No Captions, No Problem: Captionless 3D-CLIP Alignment with Hard Negatives via CLIP Knowledge and LLMs

Cristian Sbrolli

www.deib.polimi.it/eng/people/details/1273522

Matteo Matteucci

www.deib.polimi.it/eng/people/details/267262

Department of Electronics, Information and Bioengineering
Politecnico di Milano
Milano, IT

Abstract

In this study, we explore an alternative approach to enhance contrastive text-image-3D alignment in the absence of textual descriptions for 3D objects. We introduce two unsupervised methods, *I2I* and $(I2L)^2$, which leverage CLIP knowledge about textual and 2D data to compute the neural perceived similarity between two 3D samples. We employ the proposed methods to mine 3D hard negatives, establishing a multimodal contrastive pipeline with hard negative weighting via a custom loss function. We train on different configurations of the proposed hard negative mining approach, and we evaluate the accuracy of our models in 3D classification and on the cross-modal retrieval benchmark, testing image-to-shape and shape-to-image retrieval. Results demonstrate that our approach, even without explicit text alignment, achieves comparable or superior performance on zero-shot and standard 3D classification, while significantly improving both image-to-shape and shape-to-image retrieval compared to previous methods.

Our code is available at: https://github.com/mortorit/No-Captions-No-Problem.

1 Introduction

In recent years, multimodal feature alignment gathered increasing interest in the scientific community. This was primarily sparked by the introduction of CLIP (Contrastive Language-Image Pretraining) [16], which successfully aligns images and text through contrastive multimodal learning, obtaining impressive results in many downstream tasks.

Extending CLIP to 3D data is however a challenging task, primarily due to the scarcity of well-annotated text-2D-3D triplets required. Indeed, although image-3D datasets are relatively easy to obtain, for instance through rendering, there is a lack of data with well-curated and structurally meaningful descriptions for 3D models.

Initial efforts to address this issue aligned images and 3D objects, either without text or with template text prompts filled with category names or metadata-derived descriptions. More recent works have leveraged Large Multimodal Models (LMMs) capable of processing images to generate image captions, thereby generating text descriptions for the 3D object. However, these generated texts often lack descriptiveness and discriminativeness, and suffer
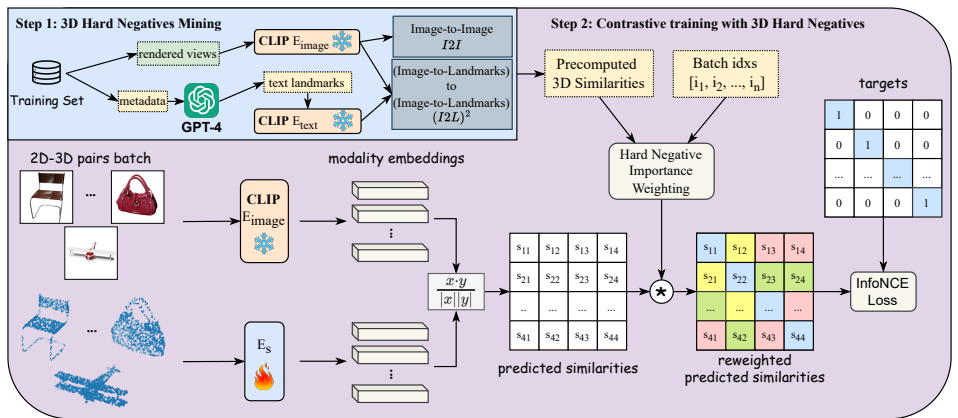
Figure 1: Overview of our proposed approach. We first precompute 3D samples similarities through our proposed neural similarity metrics, then we use the obtained scores to enhance the contrastive training with hard negatives.

from inherent LMM biases like counting inability and a propensity to fill in missing information. They also encounter problems associated with image-based queries, including the incorporation of information not obtainable from the point cloud model, such as colours, textures, and materials in the 3D model.

In light of the limitations associated with LMM-generated prompts, this study explores an alternative approach to enhance contrastive text-2D-3D alignment in the absence of labelled textual descriptions for 3D objects. Specifically, we propose to leverage CLIP knowledge about textual and 2D data to compute similarities between point cloud models, then exploit those similarities to mine hard negative samples, enhancing the contrastive training. We introduce two unsupervised approaches for computing the neural-perceived similarity between 3D samples, *I2I* and $(I2L)^2$. The first computes 3D similarities by only leveraging object views and the CLIP image encoder. The second method we propose leverages instead both images and LLM-generated texts, while addressing both the limitations of image-based approaches and the low-descriptiveness of image-captioning generated prompts. To this end, we transpose the 3D shape similarity problem into an image-text similarity space, where each object view is characterized by its similarity vector in relation to a set of fixed and detail-rich LMM-generated category prompts that do not refer to color, texture, or material.

We establish a text-image-3D contrastive pipeline that uses precomputed 3D similarities to perform hard negative sampling, and we instantiate it in different configurations by employing the proposed similarities. We rigorously test the obtained models following evaluation benchmarks and community protocols for text-image-3D contrastive learning on multiple datasets, both real and synthetic. We assess the macro-structure of the latent space by testing on 3D classification (both zero-shot and fine-tuned). We evaluate instead the fine-grained understanding of intra-category details and the alignment of 3D and images by a cross-modal retrieval benchmark, testing image-to-shape and shape-to-image retrieval. Our experiments demonstrate that our approach achieves superior or comparable performance on 3D classification, while obtaining a significant improvement in both image-to-shape and shape-to-image retrieval compared to state-of-the-art methods.

2 Related Works

Over the past decade, Contrastive Representation Learning (CRL) [7] has emerged as a robust self-supervised learning framework. It builds a well-structured representation space by learning to discern similar and dissimilar samples. The advent of InfoNCE loss [13, 21] and its application in CLIP (Contrastive Language Image Pretraining) [16] has further propelled CRL's popularity. CLIP, aligning images and text in the latent space of two separate modality encoders, has shown remarkable results in various downstream tasks, making it a key component in models like LLaVA [11] and DALLE-2 [17]. Subsequent research has enhanced the CLIP methodology with soft labels [4], sigmoid loss [22], data filtering, and hard negatives [15], leading to improved text-image alignment.

Building on the success of CLIP with text-image data, researchers have expanded this approach to other modalities, including audio [6], video [24], and even imu, depth, and heat maps [5]. The extension to 3D arrived with initial works obtaining the alignment of voxels and images [19], followed by more efficient 3D representations. Early works aligning point clouds with other modalities, such as PointCLIP [28] and its successor, PointCLIP v2 [29], use the pretrained vision encoder from CLIP to process point clouds by projecting them onto multi-view depth maps. CLIP2Point [8] also trains on depth maps projected from point clouds, but introduces a fully trainable depth encoder and an intra-modality loss for aligning different depth maps of the same 3D sample, thereby enhancing invariance to view changes. ULIP [25] presents and successfully trains a pure text-image-3D point cloud model, introducing a point cloud encoder instead of using projected depth maps, and outperforms previous works in all benchmarks. After ULIP, as with CLIP, research in this field has begun to explore ways to improve the model's ability to align concepts from 3D with other modalities. ReCON [14] achieves this by combining a generative paradigm along with the contrastive loss to ensure the latent space retains more information about the 3D shape. While effective, this approach is less parameter-efficient, as it employs an additional generative branch and module. ULIP 2 [26] maintains efficiency by focusing instead on generating more informative text captions. Indeed, while ULIP used metadata and class categories to create text labels for image-text-3D triplets, ULIP 2 employs BLIP-2 [10] to pre-generate per-sample text captions for 2D views of the training set objects. This allows for a more informative contrastive training between text and shape embeddings. However, the use of a general-purpose LMM means that the captions inherit both the strengths and limitations of the model. For instance, by analyzing the generated texts (which are publicly available), it is evident that the model lacks the specialization required to describe relevant structural elements of the considered objects. Furthermore, the captions often describe the object in a generic manner, sometimes hallucinating scenes biased toward the object's function or background color. Another challenge is the reference to color, textures, and materials in the captions, which are not visible from the point cloud. We further discuss and exemplify these aspects of the ULIP-2 captions in the supplementary.

In light of the aforementioned issues with previous models, the fundamental problem for this work is to remain parameter-efficient by not adding more models/modules, while still incorporating additional knowledge to achieve a more informative 3D contrastive training.

3   Method

3.1   Preliminaries: Contrastive Training and Hard Negatives

The objective of contrastive multimodal training, as proposed in CLIP [16], is to accurately match the same sample from different modalities among all the samples in a batch. Formally, the model takes a batch of N coupled samples (x_{m_1}, x_{m_2}) from different modalities m_1 and m_2 e.g. couples of images and their corresponding captions. Two separate modality encoders E_{m_1} and E_{m_2} process the batch producing L2-normalized modality embeddings for the batch, e_{m_1} and e_{m_2}, each with shape (N, F) with F being the number of latent features. The objective of the model is that of matching, through cosine similarity, corresponding samples from different modalities, thus aligning the corresponding embeddings in the latent space. The contrastive self-supervised loss is then defined as:

$$\mathcal{L}_{(m_1, m_2)} = \frac{1}{2} \mathcal{L}_{m_1 \to m_2} + \frac{1}{2} \mathcal{L}_{m_2 \to m_1}, \quad \mathcal{L}_{m_i \to m_j} = -\frac{1}{N} \sum_{k=1}^{N} \log \frac{\exp\left(e_{m_i}^k \cdot e_{m_j}^k / \tau\right)}{\sum_{s=1}^{N} \exp\left(e_{m_i}^k \cdot e_{m_j}^s / \tau\right)}, \quad (1)$$

with $\cdot$ being the inner product, τ a learnable temperature parameter stored as a log-parameterized scalar, and the apices indicize elements in the batch.

If all samples in the batch are vastly different, it becomes easier for the model to match samples, resulting in the model learning less detailed features. Conversely, if the batch contains similar samples, the model faces a challenge in discriminating the correct one, thereby forcing it to learn more detailed features. These samples are named as hard negatives, and their mining and sampling is a well-studied topic in classic representation learning [12, 20]. More recently, it has been applied in the context of contrastive learning, proving successful in aligning 2D images [18] and text-2D [15]. Yet, it remains unexplored in text-2D-3D contrastive representation learning, where the first step is to address hard negative mining, a non-trivial task with 3D data. In the following, we approach this problem by defining neural similarities to extract hard samples for a given 3D object in a dataset.

3.2   Similarities for 3D Hard Negative Mining: *I2I*, $(I2L)^2$

In preliminary experiments, reported in the supplementary, we tested with established 3D similarity metrics like Chamfer Distance (CD) and Earth Mover's Distance (EMD). However, these metrics did not yield satisfactory outcomes. While they excel at overall correspondence assessment, they tend to overlook finer details, favoring denser regions [3]. Moreover, they are not context-aware, thus they are not able to extract and compare context-relevant features which are paramount for the extraction of semantically challenging negatives. These limitations mainly impact the matching of local geometric structures, such as small or thin elements, holes and curvatures, all of which are crucial features in describing and distinguishing 3D samples. We propose then to build neural similarities capable of capturing relevant context-aware features by exploiting the semantically rich CLIP latent space. We introduce two similarity measures, depicted in Figure 2, built upon the publicly available OpenCLIP [2] model in its ViT-B-32 version.

I2I - Image to Image 3D similarity: Our first proposed strategy primarily utilizes images, based on the assumption that 2D views provide sufficient information for the 3D object. For each 3D sample x_{3D}^i, we consider its set of R rendered views $\{v^{i,r}\}_{r=0}^{R}$ from a set of fixed
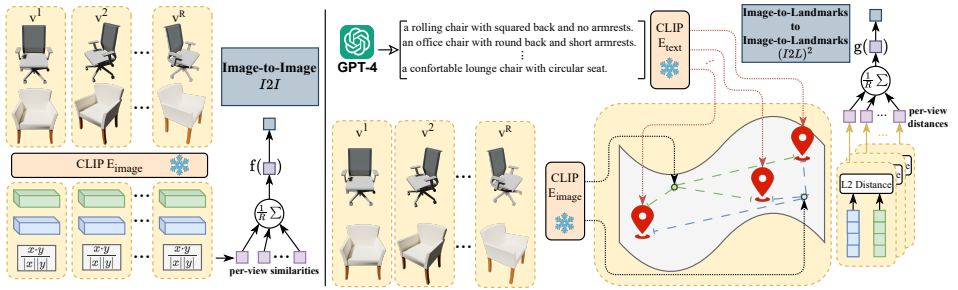
Figure 2: Our proposed similarity metrics for 3D hard negative mining.

camera poses for all samples. We employ the pretrained OpenCLIP image encoder E_i to compute the image embedding $e_{image}^{k,r}$ for each view. With a sufficient number of views, these embeddings serve as neural descriptors of the 3D object, capturing features from all sides. Given these descriptors for any pair of point cloud samples (x_{3D}^1, x_{3D}^2), we compute the *I2I* 3D similarity as:

$$\text{3D-Sim}_{\text{I2I}}(x_{3D}^1, x_{3D}^2) = f\left(\frac{1}{V}\sum_{r=1}^{V} e_{image}^{1,r} \cdot e_{image}^{2,r}\right), \quad f(x) = \frac{(x+1)}{2} \qquad (2)$$

which, considering L2-normalized embeddings, is the mean cosine similarity between corresponding views of the two objects, shifted to $[0,1]$ with $f(x)$. As demonstrated in Figure 3, this method effectively matches similar items by leveraging the fact that while CLIP is trained to align images and text, it also implicitly learns to align images containing similar concepts and features. However, in Figure 3 we also show that this approach still encounters the same color-texture-material issue discussed in Section 2, where objects with similar texture result more similar than structurally similar samples.

$(\textbf{I2L})^2$ - (Image to Landmarks) to (Image to Landmarks) 3D similarity: We propose a second strategy with two aims: (1) solving the colour-texture-material limitation and (2) leveraging both the image and text knowledge of CLIP derived from its large scale training. First, we need to define how to efficiently use text to describe 3D objects, so that we can leverage CLIP text encoder to our advantage. However, as mentioned in Section 2, using LMMs for captioning suffers from limitations in understanding structural features, resulting in not enough-detailed annotations or even the hallucination of false information, adversely affecting the accurate multimodal alignment. We propose instead to focus on fewer but useful and detail-rich texts. To this end, we use LMMs not to generate captions for each view, but to generate a small amount of detailed texts that describe hypothetical objects from a given category, covering its subsets and different instances for structural features. In particular, we leverage GPT-4 [1] to generate a set of L texts per category c with following prompt:

In the context of a 3D objects dataset, generate **L** text descriptions for the category "**c**" **<optional if metadata present:** containing "**METADATA**">. Describe one object per text, focusing on its relevant structural features and details. The list should cover a high variety of settings and types for each feature. Do not mention texture, materials, or color.
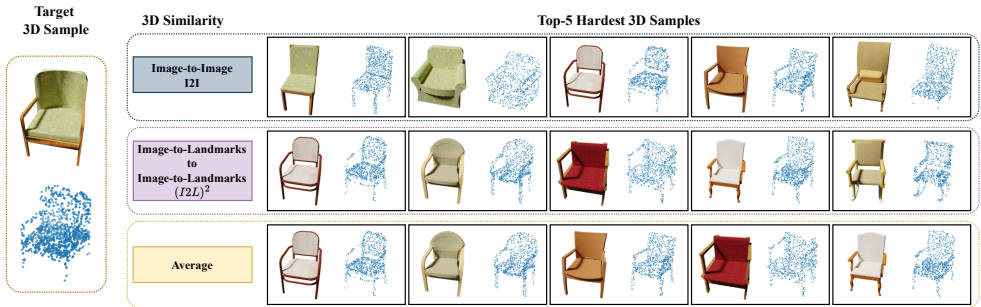
Figure 3: 3D-to-3D retrieval using our similarities on a chair sample from ShapeNet dataset.

Where METADATA, if available, contains possible subsets or information on objects in the given category. With the above prompt, we generate L texts per category in our training set and compute their text embeddings $\{e_{text}^{c,l}\}_{l=0}^{L}$ through OpenCLIP text encoder E_t. We refer to the generated texts and their embedding, respectively, as text landmarks and landmarks, as they will act as such in building 3D descriptors. Specifically, given a 3D object x_{3D}^i from category c and its views $\{v^{i,r}\}_{r=0}^{R}$, we build descriptors $d^{i,r}$ with dimension L for each view r by measuring its cosine similarity with respect to all of the landmarks for that category:

$$d^{i,r}[l] = e_{image}^{i,r} \cdot e_{text}^{c,l}, l \in 1...L \tag{3}$$

Each element of the descriptor is then a value between -1 and 1 measuring the similarity of the view with respect to one of the landmarks. When $L < F$, this process builds descriptors that live in a sparser and lower-dimensional per-category space, using the generated texts as landmarks (and thus the name) in the latent space of CLIP. Given two 3D objects (x_{3D}^1, x_{3D}^2) and their descriptors (d^1, d^2), each with shape (R, L), we can now compute $(I2L)^2$ shape similarity leveraging the mean Euclidean distance between descriptors of corresponding views:

$$\text{3D-Sim}_{(I2L)^2}(x_{3D}^1, x_{3D}^2) = g\left(\frac{1}{V}\sum_{r=1}^{V}\sqrt{\sum_{l=1}^{L}(d^{1,r}[l] - d^{2,r}[l])^2}\right), \quad g(x) = \frac{1}{1+x} \tag{4}$$

where g(x) maps distances to similarities values between 0 and 1. Despite the increased complexity in its construction, this descriptor offers several advantages. It enables the utilization of CLIP language knowledge alongside visual knowledge, and it does this efficiently by working with fewer, yet more detailed and meaningful texts. This approach also shifts the problem to a lower-dimensional representation space controlled by text landmarks, mitigating biases related to color, texture, and material. As we show in Figure 3, the $(I2L)^2$ similarity succeeds when the image-based similarity falls short, correctly preferring structurally similar samples over samples with similar texture/color.

Lastly, in Figure 3 we show that averaging I2I and $(I2L)^2$ similarities is beneficial, allowing us to benefit from CLIP visual features of the I2I method while also mitigating image biases and enforcing structural coherence between similar samples thanks to $(I2L)^2$.

3.3   Contrastive training with Hard Negative Sampling

Inspired by, successful methods with 2D and text-2D data [7, 15], we apply an importance sampling approach to increase the weight of harder samples in a batch. We leverage the similarity functions defined in the previous section to define a reweighted contrastive loss to align 2D and 3D data. We define this loss as:

$$\mathcal{L}_{\text{(2D,3D)-HN}} = -\frac{1}{2N}\sum_{i=1}^{N}\log\frac{\exp\left(e^i_{image}\cdot e^i_{shape}/\tau\right)}{\sum\limits_{s=1}^{N}w^{i,s}\exp\left(e^i_{image}\cdot e^s_{shape}/\tau\right)} - \frac{1}{2N}\sum_{s=1}^{N}\log\frac{\exp\left(e^s_{image}\cdot e^s_{shape}/\tau\right)}{\sum\limits_{i=1}^{N}w^{s,i}\exp\left(e^i_{image}\cdot e^s_{shape}/\tau\right)}, \quad (5)$$

where weights w for a given batch are defined as:

$$w^{i,s} = \frac{(N-1)sim(x^i_{3D},x^s_{3D})}{\sum\limits_{k\neq i}sim(x^i_{3D},x^k_{3D})}, \quad w^{s,i} = \frac{(N-1)sim(x^i_{3D},x^s_{3D})}{\sum\limits_{k\neq s}sim(x^k_{3D},x^s_{3D})}, \quad (6)$$

and $sim(\cdot,\cdot)$ is any positive similarity function between 3D samples. For each sample in a batch, the weights consider the mean similarity of all its negative samples. Typically, for each sample, easy negatives significantly outnumber hard negatives within a batch, causing the mean similarity to align with easy negative values. Consequently, easy negatives receive weights close to 1, while harder negatives—characterized by notably higher similarity—are upweighted in the loss function.

As similarity $sim(\cdot,\cdot)$ in Equation (6) we can employ both 3D-Sim_{I2I} and 3D-$Sim_{(I2L)^2}$. To enhance training efficiency, we precompute and store similarity values. Notably, we defined the $(I2L)^2$ similarity exclusively for samples within the same category. This choice leverages the observation that hard negatives often belong to the same category, sharing both concept and appearance. This observation allows us to significantly reduce the computational cost and memory requirements for similarity precomputation from D^2 (where D is dataset size), to $\sum_{c\in C}|c|^2$ with $|c|$ representing the number of samples in category c. We then define $sim(x,y)$ as 3D-$Sim(x,y)$ when the samples are from the same category, while we set it to a constant α otherwise, as to not obtain zero weights in Equation (6).

4   Experiments

4.1   Model and training details

We employ the pretrained OpenCLIP model, which we freeze, alongside a trainable point cloud encoder. To avoid architectural biases, we adopt the more efficient PointNeXt configuration from the state-of-the-art ULIP-2 [26], thus exclusively validating the novel training method. We also train on the same ShapeNet [2] subset, comprising 55 categories of 3D objects, each represented by 30 equally spaced views and a point cloud sampled from the mesh surface. We normalize point clouds to the unit sphere and apply random translations, rotations, and noise augmentation. The training process utilizes the loss defined in Equation (2), with the AdamW optimizer and a cosine annealing schedule featuring linear warmup. The initial learning rate is set to 10^{-7} and gets to 10^{-3} after 30 warmup epochs.

We train models using the proposed 3D-Sim_{I2I}, 3D-$Sim_{(I2L)^2}$ for hard negatives. Additionally, observing the benefits of combining the two similarities, we also train a third model denoted as 3D-Sim_{AVG}. However, instead of directly averaging the similarities, we compute

Model	#P (M)	Zero-Shot 3D Classification				Standard 3D Classification	
		ModelNet40		ScanObjectNN		ModelNet40	ScanObjectNN
		top-1	top-5	top-1	top-5	top-1	top-1
PointCLIP	0	20.2	43.7	15.4	33.9	92.0	89.1
CLIP2Point	86	49.4	66.5	23.32	57.5	94.2	90.6
ULIP-PointNet++	1.5	58.4	78.2	49.9	78.8	93.5	89.7
ULIP-PointBERT	22.1	60.4	84.0	48.5	79.9	94.1	86.4
ReCon	43.6	61.7	83.6	30.5	63.2	**94.7**	**91.3**
ULIP 2-PointNeXt	1.4	64.5	81.3	-	-	-	89.5
ULIP 2-PointBERT	22.1	**66.4**	**87.7**	-	-	-	86.7
Ours-PointNeXt-*I2I*	1.4	64.2	<u>87.1</u>	53.1	82.4	93.2	90.1
Ours-PointNeXt-$(I2L)^2$	1.4	63.7	86.9	54.8	<u>83.9</u>	<u>94.3</u>	88.9
Ours-PointNeXt-AVG	1.4	<u>65.9</u>	87.0	**<u>55.7</u>**	82.6	93.7	<u>90.2</u>

Table 1: 3D classification results. Our best model is underlined, the best overall is in bold.

weights for both 3D-Sim_{I2I}, 3D-$Sim_{(I2L)^2}$ and average them, as this is more robust to differences in similarity scale between the two employed similarities. For all experiments, we set $\alpha = 0.25$ and $L = 128$, and we ablate the latter in Section 5.

4.2   Experimental results

We adopt the evaluation pipeline of previous works [25, 26, 28, 29] by testing our models on zero-shot 3D classification and standard 3D classification on both ModelNet40 [23] and the hardest set from ScanObjectNN [22] (PB_T50_RS). Moreover, to evaluate in a more fine-grained manner the learned multimodal space, we also introduce experiments on cross-modal retrieval. For ULIP-2, we report only available results as the checkpoints for ShapeNet-pretrained models have not been released.

Zero-Shot 3D Classification: Thanks to the multimodal alignment, zero-shot 3D classification can directly be performed by computing text embeddings for the sentences "a point cloud of a <c>" for each category c, then predicting the one with the maximum similarity with respect to the point cloud embedding of the considered sample. We report top-1 accuracy and top-5 accuracy results in Table 1. On ModelNet40, our models outperform most previous models despite the lower number of parameters, being on par or better with corresponding ULIP-2 PointNeXt architectures. Our AVG model surpasses ReCon, which has 30x more parameters and is also competitive with the best ULIP-2 PointBERT architecture, which has 15x more parameters. On ScanObjectNN, our model outperforms all baselines and, based on ULIP-2 ModelNet improvements with respect to ULIP, we expect it to be comparable to ULIP 2 models too. Note that our models do not train on directly aligning 3D and text data, and thus the performance of our model also shows that, as suggested in Imagebind [5], the alignment of all couples of modalities is not needed, as aligning two modalities suffices to indirectly align the others.

Standard 3D Classification: We adopt the community evaluation protocol on 3D classification by concatenating a linear head to the point cloud encoder and finetuning it on classification. Results, reported in Table 1 show that our models obtain competitive performances with respect to state-of-the art models and similar to analogous PointNeXt architectures. This shows that our models learn a meaningful space which serves as a good initialization for fine-tuning on other 3D datasets, obtaining competitive results even with a much lower number of parameters.

Cross-modal retrieval While in ULIP [25] authors visually experiment cross-model retrieval with few images from Caltech101 dataset, which for fairness we also test and report in Figure 4, a quantitative analysis is not presented. Unlike 3D classification, which focuses solely on category prediction, cross-modal retrieval assesses the model's ability to
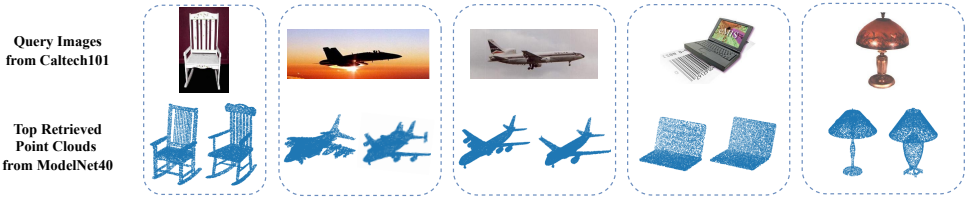
Figure 4: Example of 2D-to-3D cross modal retrieval using our model. Images are from Caltech101, point clouds are from ModelNet40 and are ranked in order of cosine similarity.

align concepts and fine-grained features across different modalities. To evaluate this aspect, we conduct zero-shot 2D-to-3D and 3D-to-2D retrieval experiments on the Pix3D dataset—a real-world 2D-3D dataset featuring common objects. Notably, Pix3D includes masks for background removal, and we evaluate both the simpler no-background and the more complex background versions. As for classification, we report top-1 and top-5 accuracy in Table 2. Our results demonstrate that our models consistently outperform previous approaches in both 2D-to-3D and 3D-to-2D retrieval tasks. The AVG model excels across most scenarios, except for image-to-shape retrieval with background, where our *I2I* model exhibits superior performance. Specifically, we achieve an average (top-1,top-5) improvement of $(5.2\%, 3.95\%)$ on test without backgrounds and $(4.65\%, 4.3\%)$ tests with backgrounds. These findings highlight the effectiveness of our approach, showing that the multimodal alignment benefits from our hard negative strategy also at a finer-grained level, leading to a better ability in discriminating correct shapes from images and vice versa.

		No Background				Background			
		Image-to-Shape		Shape-to-Image		Image-to-Shape		Shape-to-Image	
Model	#P (M)	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
PointCLIP	0	13.6	47.3	10.1	42.5	9.9	41.2	7.4	34.8
CLIP2Point	86	20.4	63.2	16.5	53.7	15.0	50.1	10.9	45.2
ULIP-PointNet++	1.5	23.9	64.6	17.2	55.9	15.1	49.2	11.4	47.3
ULIP-PointBERT	22.1	25.1	65.8	17.5	55.6	14.8	51.4	12.0	47.5
ReCon	43.6	31.4	72.8	22.8	63.1	20.7	56.4	19.6	53.2
Ours-PointNeXt-*I2I*	1.4	35.2	74.4	27.7	68.1	**26.3**	**60.4**	22.8	56.6
Ours-PointNeXt-$(I2L)^2$	1.4	33.3	74.4	26.8	66.8	25.3	59.9	22.2	55.5
Ours-PointNeXt-AVG	1.4	**35.6**	**75.1**	**29.0**	**68.7**	25.9	60.1	**23.3**	**57.8**

Table 2: Cross modal retrieval on Pix3D. The best model overall is highlighted in bold.

5 Ablation Study

We experiment with different numbers of landmarks per category, and we report top-1 accuracy results in Table 3, on ModelNet40 for zero-shot and finetuned 3D classification and on Pix3D with background for retrieval, averaged over 2D-to-3D and 3D-to-2D, while we report extended results in the supplementary. For this study, we use the $(I2L)^2$ model to test the effect of landmarks without the contribution of *I2I* similarity. As we increase the number of generated texts, their diversity decreases and they tend to repetition, becoming unuseful in discriminating differ-

L	Zero Shot	Fine Tuned	Retrieval
32	52.5	93.1	16.4
64	58.4	92.9	20.1
128	63.7	94.0	23.8
256	63.9	93.7	24.1
512	63.9	94.1	23.9

Table 3: Ablation on L value.

ent features and structural settings. Consequently, we opt for $L = 128$, as it allows a balance between efficiency and performance without compromising accuracy.

6 Conclusions

In this study, we presented an effective method to enhance the 3D alignment to the CLIP model by using CLIP itself, through its visual and textual knowledge. We propose two similarity metrics, $I2I$ and $(I2L)^2$, leveraging CLIP to compare 3D samples and extract 3D hard negatives, and use these to build an effective hard negative weighted contrastive pipeline. With $(I2L)^2$, we also address the challenges of 2D biases (texture, materials, colours) in comparing 3D shapes and we leverage LLM-generated texts in an innovative way, mitigating issues with LMM-generated captions from previous methods. We evaluate our models on zero-shot and finetuned 3D classification, obtaining results comparable or better than previous approaches. Lastly, we evaluate the fine-grained multimodal understanding of our model with cross-modal retrieval, where our models surpass all the previous approaches.

7 Acknowledgements

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] 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.

[3] Ben Fei, Weidong Yang, Wen-Ming Chen, Zhijun Li, Yikang Li, Tao Ma, Xing Hu, and Lipeng Ma. Comprehensive review of deep learning-based 3d point cloud completion processing and analysis. *IEEE Transactions on Intelligent Transportation Systems*, 23 (12):22862–22883, 2022. doi: 10.1109/TITS.2022.3195555.

[4] Yuting Gao, Jinfeng Liu, Zihan Xu, Tong Wu, Enwei Zhang, Ke Li, Jie Yang, Wei Liu, and Xing Sun. Softclip: Softer cross-modal alignment makes clip stronger. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1860–1868, 2024.

[5] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023.

[6] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE, 2022.

[7] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742, 2006. doi: 10.1109/CVPR.2006. 100.

[8] Tianyu Huang, Bowen Dong, Yunhan Yang, Xiaoshui Huang, Rynson W.H. Lau, Wanli Ouyang, and Wangmeng Zuo. Clip2point: Transfer clip to point cloud classification with image-depth pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22157–22167, October 2023.

[9] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL https://doi.org/10.5281/zenodo.5143773. If you use this software, please cite it as below.

[10] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[11] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[12] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4004–4012, 2016.

[13] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[14] 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 *International Conference on Machine Learning (ICML)*, 2023.

[15] Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenhende, Yash Patel, Yi Wen, Vignesh Ramanathan, and Dhruv Mahajan. Filtering, distillation, and hard negatives for vision-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6967–6977, June 2023.

[16] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[17] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[18] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *International Conference on Learning Representations*, 2021.

[19] Cristian Sbrolli, Paolo Cudrano, Matteo Frosi, and Matteo Matteucci. Ic3d: Image-conditioned 3d diffusion for shape generation. *arXiv preprint arXiv:2211.10865*, 2022.

[20] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015. doi: 10.1109/CVPR.2015. 7298682.

[21] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 1857–1865, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

[22] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Duc Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *International Conference on Computer Vision (ICCV)*, 2019.

[23] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015.

[24] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021.

[25] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1179–1189, June 2023.

[26] 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.

[27] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023.

[28] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8552–8562, 2022.

[29] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, and Peng Gao. Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2639–2650, 2023.