure 1: 3D Zero-Shot Classification for Aligned-Poses and Open-Poses. (a) is a 3D sample in aligned-pose from the dataset ModelNet40, while (b) and (c) are the corresponding sample in open-poses from our benchmark ModelNet40[‡].

Currently, to better leverage knowledge from the 2D domain, most state-of-the-art methods (SOTAs) (Zhang et al., 2022; Huang et al., 2023; Zhu et al., 2023; Xue et al., 2023a; Wang et al., 2023b) explore bridging Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021) to 3D classification. Their general pipeline is to project 3D objects to 2D depth maps and then to match the class names and image features via CLIP. As an initial exploration, these methods have already demonstrated remarkable ef-

fectiveness on existing benchmarks. However, all these evaluations are based on the assumption that 3D objects are in aligned-poses. Such alignment is based on prior category knowledge, thus inherently simplifying the task of classification. Unfortunately, objects in the real world are typically positioned in random orientations, which can be referred to as open-pose, such as an overturned chair or a flying bird. These open-pose scenarios present a realistic yet significant challenge for current SOTAs. As shown in Fig. 1, the effectiveness of these SOTAs is closely tied to the objects' orientations, and their performance becomes poor when the aligned 3D object is modified to the open-pose case.

To this end, we investigate a more challenging, yet practical task of 3D zero-shot classification aiming at recognizing 3D objects in open-poses. Despite its importance, this open-pose setting is rarely studied in the literature. To this end, we first revisit the current methods and datasets for 3D zero-shot classification. Furthermore, we propose to generate two new datasets to benchmark open-pose 3D zero-shot classification, by applying random rotations to each sample in the widely-used zero-shot datasets ModelNet40 (Wu et al., 2015) and McGill (Siddiqi et al., 2008). Due to the uncertain orientations of objects, the selection of projection angles becomes a crucial matter. In this context, we propose a concise pipeline with an iterative angle refinement mechanism that automatically optimizes one ideal project angle to classify these open-pose 3D objects. Specifically, this mechanism dynamically determines the projection angles based on the characteristics of the input object and the outcomes of matching processes which shows potential in handling open-pose variations.

Additionally, it is notable that existing projection-based approaches exclusively use CLIP as their text-image matching backbone, and their projections are primarily limited to depth maps and rendered images. To achieve more comprehensive evaluations, we make the first attempt to utilize a 2D pre-trained diffusion model (Rombach et al., 2022; Sohl-Dickstein et al., 2015) as the backbone for classifying 3D objects. We also incorporate the edge image, thereby expanding the variety of projection styles.

While the above pipeline can serve as the first as well as a new benchmark for open-pose 3D zero-shot classification, we also discuss the challenges and potential outlooks for 3D zero-shot classification in the open-pose setting, which may further inspire future work in this direction.

The contributions of our work can be summarized as follows:

- We propose a more challenging scenario, namely open-pose 3D zero-shot classification, to uncover the current limitations of state-of-the-art approaches. We survey a comprehensive set of 3D zero-shot classifications in the open-pose setting for the first time.
- We develop two benchmarks, ModelNet40[‡] and McGill[‡] for evaluating open-pose 3D zero-shot classification. Our empirical investigations reveal that most current SOTA 3D zero-shot methods suffer from poor performance for open-pose classification.
- We design a concise pipeline with an iterative angle refinement mechanism, which achieves a substantial improvement, thus presenting a new benchmark method in the open-pose setting.
- We discuss the challenges for open-pose 3D zero-shot classification, and

set out new directions which may inspire future research towards this direction.

2. Related Work

Classification in the 3D Domain: The mainstream deep learning methods for 3D classification can be categorized into three types (Guo et al., 2020): point-based, voxel-based, and multi-view-based methods. PointNet (Qi et al., 2017) was the pioneering network to extract point cloud features and recognize 3D objects. Point-Transformer (Zhao et al., 2021; Wu et al., 2022) utilizes transformer architecture in the 3D domain, leading to a significant improvement in classification performance. Additionally, Voxnet (Maturana and Scherer, 2015) introduces voxel representation to process point clouds, while MinkowskiNet (Choy et al., 2019) offers a sparse voxel convolution architecture based on this concept. Moreover, MVCNN (Su et al., 2015) suggests projecting 3D mesh categories from multiple views and employing 2D networks for training. LRMV (Yang and Wang, 2019) predicts the score of each projection view to select images. Some studies (Kanezaki et al., 2018; Wei et al., 2021; Zhou et al., 2024) have also delved into how viewpoint selection can influence classification results. Unlike these supervised learning methods, our work is focused on zero-shot classification, where new (unseen) classes emerge during testing.

Zero-Shot Classification in the 2D Domain: The objective of conventional zero-shot learning tasks is to classify a sample set from unseen categories (Larochelle et al., 2008). Presently, there are two primary approaches to addressing zero-shot classification tasks in the 2D domain (Yang

et al., 2022): embedding methods (Rahman et al., 2020; Zhou et al., 2022; Han et al., 2022; Ye et al., 2023), and generative methods (Li et al., 2023; Yang et al., 2023). The former seeks to learn a specific feature space that aligns with both the samples and category descriptions, while the latter transforms the zero-shot classification task into a standard classification problem by generating pseudo samples for unseen categories.

Furthermore, both CLIP (Radford et al., 2021) and the diffusion classifier (Li et al., 2023) can perform classification tasks without relying on the training set of the specific benchmark. Consequently, these methods are also considered zero-shot models. However, it’s important to note that their pre-training dataset may include samples corresponding to classes within the target dataset. Therefore, strictly speaking, these methods differ from conventional zero-shot models in classifying “unseen classes”. Nonetheless, when applied to assist in the zero-shot classification of 3D data, they can be accurately categorized as zero-shot methods, as no 3D samples are used during the training process.

3. Revisiting 3D Zero-Shot Classification

In the realm of 3D zero-shot classification, two broad methods have emerged: language pre-training model-based and language-image pre-training model-based. The former relies on Word2Vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014) to generate text embeddings for categories, while the latter primarily uses CLIP to match text embeddings with projected image embeddings. While Cheraghian et al. (Cheraghian et al., 2022) provided a comprehensive review of the language pre-training model-based

method, recent years have witnessed a surge in new 3D zero-shot classification approaches, particularly those leveraging the language-image pre-training model-based paradigm. In this section, we’ll delve into a detailed overview and analysis of existing 3D zero-shot datasets and methodologies.

3.1. 3D Zero-Shot Classification Benchmark Datasets.

The existing 3D zero-shot classification benchmark datasets primarily consist of ModelNet40 (Wu et al., 2015), ModelNet10 (Wu et al., 2015), McGill (Siddiqi et al., 2008), ScanObjectNN (Uy et al., 2019), and SHREC2015 (Lian et al., 2015). In Table 1, we summarize the division of these datasets into seen and unseen classes, as well as their partitioning into training, validation, and test sets.

Datasets	Total Classes	Seen/Unseen Classes	Train/Valid/Test Samples
ModelNet40	40	30/-	5852/1560/-
ModelNet10	10	-/10	-/-/908
McGill	19	-/14	-/-/115
ScanObjectNN	15	-/11	-/-/495
SHREC2015	50	-/30	-/-/192

Table 1: 3D Zero-Shot Classification Benchmark Datasets.

ModelNet40 & ModelNet10: ModelNet40 is widely used in various 3D classification tasks, including 3D full-supervision, weakly-supervision, and few-shot classification. It consists of 40 categories of 3D objects, with each category containing a varying number of CAD models. The distribution of models across categories is balanced. This dataset encompasses a diverse array of common object classes, including chairs, tables, airplanes, cars, and

more. Furthermore, ModelNet10 is a subset of ModelNet40, comprising only 10 categories of 3D objects. Most existing 3D zero-shot classification methods (Narayan et al., 2020; Michele et al., 2021; Cheraghian et al., 2022; Hao et al., 2023) treat the 30 classes of ModelNet40 as seen classes and the remaining 10 classes as unseen classes, namely ModelNet10. It’s worth noting that some language-image pre-training model-based methods (Xue et al., 2023a; Zhu et al., 2023; Qi et al., 2023; Huang et al., 2023) consider all classes as unseen classes, as they directly utilize language-image pre-training models for inference without the need for training on seen classes.

McGill: McGill dataset differs from ModelNet40 in its composition, focusing on various 3D biological samples such as fish, dinosaurs, spiders, octopuses, and more. Specifically, the McGill dataset comprises a total of 19 categories, with five categories overlapping with the visible classes in ModelNet40. Consequently, only the remaining 14 categories are utilized as unseen classes. Due to its high-quality point clouds and a wide range of object categories, this dataset is widely adopted for numerous 3D zero-shot classification methods (Narayan et al., 2020; Michele et al., 2021; Cheraghian et al., 2022; Hao et al., 2023).

ScanObjectNN: ScanObjectNN dataset comprises 15 categories of 3D objects derived from real-world scans, often presenting cluttered backgrounds and partial objects due to occlusions. These objects predominantly stem from indoor scenes. With four classes overlapping with ModelNet40 excluded, the remaining 11 classes are regarded as unseen classes. Due to varying qualities of point clouds and backgrounds, this dataset is frequently divided into three subsets: OBJ_ONLY, OBJ_BG, and PB_T50_RS.

SHREC2015: This dataset was initially introduced at the Eurographics workshop for a 3D object retrieval competition. It comprises a total of 50 categories, with 20 of these categories overlapping with the seen classes in ModelNet40. Consequently, the remaining 30 categories are utilized as unseen classes. Moreover, SHREC2015 remains relatively underutilized in the zero-shot classification domain, with only three studies [Cheraghian et al. \(2019b, 2020, 2019a\)](#) conducted for validation on this dataset.

However, most of these datasets were deliberately aligned or oriented during collection. This makes 3D zero-shot classification tasks detached from complex real-world scenarios, where objects may be in arbitrary poses. In light of this, we make the first attempt to build an open-pose benchmark where all unseen 3D objects are in arbitrary poses.

Taking into account the coverage and frequency of dataset usage, we decided to separately utilize the ModelNet40 and McGill datasets to create the open-pose datasets, referred to as ModelNet40[‡] and McGill[‡]. We rotate each sample in ModelNet40 and McGill to obtain the ModelNet40[‡] and McGill[‡] datasets, respectively². Although the samples in McGill inherently have less standardized orientation compared to those in ModelNet40, the range of their rotational angles is limited, lacking sufficient randomness to effectively illustrate the open-pose issue. Therefore, we also apply random angle rotations to them. Similar to previous datasets ([Wu et al., 2015](#); [Siddiqi et al., 2008](#)), we divide our open-pose datasets as shown in Tab. 2.

²The random angles and open-pose datasets are available on our GitHub.

Datasets	Total Classes	Seen/Unseen Classes	Train/Valid/Test Samples
ModelNet40 [‡]	40	30/-	5852/1560/-
ModelNet10 [‡]	10	-/10	-/-/908
McGill [‡]	19	-/14	-/-/115

Table 2: Our Open-Pose 3D Zero-Shot Classification Benchmark Datasets.

3.2. 3D Zero-Shot Classification with Language Pre-training Model

The language pre-training model-based methods draw inspiration from existing 2D zero-shot approaches, using text embeddings to establish correlations between seen and unseen classes. Cheraghian et al. (Cheraghian et al., 2019a,b, 2020, 2022) pioneered research on 3D zero-shot classification. They made the first attempt in traditional (Cheraghian et al., 2019b), inductive (Cheraghian et al., 2019a), and transductive (Cheraghian et al., 2020) zero-shot classification settings within the 3D domain, providing a standardized evaluation protocol for subsequent research work. Their proposed methods are based on text embeddings (Word2Vec (Mikolov et al., 2013) & GloVe (Pennington et al., 2014)) with two refinement loss functions (Cheraghian et al., 2022). On the other hand, 3DCZSL (Naeem et al., 2022) takes an approach that considers the 3D zero-shot classification task from the perspective of geometric structure composition. However, it relies on point-wise component labels, which currently can only be satisfied by the PartNet dataset (Mo et al., 2019). 3DGenZ (Michele et al., 2021) is proposed as the first generative 3D zero-shot classification network. It conducts Google searches for 100 images per class and utilizes a pre-training classification network to obtain image representations for each category.

3.3. 3D Zero-Shot Classification with Language-Image Pre-training Model

We make the first attempt to provide an overview of the language image pre-training model-based methods. Currently, all these methods utilize CLIP’s knowledge for zero-shot classification of 3D models. Based on their implementation approaches, we categorize them into two groups: input-optimization and encoder-distillation methods.

Input-optimization methods. These methods typically involve optimizing the text input of CLIP or rendering image inputs, as depicted in Fig. 2. PointCLIP (Zhang et al., 2022) pioneered the projection of 3D objects into 2D images and utilized CLIP for category matching. Building upon this work, PointCLIPv2 (Zhu et al., 2023) incorporates large-scale language models (such as GPT-3) (Brown et al., 2020) to automatically design more descriptive 3D-semantic prompts. Furthermore, DiffCLIP (Shen et al., 2024) proposes integrating stable diffusion with ControlNet (Zhang et al., 2023) to minimize the domain gap between rendered images and realistic images. Additionally, DILF (Ning et al., 2024) utilizes GPT-3 to generate textual prompts enriched with 3D semantics and designs a differentiable renderer with learnable rendering parameters to produce representative multi-view images. Most of these methods do not require additional training and perform direct inference using existing pre-trained models. However, they often require a considerable number of hyperparameters to achieve the best results.

Encoder-distillation methods. As depicted in Fig. 4, these methods retain the text encoder of CLIP while incorporating a visual encoder to supervise their newly designed encoder. Specifically, Ulip (Xue et al., 2023a,b) and CLIPgoes3D (Hegde et al., 2023) design a new 3D encoder

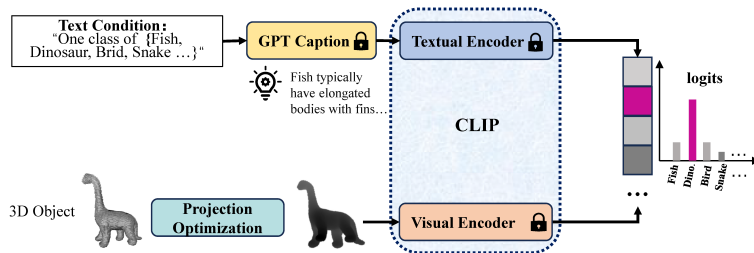


Figure 2: Input-optimization Framework

to learn the relationship between 3D features, rendered images, and textual descriptions. Furthermore, ReconCLIP (Qi et al., 2023) leverages reconstruction tasks to enhance the feature extraction of 3D encoders. Moreover, CLIP2Point (Huang et al., 2023) develops a new image encoder network to narrow the domain gap between depth maps and realistic images. Methods of this kind often require additional 3D data for pre-training, limiting their applicability to the domain of the training data. Furthermore, they heavily rely on 3D data, resulting in significant computational overhead. For instance, Ulip and CLIPgoes3D both require training on 8 A100 GPUs.

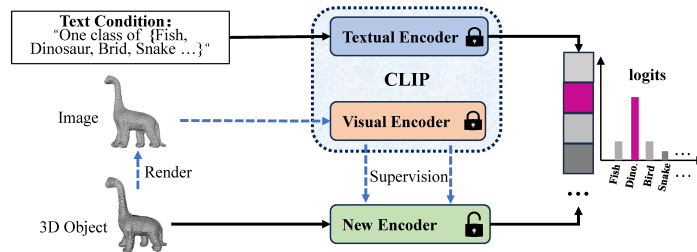


Figure 3: Encoder-distillation Framework

However, both input-optimization and encoder-distillation methods are built upon the assumption that objects are in aligned poses. As depicted in

Fig. 4, the more practical and challenging task of recognizing 3D objects in open poses is neglected. In our work, we explore the performance of many existing methods in an open-pose setting and propose a new method tailored specifically for this scenario as a baseline.

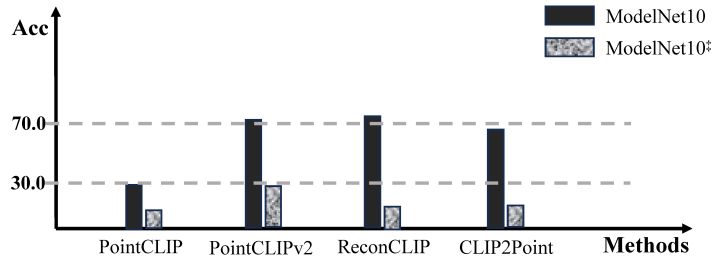


Figure 4: Performance on Aligned and Open-Pose Dataset

4. Our Main Methodology

As depicted in Fig. 5, our pipeline consists of three main components: Projection (a), Text-Image Matching (b), and Angle Selection (c). Given a 3D object, projection images can be obtained by selecting projection angles and styles. These images are then matched with text descriptions via a pre-training text-image matching backbone. Additionally, the projection angles can be pre-selected or determined based on the input object and matching outcomes. With the final settled projection angles, predictions are obtained based on the corresponding matching scores. Furthermore, (e) and (f) are optional text-image matching backbones based on Diffusion and CLIP, respectively. In this section, we provide detailed descriptions of this pipeline.

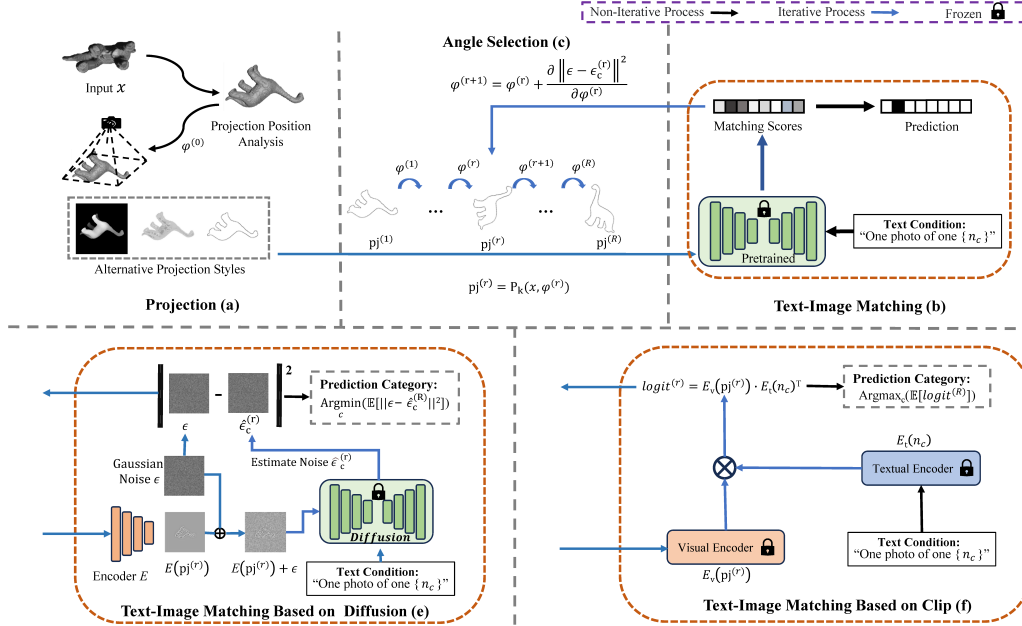


Figure 5: Overview of Our Pipeline

4.1. Projection

We utilize a perspective projection to transform the 3D input x into 2D images. In the projection phase, our main focus lies on two aspects: projection position analysis and projection style selection. Here, we concentrate on exploring various styles for projection. The analysis of projection positions, closely linked with angle selection, will be elaborated in Section 4.3.

In addition to the depth maps and rendered images commonly used in existing works, we introduce edge maps as an additional projection style. Given that the primary data representations for 3D objects are point clouds and meshes, our method considers both inputs. Due to the sparsity of point clouds, their projection style differs significantly from that of meshes. Hence, we convert images projected from both point clouds and meshes into edge im-

ages for zero-shot classification. As depicted in Fig. 6, the two different input data formats yield distinct projected images, which are ultimately extracted into similar edge images.

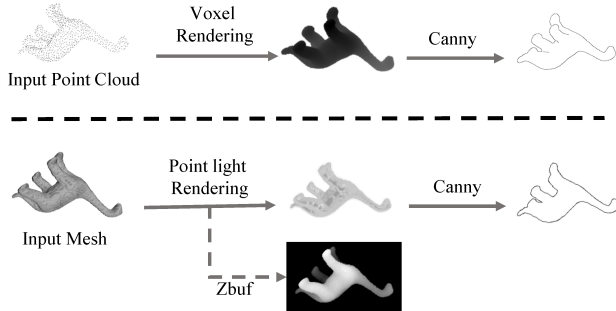


Figure 6: Projection Styles

Specifically, when the input is point-cloud data, direct projection yields a set of discrete pixels, making it challenging to adequately represent its semantic information. Following the approach of PointCLIPv2 (Zhu et al., 2023), we adopt voxel projection, which involves four steps: voxelization, densification, smoothing, and squeezing, to obtain a continuous pixel projection image. Furthermore, we utilize the Canny algorithm to compute pixel gradients, thereby obtaining the edge image. On the other hand, mesh data includes additional face information, specifically triangular surfaces, compared to point-cloud data. In this regard, we can directly achieve the continuous pixel projection image by point light projecting. To simulate parallel light, we position the point light source far away from the camera and the 3D mesh. To address noise on the surface of mesh data, we increase the Canny gradient threshold (Canny, 1986) to mitigate this impact, which also results in the removal of some detailed information. For instance, the edge image of

the mesh lacks the boundary of the legs compared to the edge image of the point cloud in Fig. 6.

4.2. Text-Image Matching

Since CLIP is originally designed for text-image matching, its performance in 3D zero-shot classification tasks has been extensively explored in numerous studies. Conversely, the diffusion process has not yet been applied to this area. Therefore, we revisit how the diffusion process, as a generative model, accomplishes this task.

Given a clean feature sample \mathbf{f}_0 and a variance schedule β_1, \dots, β_T , we can define a Markov chain that gradually adds Gaussian noise to the data as follows:

$$q(\mathbf{f}_t | \mathbf{f}_{t-1}) = \mathcal{N}(\mathbf{f}_t; \sqrt{1 - \beta_t} \mathbf{f}_{t-1}, \beta_t \mathbf{I}). \quad (1)$$

The denoising diffusion probabilistic models (Rombach et al., 2022) learn the reverse process $p_\theta(\mathbf{f}_{t-1} | \mathbf{f}_t, s)$ with the corresponding semantic description s . Typically, the diffusion model can be interpreted as a denoising autoencoder $\epsilon_\theta(\mathbf{f}_t, t, s)$ with a training target that minimizes the objective function:

$$L_{DM} = \mathbb{E}_{\mathbf{f}_0, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t} [\|\epsilon - \epsilon_\theta(\mathbf{f}_t, t, s)\|_2^2]. \quad (2)$$

In the Diffusion Classifier (Li et al., 2023), the posterior probability of the latent feature of the sample conditioned on the specific semantic is calculated based on the denoising performance, thus enabling the classification task. With a semantic description constructor $S(c)$ for each class and a 2D feature extractor $\epsilon_\theta()$, such probability can be approximately estimated as follows:

$$p_\theta(c | \mathbf{f}_0) = \frac{\exp\{-\mathbb{E}_{t, \epsilon} [\|\epsilon - \epsilon_\theta(\mathbf{f}_t, t, S(c))\|_2^2]\}}{\sum_j \exp\{-\mathbb{E}_{t, \epsilon} [\|\epsilon - \epsilon_\theta(\mathbf{f}_t, t, S(j))\|_2^2]\}}. \quad (3)$$

Here, we directly adopt such a pre-training diffusion framework with the encoder $E(\cdot)$. Denoting the multi-style projection process as the function $P_k(\cdot, \varphi)$ with style $k \in K$ and angles φ , we define the text-image matching score with the latent denoising module as below:

$$\begin{aligned}
 MS_{\theta, K}(\mathbf{x}, \varphi, c) &= \exp \left\{ -\mathbb{E}_{t, \epsilon, k} \left[\|\epsilon - \hat{\epsilon}_{\theta, k}(\mathbf{x}, \varphi, c)\|_2^2 \right] \right\} \\
 \hat{\epsilon}_{\theta, k}(\mathbf{x}, \varphi, c) &= \epsilon_{\theta}(\mathbf{f}_{t, k}(\mathbf{x}, \varphi), t, S_k(c)), \\
 \mathbf{f}_{t, k}(\mathbf{x}, \varphi) &= \sqrt{\bar{\alpha}_t} E(P_k(\mathbf{x}, \varphi)) + \sqrt{1 - \bar{\alpha}_t} \epsilon,
 \end{aligned} \tag{4}$$

where following (Ho et al., 2020; Rombach et al., 2022), $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_i^t \alpha_i$ is defined to construct the noised feature for timestep t . The closer the predicted noise to the actual noise, the higher the matching score. For consistency in representation, the output of CLIP can also be considered as the matching score. Then, the following estimated probability and prediction for a 3D point cloud sample under a single projection angle can be attained:

$$p_{\theta, K}(c|\mathbf{x}, \varphi) = \frac{MS_{\theta, K}(\mathbf{x}, \varphi, c)}{\sum_j MS_{\theta, K}(\mathbf{x}, \varphi, j)} \tag{5}$$

$$\hat{y} = \arg \max_c p_{\theta, K}(c|\mathbf{x}, \varphi). \tag{6}$$

More details about the constructed semantic descriptions for different projection types can be found in Section 5.3.

4.3. Angle Selection

To enhance the adaptability of the text-image matching framework to unaligned 3D point cloud data, the appropriate projection angles are indispensable. The choice of projection angles can be either predefined fixed angles, derived from the information of the 3D object, or refined according to preliminary matching scores.

4.3.1. Pre-defined Fixed Angles

Besides pre-selected single angles, existing pre-training-based 3D zero-shot classification methods only adopt circular and cube projections. Concerning circular projection, the camera is positioned obliquely above the object, which is adjusted by an angle φ_1 measuring the camera’s inclination with respect to the x-y plane. Multi-view images are obtained by changing the azimuthal angle φ_2 of the camera’s projection on the x-y plane relative to the x-axis. On the other hand, cube projection places the camera along the six directions of the three-dimensional coordinate axis. All cameras are oriented towards the center of the object and at a distance r_p from the object. Both circular and cube projections keep the object pose unchanged, only adjusting the camera position, as shown in Fig. 7. In this context, these two methods can achieve suitable views for open-pose 3D objects by fine-tuning the hyperparameters.

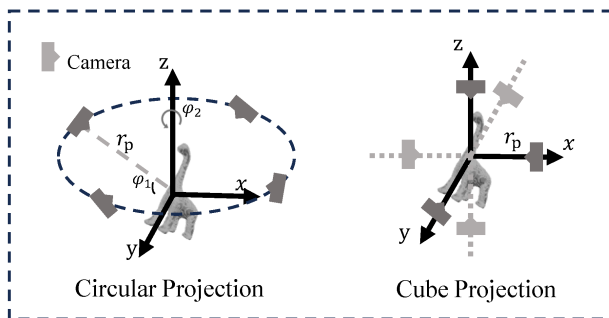


Figure 7: Basic Projection Methods

4.3.2. Iterative Angle Refinement Mechanism (IARM)

To enhance the identification of optimal projection angles, we propose an advanced method that integrates information from 3D objects and text-

image matching. In instances where improper projection angles result in significantly low matching scores for specific classes, a high prediction probability does not guarantee a reasonable outcome. Thus, we introduce an iterative angle refinement module utilizing the projection angles $\boldsymbol{\varphi}_c = [\varphi_{c,1}, \varphi_{c,2}]$, aimed at maximizing the text-image matching scores for each class c . This proposed module is expected to yield a more reasonable estimation of probability as depicted below:

$$p_{\boldsymbol{\theta},K}^*(c|\mathbf{x}) = \frac{MS_{\boldsymbol{\theta},K}(\mathbf{x}, \boldsymbol{\varphi}_c, c)}{\sum_j MS_{\boldsymbol{\theta},K}(\mathbf{x}, \boldsymbol{\varphi}_j, j)} \quad (7)$$

$$\boldsymbol{\varphi}_c = \arg \max_{\boldsymbol{\varphi}} MS_{\boldsymbol{\theta},K}(\mathbf{x}, \boldsymbol{\varphi}, c). \quad (8)$$

Assuming both the projection and text-image matching processes are differentiable, we can ascertain the projection angles yielding the highest matching score for each class through the gradient descent algorithm. Starting with an initial projection angle vector $\boldsymbol{\varphi}^{(0)}$, the entire iterative optimization process, incorporating a varying scaling factor $[\boldsymbol{\eta}_r]$, is outlined as follows:

$$\begin{aligned} \boldsymbol{\varphi}_c^{(0)} &= \boldsymbol{\varphi}^{(0)} \\ \boldsymbol{\varphi}_c^{(r+1)} &= \boldsymbol{\varphi}_c^{(r)} + \boldsymbol{\eta}_r \cdot \text{sign} \left(\frac{\partial MS_{\boldsymbol{\theta},K}(\mathbf{x}, \boldsymbol{\varphi}_c^{(r)}, c)}{\partial \boldsymbol{\varphi}_c^{(r)}} \right). \end{aligned} \quad (9)$$

Upon completion of R optimization steps, we utilize the estimated affine angle of each category $\widehat{\boldsymbol{\varphi}}_c = \boldsymbol{\varphi}_c^{(R)}$ to formulate the final classification prediction for the 3D point cloud samples, as outlined below:

$$\widehat{y}^* = \arg \max_c \frac{MS_{\boldsymbol{\theta},K}(\mathbf{x}, \widehat{\boldsymbol{\varphi}}_c, c)}{\sum_j MS_{\boldsymbol{\theta},K}(\mathbf{x}, \widehat{\boldsymbol{\varphi}}_j, j)}. \quad (10)$$

The prediction process aims to minimize the global risk associated with decision-making. Specifically, the confidence assigned to the angle most conducive to the correct classification of an individual object should surpass the confidences of other classes across all angles.

In practice, the distribution of the matching score computed by the pre-training model often exhibits non-smoothness within the projection angle space. Additionally, certain semantic image matching models, such as the Stable Diffusion utilized in this study, entail a considerable number of parameters, rendering optimization impractical and computationally expensive in such scenarios.

To streamline this optimization process, we propose a succinct strategy. Initially, we normalize and translate the sample coordinates to the coordinate origin to derive the covariance matrix Σ :

$$\Sigma = (\mathbf{x}^T \cdot \mathbf{x})/N, \quad \Sigma \in \mathbb{R}^{3 \times 3}, \quad (11)$$

where \mathbf{x} represents the coordinates and N denotes the number of vertices in a sample. By performing eigenvalue decomposition on this matrix Σ , we can derive the three eigenvectors: $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3 \in \mathbb{R}^{1 \times 3}$. Subsequently, we establish a new coordinate system for the sample, yielding the transformed coordinates as follows:

$$\mathbf{x}' = \mathbf{x} \cdot [\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3]. \quad (12)$$

According to principal component analysis, fixing the camera on the z-axis in the new coordinates ensures the largest variance in each dimension of the projected image. This perspective may result in capturing more classification information. Furthermore, this dimensional reduction operation

Algorithm 1: IARM on Diffusion Classifier

Data: 3D objects with open-pose: \mathbf{x}

Semantic description for each class: $S(c)$

Pre-training diffusion framework: $\epsilon_{\theta}(), E()$

Project function for each style: $P_k()$

Number of iterations: R

Class number: C

Refine scales: $[\eta_r]$

Distance from the camera to the object: r_p

- 1: Calculate the covariance matrix Σ through Eq. 11 and calculate its eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$
 - 2: Establish a new coordinate system and obtain the coordinates \mathbf{x}' through Eq. 12
 - 3: **for** c **in** $\text{range}(C)$ **do**
 - 4: Initialize the projection angle $\varphi_c^{(0)} = 0$
 - 5: **for** r **in** $\text{range}(R)$ **do**
 - 6: Sample random Gaussian noises ϵ and calculate the matching score $MS_{\theta,K}(\mathbf{x}', \varphi_c^{(r)}, c)$ through Eq. 4
 - 7: Calculate the partial derivative of matching score and update the $\varphi_c^{(r)}$ through Eq. 9
 - 8: **end for**
 - 9: Calculate the matching score $MS_{\theta,K}(\mathbf{x}', \varphi_c^{(R)}, c)$ through Eq. 4
 - 10: **end for**
 - 11: Obtain the final prediction \hat{y}^* through Eq. 10
 - 12: **return** predicted label \hat{y}^*
-

reduces the projection angles variable from two to one dimension (only the azimuthal angle $\varphi_{c,2}$). Instead of directly applying this rotation angle to the 3D data, it can be equivalently applied to the projected image. Following this initial adjustment of the projection angle, we can proceed to refine the angle through the previously introduced mechanism within a more simplified single-dimension case. The detailed steps for the proposed angle refinement strategy are outlined in Algorithm 1.

5. Experiment

5.1. Experiment Setting

Evaluation Metric. Following the SOTA approaches (Cheraghian et al., 2022; Hao et al., 2023; Qi et al., 2023), we utilize top-1 accuracy and employ mAcc to calculate the average accuracy across all categories, thereby providing a comprehensive reflection of the model’s classification performance across different categories. Given that the pre-training-based methods discussed in this section do not necessitate training on native 3D datasets, there are no designated ”seen” classes. To ensure a fair and uniform comparison, we exclusively evaluate the performance of conventional zero-shot classification among unseen classes.

Implementation Details. We conduct our end-to-end inference process on a single RTX3090 card. The Stable Diffusion model (Rombach et al., 2022) is employed as the 2D pre-training model to predict the noise. Following Stable Diffusion, we set the seed to 42. For the hyperparameters of the diffusion classifier, we empirically tune the time steps and trials to 600 and 30, respectively. Additionally, we adjust the projected camera dis-

tance r_p to 2.2 and the angle refinement parameters R and $[\boldsymbol{\eta}_r]$ to 10 and $[20, 18, 16, \dots, 2]$, respectively. Regarding projection, we utilize single-point light source projection from PyTorch3D (Ravi et al., 2020) for mesh samples and voxel projection for point-cloud samples.

5.2. Analysis on Open-Pose 3D Zero-Shot Classification

5.2.1. Comparison to SOTAs

We evaluate five recent SOTA methods: PointCLIP (Zhang et al., 2022), Ulip (Xue et al., 2023a), ReconCLIP (Qi et al., 2023), CLIP2Point (Huang et al., 2023), and PointCLIPv2 (Zhu et al., 2023), on our open-pose benchmark McGill[‡] and ModelNet10[‡] for 3D zero-shot classification. All reproduction codes and pre-trained models are obtained from the official GitHub repository of the respective papers. The current state-of-the-art methods employ various pre-training models for zero-shot classification, such as GPT, CLIP, etc. Detailed information and results are provided in Table 3. “Ours-CLIP” and “Ours-Diffusion” indicate that we utilize CLIP and Diffusion as the pre-training text-image matching models in our pipeline (see Figure 5), respectively.

Results on the Open-Pose 3D Zero-Shot Benchmark McGill[‡]. We present the results of our method on the McGill[‡] benchmark in Table 3. Based on our pipeline, both Ours-CLIP and Ours-Diffusion exhibit compelling performance. Particularly, Ours-Diffusion demonstrates notable improvements of 11.3% and 15.8% on Accuracy (Acc) and mean Accuracy (mAcc), respectively. Furthermore, our approach relies solely on a single pre-training model (CLIP or Diffusion), rendering it a concise solution for initial open-pose 3D zero-shot classification.

	Venue	Pre-training				McGill [‡]		ModelNet10 [‡]	
		GPT	CLIP	Diffusion	3D-pm	Acc	mAcc	ACC	mACC
PointCLIP	CVPR'22		✓			12.2	13.3	17.7	16.4
Ulip	CVPR'23		✓		✓	14.8	16.1	14.4	13.8
ReconCLIP	ICML'23		✓		✓	15.7	17.3	15.6	14.3
CLIP2Point	ICCV'23		✓		✓	14.8	17.4	19.7	18.1
PointCLIPv2	ICCV'23	✓	✓			27.8	28.9	19.9	18.2
Ours-CLIP	-		✓			<u>31.3</u>	<u>34.6</u>	26.3	24.2
Ours-Diffusion	-			✓		39.1	44.7	<u>22.6</u>	<u>21.7</u>

Table 3: Comparison to Current SOTAs on the Open-Pose 3D Zero-Shot Classification. 3D-pm stands for the **3D** pre-training model.

Results on the Open-Pose 3D Zero-Shot Benchmark ModelNet10[‡].

As indicated in Table 3, our method also surpasses SOTAs with improvements of 6.4% and 6.0% on Accuracy (Acc) and mean Accuracy (mAcc), respectively. Unlike the results on McGill[‡], the performance of Ours-Diffusion on ModelNet10[‡] is lower than that of Ours-CLIP. We observe that CLIP exhibits particularly high accuracy in the 'toilet' and 'monitor' categories, whereas Diffusion demonstrates more evenly distributed performance across all categories. Due to the smaller number of classes, CLIP achieves better results on ModelNet10[‡]. Since CLIP utilizes feature similarity for measurement while Diffusion adopts noise MSE distance, it is challenging to directly compare or ensemble these two methods. In future work, we will explore methods to effectively combine CLIP and Diffusion to yield robust inference results. Notably, samples in these datasets tend to be boxy and symmetrical, with minimal variance between some classes, thus posing a challenging task in the open-pose scenario.

5.2.2. Analysis on Views Selection

We compare our Iterative Angle Refinement Mechanism (IARM) with commonly used projection methods, namely cube and circular views. The visualization results of these methods for a single sample, including angles, projections, and final predictions, are depicted in Figure 8. Additionally, we incorporate a fixed single-view approach utilizing only the top view for further context. The results, detailed in Table 4, reveal significant disparities. In the open-pose scenario, the fixed single-view perspective yields notably poor results, with an accuracy of merely 6.6%. While the cube and circular methods, as multi-view ensemble approaches, do show improvement over a single-view perspective, their performance is still hindered by the inherent randomness in open poses. In contrast, our angle refinement mechanism offers a more advantageous approach to selecting views conducive to classification. It leads to substantial gains, as evidenced by our method showing 10.5% and 10.0% improvements in Accuracy (Acc) and mean Accuracy (mAcc), respectively, on the open-pose McGill[‡] benchmark.

	Ant	Bird	Crab	Dino.	Dolp.	Fish	Hand	Octo.	Plier	Quad.	Snake	Spec.	Spider	Teddy	Acc	mAcc
Top View	0.0	28.6	0.0	0.0	0.0	0.0	0.0	0.0	42.9	0.0	0.0	11.1	9.1	0.0	6.1	6.6
Cube	0.0	57.1	0.0	0.0	100.0	0.0	14.3	25.0	85.7	0.0	66.7	0.0	18.2	0.0	21.7	26.2
Circular	0.0	57.1	0.0	0.0	75.0	12.5	57.1	37.5	100.0	9.1	55.6	0.0	9.1	0.0	25.2	29.5
Ours-IARM	0.0	71.4	10.0	0.0	100.0	12.5	71.4	37.5	71.4	27.3	77.8	0.0	45.5	28.6	35.7	39.5

Table 4: Projection Angle Update on the McGill[‡]. In order to reduce the length of the table, we take abbreviations for some category names: Dino. vs Dinosaur, Dolp. vs Dolphin, Octo. vs Octopus, Quad. vs Quadruple, Spect. vs Spectacle.

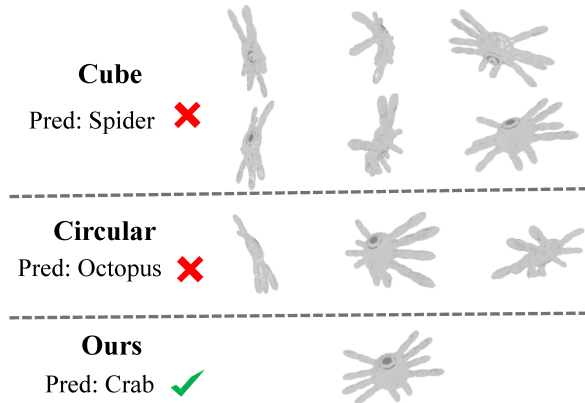


Figure 8: Final Views with the Corresponding Prediction.

5.3. Analysis on Bridging Pre-training Diffusion

Unlike current SOTAs, we are the first to utilize 2D pre-training with diffusion as the 3D zero-shot classifier instead of CLIP. In the preceding section, we presented its final performance in the open-pose setting. Given the complexity inherent in the open-pose setting, we validate its performance on the generic aligned-pose ModelNet10 dataset to provide insights into its potential in detail.

5.3.1. Comparison on Prompts

First, we investigate the influence of semantic descriptions on the matching of various styled projections. We set the trials and camera distance r_p to 30 and 2.2, respectively. We design multiple prompts for three style images: Render Image (Render I.), Depth Image (Depth I.), and Edge Image (Edge I.), and report the corresponding mAcc results on ModelNet10 in Table 5. Specifically, “one line-drawn $[n_c]$ ” is the optimal prompt for depth images, “one model of $[n_c]$ in linear composition” is the best prompt for ren-

der images, while “one edge map of one standalone $[n_c]$ ” is the most effective prompt for edge images. Moreover, the edge images exhibit the best results among the three types of images.

Prompts	Render I.	Depth I.	Edge I.
<i>one model of $[n_c]$</i>	69.3	58.8	72.3
<i>one line-drawn $[n_c]$</i>	69.9	59.8	72.0
<i>one photo of one $[n_c]$</i>	69.6	53.2	73.7
<i>one photo of one standalone $[n_c]$</i>	63.4	51.6	71.1
<i>one depth map of one standalone $[n_c]$</i>	67.2	48.5	77.4
<i>one edge map of one standalone $[n_c]$</i>	66.1	51.9	77.6
<i>one render image of one standalone white $[n_c]$</i>	68.0	55.6	66.3
<i>one sketch photo of one standalone white $[n_c]$</i>	65.8	53.0	74.7
<i>one model of $[n_c]$ in linear composition</i>	73.5	59.5	73.5
<i>one photo of one $[n_c]$ in linear composition</i>	73.1	59.7	73.7

Table 5: Prompts for Multiple Style Images. $[n_c]$ stands for the name text of each category.

5.3.2. Comparison to the CLIP

Subsequently, we compare the effectiveness of the diffusion classifier with that of the CLIP classifier on ModelNet10. To ensure a fair comparison, we only exchange the classifier while keeping all other conditions the same. The comparison results in terms of mAcc are reported in Table 6. For CLIP, we utilize four common pre-training model structures: CLIP-VIT-B\16 (Dosovitskiy et al., 2021), CLIP-VIT-B\32 (Dosovitskiy et al., 2021), CLIP-ResNet50 (He et al., 2016), and CLIP-ResNet101 (He et al., 2016). Overall, CLIP-VIT-B\16 yields the best performance among these CLIP models. However, the diffusion classifier demonstrates more powerful effectiveness, particularly when the trial is set to 30 (Tr.30).

It is worth noting that the diffusion classifier is considerably slower than the CLIP classifier, primarily due to its more complex computational process. Theoretically, selecting a larger number of trials and time steps in diffusion could potentially lead to further enhancements in performance. Therefore, opting for the diffusion method entails a trade-off, wherein significant computational resources are required to achieve performance gains.

	Render I.	Depth I.	Edge I.	Avg.	Times (s)
CLIP-VIT-B\16	54.7	59.7	34.2	49.5	0.025
CLIP-VIT-B\32	49.6	52.7	38.0	46.8	0.028
CLIP-ResNet50	45.2	41.9	37.5	41.5	0.037
CLIP-ResNet101	48.4	54.3	40.8	47.8	0.042
Diffusion-Tr.10	66.2	53.5	70.4	63.4	8.771
Diffusion-Tr.20	71.6	58.6	74.3	68.2	17.461
Diffusion-Tr.30	73.5	59.8	77.6	70.3	26.078

Table 6: Comparison to the CLIP. Times denotes the averaged inference time of a single projected image.

5.3.3. Ablation on Projection Styles

In the ablation studies of the three style images (Render, Depth, Edge) for both the CLIP and Diffusion models on the ModelNet10 dataset, the results are reported in Tables 6 and 7. The CLIP model exhibits the best performance for depth images and the worst for edge images, whereas the diffusion model performs inversely. Additionally, we explore the ensemble of these styles to achieve better performance. Specifically, as introduced in Section 4.2, the ensemble involves taking the average over projection styles when calculating the matching score. From Table 7, the CLIP model achieves better mAcc results by combining render and depth images. Conversely, the

Diffusion model demonstrates the best performance by combining render and edge images.

	Render	Depth	Edge	mACC
CLIP-VIT-B\16	✓	✓		61.2
CLIP-VIT-B\16		✓	✓	53.6
CLIP-VIT-B\16	✓		✓	50.2
CLIP-VIT-B\16	✓	✓	✓	53.6
Diffusion-Tr.30	✓	✓		73.2
Diffusion-Tr.30		✓	✓	79.4
Diffusion-Tr.30	✓		✓	81.7
Diffusion-Tr.30	✓	✓	✓	79.4

Table 7: Ablation on Projection Styles

6. Challenges

The open-pose setting poses greater difficulty for 3D zero-shot classification and holds more practical significance. In this context, we make the first attempt to introduce the 3D open-pose zero-shot classification task and provide one effective method as the baseline for subsequent studies. Clearly, this task presents numerous challenges that warrant further research and investigation. In this section, we focus on exploring several key challenges and potential solutions.

Insufficient samples for 3D seen classes. The scarcity of samples for 3D seen classes presents a significant obstacle to training accurate classification models. Unlike 2D data, the scarcity of 3D data arises from higher acquisition costs, increased processing and storage requirements, limited access channels, and difficulties in annotation. To overcome this challenge,

researchers may explore data augmentation techniques such as scaling and translation to generate synthetic samples and enrich the training dataset. Additionally, leveraging text-to-3D generation models (Lin et al., 2023; Chen et al., 2023b) or image-to-3D generation models (Liu et al., 2023) to expand 3D data holds promising prospects.

Viewpoint similarity among the classes. The presence of viewpoint similarity among different classes complicates the 3D zero-shot classification process, particularly for projection-based methods (Huang et al., 2023; Zhu et al., 2023; Xue et al., 2023a; Wang et al., 2023b). Instances like tables and beds can appear remarkably similar from certain viewpoints, a common occurrence in the open-pose setting. One possible solution is to incorporate viewpoint augmentation during training, exposing the model to diverse viewpoints of the same object class to improve its robustness against viewpoint variations. For methods that do not require additional training, iterating multiple times may help determine the optimal viewpoint.

Attribute relationship between seen and unseen classes. Establishing an effective relationship between seen and unseen classes is crucial for generalizing the classification model to unseen classes (Zhou et al., 2023; Wang et al., 2023a). Transfer learning techniques, such as feature alignment and domain adaptation, can be explored to leverage knowledge from seen classes and transfer it to unseen classes. Furthermore, fundamental knowledge from disciplines such as physics and biology could be utilized to generalize the attributes of seen and unseen classes.

Bias in the distribution of pose data. In training data for large models, there is typically a lower proportion of open-pose data compared to

aligned-pose data (Radford et al., 2021; Rombach et al., 2022; Sohl-Dickstein et al., 2015). The imbalance in the distribution of open-pose data in training datasets hinders the model’s ability to learn representative features for open-pose classification. Addressing this challenge may involve collecting additional open-pose data or exploring techniques to balance the distribution of open-pose data in the training process.

Our open-pose setting introduces additional complexities, such as variations in object orientations and viewpoints, which are not adequately addressed by existing 3D zero-shot classification models. By incorporating these considerations into the design and training of 3D large-scale models, there will be an enhancement in their adaptability and robustness for handling diverse 3D data in reality. Our work serves as an important clue, highlighting the critical need for open-pose scenarios to promote the development of 3D zero-shot learning.

7. Conclusion

This paper provides an overview of the current progress in 3D zero-shot classification and proposes a more challenging benchmark for 3D zero-shot classification, aiming to recognize unseen 3D objects with open poses. Correspondingly, we validate the effectiveness of different strategies and design a concise pipeline with a concise angle refinement mechanism to present the preliminary solution. However, due to the significantly higher difficulty, our approach, being the first exploration in open-pose situations, does not achieve as remarkable results as in the aligned-pose case. Furthermore, we pioneer the exploration of knowledge transfer using pre-training Diffusion, broaden-

ing the scope of validation beyond existing CLIP-based methods. Finally, we also set out challenges and potential exploration strategies for 3D zero-shot classification in the open-pose setting.

References

- Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D., 2020. Language models are few-shot learners, in: NeurIPS, pp. 1877–1901.
- Canny, J., 1986. A computational approach to edge detection. *IEEE TPAMI*, 679–698.
- Chen, J., Yang, M., Velipasalar, S., 2023a. Viewnet: A novel projection-based backbone with view pooling for few-shot point cloud classification, in: CVPR, pp. 17652–17660.
- Chen, R., Chen, Y., Jiao, N., Jia, K., 2023b. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation, in: ICCV, pp. 22246–22256.
- Cheraghian, A., Rahman, S., Campbell, D., Petersson, L., 2019a. Mitigating the hubness problem for zero-shot learning of 3d objects, in: BMVC, pp. 41–53.

- Cheraghian, A., Rahman, S., Campbell, D., Petersson, L., 2020. Transductive zero-shot learning for 3d point cloud classification, in: WACV, pp. 923–933.
- Cheraghian, A., Rahman, S., Chowdhury, T.F., Campbell, D., Petersson, L., 2022. Zero-shot learning on 3d point cloud objects and beyond. IJCV 130, 2364–2384.
- Cheraghian, A., Rahman, S., Petersson, L., 2019b. Zero-shot learning of 3d point cloud objects, in: MVA, pp. 1–6.
- Choy, C., Gwak, J., Savarese, S., 2019. 4d spatio-temporal convnets: Minkowski convolutional neural networks, in: CVPR, pp. 3075–3084.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2021. An image is worth 16x16 words: Transformers for image recognition at scale, in: ICLR.
- Guo, Y., Wang, H., Hu, Q., Liu, H., Liu, L., Bennamoun, M., 2020. Deep learning for 3d point clouds: A survey. IEEE TPAMI 43, 4338–4364.
- Han, Z., Fu, Z., Chen, S., Yang, J., 2022. Semantic contrastive embedding for generalized zero-shot learning. IJCV 130, 2606–2622.
- Hao, Y., Su, Y., Lin, G., Su, H., Wu, Q., 2023. Contrastive generative network with recursive-loop for 3d point cloud generalized zero-shot classification. PR 144, 109843.

- He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: CVPR, pp. 770–778.
- Hegde, D., Valanarasu, J.M.J., Patel, V., 2023. Clip goes 3d: Leveraging prompt tuning for language grounded 3d recognition, in: ICCV, pp. 2028–2038.
- Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. NeurIPS 33, 6840–6851.
- Huang, T., Dong, B., Yang, Y., Huang, X., Lau, R.W., Ouyang, W., Zuo, W., 2023. Clip2point: Transfer clip to point cloud classification with image-depth pre-training, in: ICCV, pp. 22157–22167.
- Jiang, C., Huang, K., Wu, J., Wang, X., Xiao, J., Hussain, A., 2023. Pointgs: Bridging and fusing geometric and semantic space for 3d point cloud analysis. IF 91, 316–326.
- Kanezaki, A., Matsushita, Y., Nishida, Y., 2018. Rotationnet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints, in: CVPR, pp. 5010–5019.
- Larochelle, H., Erhan, D., Bengio, Y., 2008. Zero-data learning of new tasks, in: AAAI, pp. 646–651.
- Li, A.C., Prabhudesai, M., Duggal, S., Brown, E., Pathak, D., 2023. Your diffusion model is secretly a zero-shot classifier, in: ICCV, pp. 2206–2217.
- Lian, Z., Zhang, J., Choi, S., ElNaghy, H., El-Sana, J., Furuya, T., Giachetti, A., Guler, R.A., Lai, L., Li, C., Li, H., Limberger, F.A., Martin, R., Nakan-

- ishi, R.U., Neto, A.P., Nonato, L.G., Ohbuchi, R., Pevzner, K., Pickup, D., Rosin, P., Sharf, A., Sun, L., Sun, X., Tari, S., Unal, G., Wilson, R.C., 2015. Non-rigid 3D Shape Retrieval, in: Eurographics Workshop on 3D Object Retrieval.
- Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y., 2023. Magic3d: High-resolution text-to-3d content creation, in: CVPR, pp. 300–309.
- Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C., 2023. Zero-1-to-3: Zero-shot one image to 3d object, in: ICCV, pp. 9298–9309.
- Maturana, D., Scherer, S., 2015. Voxnet: A 3d convolutional neural network for real-time object recognition, in: IROS, pp. 922–928.
- Michele, B., Boulch, A., Puy, G., Bucher, M., Marlet, R., 2021. Generative zero-shot learning for semantic segmentation of 3d point clouds, in: 3DV, pp. 992–1002.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. NeurIPS 26.
- Mo, K., Zhu, S., Chang, A.X., Yi, L., Tripathi, S., Guibas, L.J., Su, H., 2019. PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding, in: CVPR, pp. 909–918.
- Naeem, M.F., Khan, M.G.Z.A., Xian, Y., Afzal, M.Z., Stricker, D., Van Gool,

- L., Tombari, F., 2023. I2mvformer: Large language model generated multi-view document supervision for zero-shot image classification, in: CVPR, pp. 15169–15179.
- Naeem, M.F., Örnek, E.P., Xian, Y., Van Gool, L., Tombari, F., 2022. 3d compositional zero-shot learning with decompositional consensus, in: ECCV, pp. 713–730.
- Narayan, S., Gupta, A., Khan, F.S., Snoek, C.G., Shao, L., 2020. Latent embedding feedback and discriminative features for zero-shot classification, in: ECCV, pp. 479–495.
- Ning, X., Yu, Z., Li, L., Li, W., Tiwari, P., 2024. Dilf: Differentiable rendering-based multi-view image–language fusion for zero-shot 3d shape understanding. IF 102, 102033.
- Pennington, J., Socher, R., Manning, C.D., 2014. Glove: Global vectors for word representation, in: EMNLP, pp. 1532–1543.
- Qi, C.R., Su, H., Mo, K., Guibas, L.J., 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation, in: CVPR, pp. 652–660.
- Qi, Z., Dong, R., Fan, G., Ge, Z., Zhang, X., Ma, K., Yi, L., 2023. Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining, in: ICML, pp. 28223–28243.
- 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: ICML, pp. 8748–8763.

- Rahman, S., Khan, S.H., Porikli, F., 2020. Zero-shot object detection: joint recognition and localization of novel concepts. *IJCV* 128, 2979–2999.
- Ravi, N., Reizenstein, J., Novotny, D., Gordon, T., Lo, W.Y., Johnson, J., Gkioxari, G., 2020. Accelerating 3d deep learning with pytorch3d. [arXiv:2007.08501](https://arxiv.org/abs/2007.08501) .
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-resolution image synthesis with latent diffusion models, in: *CVPR*, pp. 10684–10695.
- Shen, S., Zhu, Z., Fan, L., Zhang, H., Wu, X., 2024. Diffclip: Leveraging stable diffusion for language grounded 3d classification, in: *WACV*, pp. 3596–3605.
- Siddiqi, K., Zhang, J., Macrini, D., Shokoufandeh, A., Bouix, S., Dickinson, S., 2008. Retrieving articulated 3-d models using medial surfaces. *MVA* 19, 261–275.
- Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S., 2015. Deep unsupervised learning using nonequilibrium thermodynamics, in: *ICML*, pp. 2256–2265.
- Su, H., Maji, S., Kalogerakis, E., Learned-Miller, E., 2015. Multi-view convolutional neural networks for 3d shape recognition, in: *ICCV*, pp. 945–953.
- Uy, M.A., Pham, Q.H., Hua, B.S., Nguyen, D.T., Yeung, S.K., 2019. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data, in: *ICCV*, pp. 1588–1597.

- Wang, Q., Liu, L., Jing, C., Chen, H., Liang, G., Wang, P., Shen, C., 2023a. Learning conditional attributes for compositional zero-shot learning, in: CVPR, pp. 11197–11206.
- Wang, W., Zheng, V.W., Yu, H., Miao, C., 2019. A survey of zero-shot learning: Settings, methods, and applications. TIST 10, 1–37.
- Wang, Y., Huang, S., Gao, Y., Wang, Z., Wang, R., Sheng, K., Zhang, B., Liu, S., 2023b. Transferring clip’s knowledge into zero-shot point cloud semantic segmentation, in: ACM MM, pp. 3745–3754.
- Wei, X., Gong, Y., Wang, F., Sun, X., Sun, J., 2021. Learning canonical view representation for 3d shape recognition with arbitrary views, in: ICCV, pp. 407–416.
- Wu, X., Lao, Y., Jiang, L., Liu, X., Zhao, H., 2022. Point transformer v2: Grouped vector attention and partition-based pooling. NeurIPS 35, 33330–33342.
- Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J., 2015. 3d shapenets: A deep representation for volumetric shapes, in: CVPR, pp. 1912–1920.
- Xue, L., Gao, M., Xing, C., Martín-Martín, R., Wu, J., Xiong, C., Xu, R., Niebles, J.C., Savarese, S., 2023a. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding, in: CVPR, pp. 1179–1189.
- Xue, L., Yu, N., Zhang, S., Li, J., Martín-Martín, R., Wu, J., Xiong, C., Xu,

- R., Niebles, J.C., Savarese, S., 2023b. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. arXiv preprint arXiv:2305.08275 .
- Yang, F., Lee, Y., Lin, C., Wang, Y.F., 2023. Semantics-guided intra-category knowledge transfer for generalized zero-shot learning. IJCV 131, 1331–1345.
- Yang, G., Ye, Z., Zhang, R., Huang, K., 2022. A comprehensive survey of zero-shot image classification: methods, implementation, and fair evaluation. AIMS-ACI 2, 1–31.
- Yang, Z., Wang, L., 2019. Learning relationships for multi-view 3d object recognition, in: ICCV, pp. 7505–7514.
- Ye, Z., Hu, F., Lyu, F., Li, L., Huang, K., 2021. Disentangling semantic-to-visual confusion for zero-shot learning. IEEE TMM 24, 2828–2840.
- Ye, Z., Yang, G., Jin, X., Liu, Y., Huang, K., 2023. Rebalanced zero-shot learning. IEEE TIP .
- Zhang, L., Rao, A., Agrawala, M., 2023. Adding conditional control to text-to-image diffusion models, in: ICCV, pp. 3836–3847.
- Zhang, R., Guo, Z., Zhang, W., Li, K., Miao, X., Cui, B., Qiao, Y., Gao, P., Li, H., 2022. Pointclip: Point cloud understanding by CLIP, in: CVPR, pp. 8542–8552.
- Zhao, H., Jiang, L., Jia, J., Torr, P.H., Koltun, V., 2021. Point transformer, in: CVPR, pp. 16259–16268.

- Zhao, W., Yan, Y., Yang, C., Ye, J., Yang, X., Huang, K., 2023. Divide and conquer: 3d point cloud instance segmentation with point-wise binarization, in: ICCV, pp. 562–571.
- Zhou, L., Liu, Y., Bai, X., Li, N., Yu, X., Zhou, J., Hancock, E.R., 2023. Attribute subspaces for zero-shot learning. PR 144, 109869.
- Zhou, L., Liu, Y., Zhang, P., Bai, X., Gu, L., Zhou, J., Yao, Y., Harada, T., Zheng, J., Hancock, E., 2022. Information bottleneck and selective noise supervision for zero-shot learning. ML , 1–23.
- Zhou, W., Zheng, F., Zhao, Y., Pang, Y., Yi, J., 2024. Msdcnn: A multiscale dilated convolution neural network for fine-grained 3d shape classification. NN 172, 106141.
- Zhu, X., Zhang, R., He, B., Zeng, Z., Zhang, S., Gao, P., 2023. Pointclip v2: Adapting clip for powerful 3d open-world learning, in: ICCV, pp. 2639–2650.