Single Image 3D Shape Retrieval via Cross-Modal Instance and Category Contrastive Learning

Ming-Xian Lin[1,3] **Jie Yang**[1,3] **He Wang**[4] **Yu-Kun Lai**[5]
Rongfei Jia[6] **Binqiang Zhao**[6] **Lin Gao**[1,2,3*]
[1] Beijing Key Laboratory of Mobile Computing and Pervasive Device,
Institute of Computing Technology, Chinese Academy of Sciences
[2] Zhejiang Lab [3] University of Chinese Academy of Sciences
[4] Center on Frontiers of Computing Studies, Peking University
[5] School of Computer Science and Informatics, Cardiff University
[6] Tao Technology Department, Alibaba Group
{linmingxian20g, yangjie01, gaolin}@ict.ac.cn hewang@pku.edu.cn
LaiY4@cardiff.ac.uk {rongfei.jrf, binqiang.zhao}@alibaba-inc.com

Abstract

In this work, we tackle the problem of single image-based 3D shape retrieval (IBSR), where we seek to find the most matched shape of a given single 2D image from a shape repository. Most of the existing works learn to embed 2D images and 3D shapes into a common feature space and perform metric learning using a triplet loss. Inspired by the great success in recent contrastive learning works on self-supervised representation learning, we propose a novel IBSR pipeline leveraging contrastive learning. We note that adopting such cross-modal contrastive learning between 2D images and 3D shapes into IBSR tasks is non-trivial and challenging: contrastive learning requires very strong data augmentation in constructed positive pairs to learn the feature invariance, whereas traditional metric learning works do not have this requirement. Moreover, object shape and appearance are entangled in 2D query images, thus making the learning task more difficult than contrasting single-modal data. To mitigate the challenges, we propose to use multi-view grayscale rendered images from the 3D shapes as a shape representation. We then introduce a strong data augmentation technique based on color transfer, which can significantly but naturally change the appearance of the query image, effectively satisfying the need for contrastive learning. Finally, we propose to incorporate a novel category-level contrastive loss that helps distinguish similar objects from different categories, in addition to classic instance-level contrastive loss. Our experiments demonstrate that our approach achieves the best performance on all the three popular IBSR benchmarks, including Pix3D, Stanford Cars, and Comp Cars, outperforming the previous state-of-the-art from 4% - 15% on retrieval accuracy.

1. Introduction

Multimedia retrieval including image retrieval and 3D shape retrieval is one of the fundamental problems in computer vision. Thanks to the development of deep learning and 3D shape datasets with rich object categories such as ShapeNet [7], 3D shape retrieval from single realistic images has recently gained more attention, owing to its wide range of applications, including scene reconstruction, 3D printing, virtual reality and e-commerce platforms.

However, despite significant progress achieved with pioneering works, using single images to retrieve the corresponding 3D shapes is still a challenging problem because of the domain gap. In order to handle this gap, one common direction in previous works is to address the retrieval task by mapping 3D shapes and query images into a common embedding space. [15] embeds 3D shapes and 2D images into a common low-level representation space using location fields. [12] assigns a texture to a 3D shape based on a texture code encoded from the 2D image to generate hard samples. Both [15, 12] are modified from triplet loss [56] for metric learning, which nearly always requires hard-negative mining for good performance, as proved by [48].

[30] proves that triplet loss is a special case of the contrastive loss when the numbers of positives and negatives are both one, which means that batch contrastive approaches subsume or significantly outperform traditional

[*] Corresponding Author is Lin Gao

triplet loss. The use of many negatives in contrastive learning for each anchor helps the model achieve state of the art performance without the need for hard-negative mining, which can be difficult to tune properly. Therefore, we introduce contrastive learning into the area of IBSR. The data studied by traditional contrastive learning are all of the same type, which is different from our task. [27, 28] perform cross-modal retrieval by applying contrastive learning, and aim to learn discriminative and modal-invariant features for data from different modalities. However, the retrieval performance will drop in the more fine-grained IBSR task if different rendered images are mapped to the one center embedding. Therefore, the introduction of contrastive learning into the task of IBSR is worth exploring.

Data augmentation plays a critical role in defining effective predictive tasks with contrastive learning. In [10], it is proved that contrastive learning needs stronger data augmentation, which helps to avoid the high complexity of the network architecture. It is also proved that it is critical to compose cropping with color augmentation in order to learn generalizable features. From another aspect, [12] finds that objects in 2D images entangled with the color make the network ineffective to push away negative pairs in the IBSR task. In order to combine the above two mentioned points at the same time, we introduce the color transfer mechanism [44] as a simple but powerful solution, which applies the colors of one image to another. The color transfer mechanism not only performs data augmentation on the input query images, but also effectively decouples the object and color in 2D images.

With the help of contrastive learning accompanied by the color transfer mechanism, we propose an efficient approach to image-based 3D shape retrieval task from both instance and category levels. 3D shapes are first converted into multi-view grayscale images. Instead of mapping multi-view images of the same object into one center embedding, our approach processes them by an attention mechanism with query image into query-specific embeddings. Inspired by [30], we design an instance loss based on self-supervised contrastive loss to pull augmented embeddings of query image closer to embeddings of its ground-truth 3D shape renderings than embeddings of all other 3D shapes. The instance loss ensures the exact shape retrieval accuracy. Similar to the instance loss, the category loss pulls embeddings of 3D shape renderings with the same category label as query image closer than embeddings of 3D shapes with different labels. The category loss is a cross-modal supervised loss, which effectively leverages the label information to push apart embeddings from different categories. In such a way, both instance and category losses avoid hard example mining existing in the triplet loss.

In order to evaluate the performance of our novel image-based 3D shape retrieval approach, we evaluate it on three challenging real-world datasets: Pix3D [50] (bed, chair, sofa, table), Comp [55] (car), and Stanford [55] (car). Quantitative results show that our approach significantly outperforms the state-of-the-art.

In summary, the key contributions of this work are:

- We propose a novel approach with a cross-modal Instance-Category loss, which is based on contrastive learning from instance and category levels, for image-based 3D shape retrieval.

- We introduce the color transfer mechanism into contrastive learning, which is a more powerful color augmentation that augmenting training images. It applies another training image as a reference, improving the robustness of the network and helping the network extract color-independent features.

- Our proposed novel approach outperforms the previous state-of-the-arts (SOTAs) on standard real-world benchmark datasets by 4% - 15% on the retrieval accuracy.

2. Related Work

3D shape retrieval from a single image has received significant attention in computer vision and graphics. Many previous works discussed how to improve the accuracy of shape retrieval by learning some critical image/shape features or reasonable embedding space with metric learning. A complete survey is beyond the scope of this paper; please refer to [51, 59] for more comprehensive discussions. In this section, we focus our discussion on recent works of 3D shape retrieval, metric learning and contrastive learning.

3D Shape Retrieval There have been growing interests in 3D shape retrieval algorithms, which include two main streams: 1) 3D model-based shape retrieval; 2) image-based shape retrieval. The first stream s3D model-based shape retrieval aims to retrieve 3D shapes based on a query shape. These methods [8, 17, 26] extract the representative 3D shape features and measure the similarity of these features, which achieve high performance (retrieval accuracy) on 3D shape retrieval or 3D shape classification tasks. However, these methods always need a 3D query shape for retrieval, which is not easy to obtain in the real world.

The second stream is image-based shape retrieval, which generally renders 3D shapes in a single view manner by estimating 3D poses from 2D images [5, 14]. [50, 31, 4] combine 2.5D sketches and a shape prior learned from past experience to reconstruct a full 3D shape for retrieval. [14] recovers the object pose using a PnP algorithm and renders depth images from 3D shapes under the estimated pose.

Although the above methods could achieve satisfactory results, 2.5D sketches prediction is itself still an open problem. Another direction in IBSR is to render 3D shapes in a multi-view manner [49, 33, 21] and then jointly map 3D shapes and RGB images into a common embedding space [34, 3, 15, 42, 52] to reduce domain gap. However, both these methods are only trained on synthetic data, and thus do not generalize well to real data due to the domain gap between query images and rendered images. [15] embeds 3D shapes and 2D images into a common low-level representation space using location fields. Despite that location fields have rich information, they are position sensitive and image location field prediction is also an open problem. [15, 12], the previous SOTA, are based on triplet loss, which has the need for hard-negative mining. Moreover, in order to consider texture information, [12] applies a texture synthesis module for 3D shapes to generate hard-negative mining, which is a learning-based network that requires additional training. Since not all 3D shapes contain color/texture information, we focus on extracting a feature from a single image, which is able to describe the 3D shape information or disentangle the appearance and shape information from the 2D images.

Deep Metric Learning (DML) The task of DML is to learn to embed the input to an embedding space such that the distribution in the embedding space is closest to the distribution of inputs for the given task, with the help of deep neural networks. The key to the success of DML involves setting appropriate sampling strategy and loss function. The deep neural model is expected to be able to distinguish different types of objects across the entire data set. However, we usually apply mini-batches in the training stage, which makes it difficult for the model to learn the global distribution of the data well, especially for relatively large data sets. Many methods have been proposed to make the mini-batch more expressive [45, 48] and some efforts try to use proxies [39] to speed up training process. Various losses have been used in recent work. Triplet loss [48], its extension to soft triplet [43], hierarchical structure [13], and classification based loss [36, 57] show superiority in the task of classification and retrieval. However, [30] proves that triplet loss is a special case of the contrastive loss when the numbers of positives and negatives are both one, which means that batch contrastive approaches subsume or significantly outperform traditional triplet loss.

Contrastive Learning Contrastive learning is a self-supervised learning technique to learn the general features by making the network learn to distinguish the data, which means the similarity or difference between two data samples. Recently, thanks to the capability of self-learning on the data without any annotations or labels, especially for the applications where labels need to be annotated in a professional and time-consuming way, the contrastive learning makes a great success and draws significant attention on 2D tasks and 3D tasks. There are many works [10, 18, 22, 23, 37, 38, 54, 53, 58] that make use of the contrastive learning to learn an embedding space which can make similar points cluster together and different points contrast. Following this, recent works attempt to develop the self-supervised learning representation on some 2D/3D tasks, such as unpaired image translation [2, 40], image generation [11, 32, 63, 29], image segmentation/classification [6, 9], 3D-base shape retrieval/classification [47, 16], object detection [60], shape analysis and understanding [61]. Inspired by the recent contrastive learning based works, our work leverages the devised category-level contrastive loss to distinguish the objects from different categories and incorporates the traditional instance-level loss, which jointly teaches the model which data is similar or different from category level to instance level. We refer to the survey [25] for more comprehensive discussions on contrastive learning.

3. Proposed Approach

We propose a novel framework for the single image 3D shape retrieval task by jointly training multiple modalities using the proposed Instance-Category loss with contrastive learning. An overview of the proposed approach is illustrated in Figure 1. Given a single query image and 3D shape databases, the task is to retrieve a 3D shape for the object in the image. For this purpose, a 3D shape is firstly converted into multi-view grayscale images by flat shading. The task is then transferred into the retrieval from the image to the image set.

3.1. Review of Color Transfer

The most commonly used method to generate positive and negative examples on input data is data augmentation [18, 10]. However, the contribution of common data augmentation, such as affine transformation, is not enough in the image-based 3D shape retrieval task because the network often learns color-related features. In order to minimize the impact of the color on retrieval, the color transfer mechanism [44] is introduced for applying the colors of one image to another. The goal of this mechanism is to do color augmentation in a simple but more realistic way, and the core strategy is to choose a suitable color space and then to apply simple operations there. Ruderman et al. [46] developed a color space, called $l\alpha\beta$, which minimizes correlation between channels for many natural scenes compared to RGB space. This space is based on data-driven human perception research that assumes the human visual system is ideally suited for processing natural scenes. In order to apply color in the $l\alpha\beta$ space, the image in RGB space needs to be converted into $l\alpha\beta$ space described in [44]. The trans-

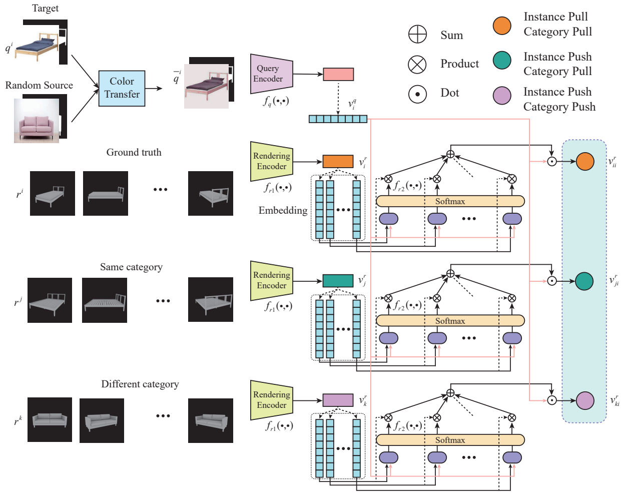

Figure 1. **The pipeline of our proposed approach.** Our network contains one query image encoder, one rendered image encoder and a learnable attention module. The query image is augmented with color transfer which uses the color of another training image as a reference before being embedded. Query encoder takes in one query image and its mask to generate query embeddings. Rendering encoder takes in one view of renderings of a 3D shape to generate view-specific embeddings. The attention module takes in one query embedding and multi-view embeddings of a 3D shape to merge multi-view embeddings into one query-specific embedding. With the category and instance level losses, the proposed framework can pull and push the embeddings of 3D shapes. The ground-truth shape is pulled at both category and instance levels. The 3D shapes with different categories are pushed at both category and instance levels. The 3D shapes with the same category are pushed at the instance level while pushed at the category level.

ferred image in $l\alpha\beta$ space is calculated by the following equation:

$$
\begin{aligned}
l^{'} &= \frac{\sigma_t^l}{\sigma_s^l}(l-\mu_t^l)+\mu_s^l \\
\alpha^{'} &= \frac{\sigma_t^\alpha}{\sigma_s^\alpha}(\alpha-\mu_t^\alpha)+\mu_s^\alpha \\
\beta^{'} &= \frac{\sigma_t^\beta}{\sigma_s^\beta}(\beta-\mu_t^\beta)+\mu_s^\beta
\end{aligned}
\tag{1}
$$

$l\alpha\beta$ and $l^{'}\alpha^{'}\beta^{'}$ are colors before and after color transfer. μ and σ stand for means and standard deviations respectively. The subscripts s and t mean the source color image and the target geometry image respectively. Finally, the transformed augmented image is then converted back into the RGB space.

3.2. Problem Formulation

Assuming dataset S contains $|S|$ instances where the i-th instance s_i consists of a query image q_i, multi-view rendering image r_i including M views, and a semantic label y_i. Formally,

$$S = \{s_i\}_{i=1}^{|S|}, s_i = (q_i, r_i, y_i), r_i = \{t_i^m\}_{m=1}^{M} \tag{2}$$

In which, $\{t_i^m\}_{m=1}^{M}$ are M views of rendered images for the i-th instance.

In order to make each query image q_i obtain a different source color during each epoch, the source color of each query image is randomly selected from other query images q_j within the same batch in a mini-batch. Here both i and j belong to B, which is the size of mini-batch. The transferred image is denoted as $\bar{q}_i$ from q_i.

Unlike r_i, whose contents are clean, q_i contains much redundant information, which may prevent the embedding module from extracting the key information, including structure of contents. For the redundant information from the background, we calculate the mask of q_i by Mask R-CNN [19] and OCRNet [62] as a guidance for its attention to the corresponding 3D shape in the 2D image. The mask of q_i is denoted as k_i. Therefore, the expression in Eq. 2 can be re-described as follows within a mini-batch situation:

$$S = \{s_i\}_{i=1}^{|B|}, s_i = (\bar{q}_i, r_i, y_i, k_i), r_i = \{t_i^m\}_{m=1}^{M} \tag{3}$$

3.3. Contrastive Learning from Two Levels

We first extract the embeddings of the input query image q_i by the image encoder $f_q(,)$. Since the model needs to extract color-invariant embeddings of the query image, the input to $f_q(,)$ is the augmented image $\bar{q}_i$. This process can be expressed as $v_i^q = f_q(\bar{q}_i, \theta_q)$. θ_q and v_i^q stand for the network parameters of $f_q(\cdot,\cdot)$ and the embeddings of q_i respectively.

The embeddings collection of the multi-views rendering images r_i is extracted by another image encoder $f_{r1}(,)$. This process can be described as $v_i^r = f_{r1}(r_i, \theta_{r1})$. θ_{r1} and v_i^r stand for the network parameters of $f_{r1}(\cdot,\cdot)$ and the embeddings of r_i respectively.

Since v_i^r is now a collection of embeddings, we merge the multi-view embeddings and get the instance-wise query-specific embeddings of r_i through an attention module $f_{r2}(,)$. This process can be described as $v_{ij}^r = f_{r2}(v_i^r, v_j^q, \theta_{r2})$. θ_{r2} and v_{ij}^r stand for the network parameters of $f_{r1}(\cdot,\cdot)$ and the embeddings of r_i for the query image q_j respectively.

Instance Loss: The core of image-based 3D shape retrieval is to pull the embeddings of the query image v_i^q closer to the embeddings of its corresponding rendered images v_{ii}^r than embeddings of rendered images from different instance v_{ji}^r where $j \in S \setminus \{S_i\}$. This motivation fits well with contrastive self-supervised learning, which requires that the features from the same class are pulled closer together than the features from different classes. In formulation, v_i^q is closer to v_{ji}^r, where $j = i$, while further when

$j \in B \setminus \{B_i\}$ in a mini-batch. Inspired by self-supervised contrastive learning (e.g., [30]), the instance loss takes the following form.

$$L_{inst} = -\sum_{i \in B} \log \frac{\exp(v_i^q \cdot v_{ii}^r / \tau)}{\sum_{j \in B} \exp(v_i^q \cdot v_{ji}^r / \tau)} \tag{4}$$

Here, the $\cdot$ symbol denotes the inner (dot) product, which is an implementation of the similarity metric. $\tau \in R^+$ is a scalar temperature parameter. Note that for each query image q_i, there is one positive pair r_i and $|B| - 1$ negative pairs r_j, where $j \in B \setminus \{B_i\}$.

Category Loss: Some datasets provide labels at the category level. In order to distinguish similar objects from different categories, we leverage category labels to improve the accuracy of category retrieval on cross-category datasets. Inspired by the supervised contrastive learning work [30], we introduce a category loss here. The category loss takes the following form:

$$L_{cats} = \sum_{i \in B} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(v_i^q \cdot v_{pi}^r / \tau)}{\sum_{j \in B} \exp(v_i^q \cdot v_{ji}^r / \tau)} \tag{5}$$

Here, $P(i)$ means $\{j | j \in B \setminus \{B_i\}$ and $y_j = y_i\}$. Note that for each query image q_i, there can be many positive pair r_i and many negative pairs r_j in category level compared to only one positive pair in the instance level loss. The overall loss of the proposed approach is as follows:

$$L_{total} = L_{inst} + \beta_1 \cdot L_{cats} \tag{6}$$

where β_1 denotes weights that balances two loss terms.

3.4. Framework Architecture

Image Encoder: The query image encoder, denoted as $f_q(,)$, applies ResNet50 [20] as backbone architecture to extract the representation embedding for the query image q_i. The first convolution layer is revised from 3 channels into 4 channels to make one more channel for accepting mask image k_i. The last fully-connected layer consists of one batch-normalization layer and one linear layer with our desired output dimension, which is 128 here. The rendered image encoder, denoted as $f_{r1}(,)$, applies ResNet34 [20] as backbone architecture to extract the representation embeddings for the multi-view rendered images r_i. The modifications here are the same as the query image encoder, except that the channel number of the first layer is changed to one. The view number M is 12 in our implementation.

Attention Module: The multi-view embeddings are merged into one query-image-specific embedding with the dimension of 128-d using attention mechanism instead of a query-irrevelant center embedding. This attention module contains one MLP layer before dot product with multi-view embeddings to generate weighted embeddings of each 3D shape for a specific query image.

3.5. Retrieval Process

In the retrieval process, the query image would be embedded into a vector directly by the image encoder without the color transfer module. Then the shape whose rendered images are embedded with highest similarity to the embeddings of the query image is picked as the matched model.

4. Experiments

Datasets: To validate our proposed method, we perform experiments on three challenging real world datasets with different object categories: Pix3D [50] (bed, chair, sofa, table), Comp [55] (car), and Stanford [55] (car) following the experiment setting of [15, 12]. For Pix3D datasets, the masks are generated by OCRNet [62] trained in Pix3D. Experiments were conducted on categories that contain more than 300 non-occluded and non-truncated samples. Therefore, there are 5,118 images and 322 shapes, with 2,648 for training and 2,470 for evaluation. For Comp Cars and Stanford Cars, we directly use the Mask R-CNN [19] pretrained on COCO [35] to generate the masks. Stanford Cars and Comp Cars focus more on challenging fine-grained retrieval. The two datasets already provide a train-test split. Stanford Cars provides 134 3D car shapes with 16,185 images (8,144 for training and 8,041 for evaluation). There are 94 3D shapes with 5,696 images (3,798 for training and 1,898 for test) in Comp Cars. Additionally, we also provide evaluation results for 3D shape retrieval using query images from seen datasets and shapes from unseen datasets, such as ShapeNet [7].

4.1. Implementation Details

Training Details. The size of the input images of the image encoder is 224×224. Each 3D shape contains 12 views of rendered images. We choose temperature $\tau = 0.1$ and $\beta_1 = 0.2$. We use the Adam optimizer with a learning rate of 5×10^{-5}, betas of $(0.5, 0.999)$ and batch size of 60. The total training epoch number is 500. We implemented our network on PyTorch [41].

Evaluation. For images with corresponding 3D shape annotations, $Acc_{\text{Top-1}}$ and $Acc_{\text{Top-10}}$ are used to measure the retrieval accuracy. $Acc_{\text{Top-1}}$ means the ratio of the first 3D shape predicted being the same as the annotation. $Acc_{\text{Top-10}}$ stands for the ratio of the ground truth shape within first 10 predicted shapes. To measure the distance between two 3D shapes, we adopt two metrics, HAU (mean modified Hausdorff Distance) and IoU (Intersection over Union) as suggested in [15] to report.

Method	Dataset	Category	seen 3D models				unseen 3D models	
			$Acc_{\text{Top-1}}$	$Acc_{\text{Top-10}}$	d_{HAU}	d_{IoU}	d_{HAU}	d_{IoU}
UDF-CGI [1]	Pix3D	bed	19.4%	46.6%	0.0821	0.3397	0.0960	0.2487
Grabner *et al.* [14]			35.1%	83.2%	0.0385	0.5598	0.0577	0.3013
LFD [15]			64.4%	89.0%	0.0152	0.8074	0.0448	0.3490
HEG-TS [12]			65.3%	95.4%	0.0122	0.8213	0.0425	0.3684
Ours			**73.3%**	**96.1%**	**0.0093**	**0.8927**	**0.0408**	**0.3999**
UDF-CGI [1]	Pix3D	chair	17.3%	49.1%	0.0559	0.3027	0.0843	0.1334
Grabner *et al.* [14]			41.3%	73.9%	0.0305	0.5469	0.0502	0.1965
LFD [15]			58.1%	81.8%	0.0170	0.7169	0.0375	0.2843
HEG-TS [12]			**87.9%**	**97.9%**	**0.0041**	**0.9063**	**0.0152**	**0.7482**
Ours			79.4%	96.3%	0.0080	0.8661	0.0190	0.6384
UDF-CGI [1]	Pix3D	sofa	21.7%	52.2%	0.0503	0.3824	0.0590	0.3493
Grabner *et al.* [14]			44.1%	89.9%	0.0197	0.7762	0.0294	0.6178
LFD [15]			67.0%	94.4%	0.0075	0.9028	0.0178	0.7472
HEG-TS [12]			72.8%	**97.7%**	0.0047	0.9070	0.0156	0.7963
Ours			**80.7%**	97.1%	**0.0045**	**0.9329**	**0.0151**	**0.8017**
UDF-CGI [1]	Pix3D	table	12.0%	34.2%	0.1003	0.1715	0.1239	0.1047
Grabner *et al.* [14]			33.9%	66.1%	0.0607	0.4500	0.0753	0.1730
LFD [15]			53.3%	80.1%	0.0288	0.6383	0.0482	0.2573
HEG-TS [12]			73.7%	92.4%	0.0170	0.7667	0.0228	0.4391
Ours			**76.9%**	**93.5%**	**0.0168**	**0.8088**	**0.0213**	**0.4701**
UDF-CGI [1]	Pix3D	mean	17.6%	45.5%	0.0722	0.2991	0.0908	0.2090
Grabner *et al.* [14]			38.6%	78.3%	0.0374	0.5832	0.0531	0.3222
LFD [15]			60.7%	86.3%	0.0171	0.7663	0.0370	0.4095
HEG-TS [12]			74.9%	95.8%	0.0095	0.8503	0.0240	0.6081
Ours			**78.9%**	**96.1%**	**0.0086**	**0.8746**	**0.0202**	**0.6317**
UDF-CGI [1]	Comp	car	2.4%	18.2%	0.0207	0.7224	0.0271	0.6344
Grabner *et al.* [14]			10.2%	36.9%	0.0158	0.7805	0.0194	0.7230
LFD [15]			20.5%	58.0%	0.0133	0.8142	0.0165	0.7707
HEG-TS [12]			67.1%	93.7%	0.0035	0.9256	0.0092	0.8591
Ours			**77.8%**	**94.1%**	**0.0023**	**0.9399**	**0.0080**	**0.9053**
UDF-CGI [1]	Stanford	car	3.7%	20.1%	0.0198	0.7169	0.0242	0.6526
Grabner *et al.* [14]			11.3%	42.2%	0.0153	0.7721	0.0183	0.7201
LFD [15]			29.5%	69.4%	0.0110	0.8352	0.0150	0.7744
HEG-TS [12]			68.4%	92.1%	0.0034	0.9210	0.0074	0.8735
Ours			**83.4%**	**96.4%**	**0.0021**	**0.9431**	**0.0060**	**0.9217**

Table 1. **Results on the Pix3D, Comp Cars, and Stanford Cars datasets for image-based 3D shape Retrieval.** Seen parts have ground truth so that retrieval accuracy is main metric. For evaluation on unseen shapes on the ShapeNet, Haus-distance and 3D IoU are the main measurement criteria. We also test the unseen model on the ShapeNet [7] dataset.

4.2. Benchmark Performance

Quantitative Results. We compare our approach to baselines [1, 14] and SOTA methods [15, 12] in Pix3D, Comp and Stanford Cars, which is reported in Table 1. In General, the retrieval performance of our method on these three seen datasets far exceeds baselines and SOTAs over all metrics. For Top-1 accuracy metrics, it outperforms the SOTA by about $4\% - 15\%$ across all three datasets. We notice that HEG-TS [12] achieves better performance in chair category on Pix3D dataset. We find that chairs in most of the query images on Pix3D dataset almost show one color as a whole, which is very close to the style of the result converted by TSM in [12] (see Fig. 3 in [12]). For other category in Pix3D or other datasets with more fine-grained color patterns, the performance of [12] drops significantly. For cars datasets, our approach still has almost 80% Top1 retrieval accuracy while the performance of both previous SOTAs reduces substantially compared to Pix3D dataset. For unseen dataset ShapeNet, our approach achieves much better IoU and HAU compared with previous SOTAs.

Qualitative Results. Visualization results of some query image by our retrieval approach on ShapeNet dataset are reported in Figure 2. It shows that our proposed approach could retrieve shapes that are similar to the target.

4.3. Ablation studies

Contrastive Learning vs. Triplet Loss: We first discuss the benefit of the introduction of contrastive loss into IBSR task. We do experiment with the traditional triplet loss com-

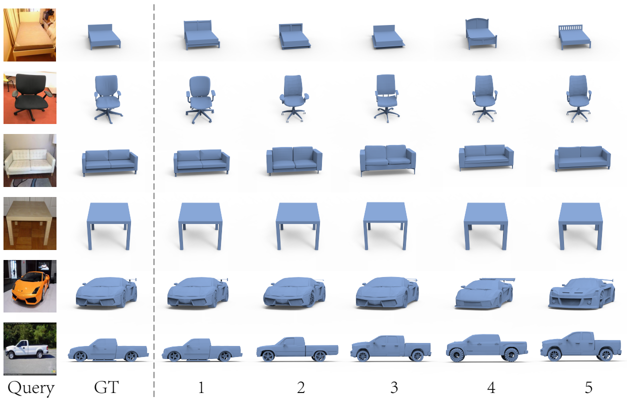

Figure 2. **Qualitative results of shape retrieval on ShapeNet.** RGB query images and ground-truth are in the first and second columns. The top-5 ranked models retrieved are shown in the following five columns.

Dataset	Aug	Instance	Category	Acc_{Top-1}
Pix3D	CT	Contrastive	✓	**78.9%**
Pix3D	×	Contrastive	✓	75.2%
Pix3D	HSV	Contrastive	✓	76.6%
Pix3D	CT	Contrastive	×	74.7%
Pix3D	×	Contrastive	×	71.4%
Pix3D	CT	Triplet	✓	66.5%
Pix3D	CT	Triplet	×	62.7%
Comp	CT	Contrastive	N/A	**77.8%**
Comp	HSV	Contrastive	N/A	73.0%
Comp	×	Contrastive	N/A	67.3%
Comp	CT	Triplet	N/A	64.0%
Stanford	CT	Contrastive	N/A	**83.4%**
Stanford	HSV	Contrastive	N/A	81.5%
Stanford	×	Contrastive	N/A	81.7%
Stanford	CT	Triplet	N/A	80.3%

Table 2. **Results of ablation studies on augmentation method, instance loss and category loss**. 'Aug' and 'CT' are the abbreviation of augmentation and color transfer. ✓ and × in Category means with category and without category loss. × in Aug means that only the traditional affine transformation is used for data augmentation instead of any other data augmentation algorithm about color. We can see that CT + Contrastive instance loss + Category loss is the best choice on cross category dataset and CT + Contrastive instance loss is the best choice on single category datasets.

Dataset	Category	$Acc_{\text{Top-1}}$	Acc_{Cats}
Pix3D	✓	**78.9%**	**96.1%**
Pix3D	×	75.2%	95.9%

Table 3. **Results of ablation studies on the rationality of existence of category loss**. ✓ and × in 'Category' means with and without category loss respectively. We apply Top-1 retrieval accuracy at both instance and category levels to evaluate the performance of category loss.

pared to our method across all 3 datasets. As mentioned in FaceNet [48], it is crucial to select hard and semi-hard triplets, that are active and can therefore contribute to improving the network. For a fair comparison, we design a similar Instance-Category loss with efficient sample strategy for triplet loss. For each query image, we search all shapes to find all semi-hard and hard samples at the instance level in one mini-batch. We could not consider all the shapes of different categories like our previous proposed approach in category level because triple loss uses one positive and one negative as input while our approach uses many positive and many negative samples. As a remedy, we constrain the farthest embeddings of 3D shapes with the same category and the nearest embeddings of 3D shapes with the different category label to form a tuple of category level triplets for each query image. The experimental results are shown in Table 2. The Top-1 retrieval accuracy of triplet loss combined with category supervision and color transfer mechanism is 66.5%, which is about 10% lower than our method. It proves that the contrastive loss fits better than triplet loss in IBSR task. The number of positive and negative pairs is the key distinctive statistic between triplet loss and contrastive loss. Triplet loss uses exactly one positive and one negative pair per anchor while contrastive loss could apply many positive and many negative pairs per anchor. This means that contrastive loss could contain more information and avoid relying on hard example mining.

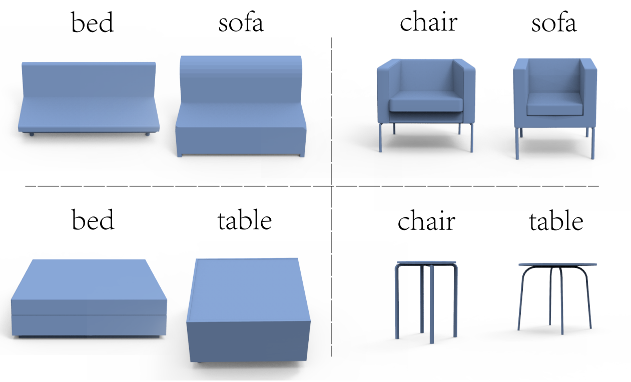

Figure 3. **Visualization of similar shapes with different category labels in Pix3D dataset.** The shape pairs in this figure are hard to distinguish from instance level. With the category level label information, the shape pairs could be distinguished.

Category Loss: There are two ablation studies here for the category loss. One is for the rationality of the existence of the category loss. We design the category loss to help distinguish similar objects from different categories. Figure 3 shows some similar objects with different categories. The experimental results reported in Table 3 show that the category loss leads to a better Top-1 retrieval accuracy in both the instance and category level. We visualized the aggregation of the embeddings of all shapes on Pix3D in 2D

Dataset	w/ ground-truth shape	Acc_{Top-1}
Pix3D	✓	**78.9%**
Pix3D	×	76.7%

Table 4. **Results of ablation studies on the formulation of category loss**. 'w/ ground-truth shape' means whether to take ground-truth shape corresponding to query image into consideration in Eq. 5. ✓ and × means with and without ground-truth shape respectively. It shows that consideration with ground-truth shape at the category level leads to a better result.

for one query image in Figure 4. We can see that with the help of the category loss, shapes belonging to the same category are pulled together in the embedding space, while simultaneously pushing apart different categories. The above experiments prove the rationality of the category loss.

Another one is for the formulation of the category loss. We notice that the positive pairs do not have a bias in [30] while the ground-truth shape corresponding to the query image has quite higher similarity than other shapes with the same category label to query image. Formally, the question is whether to remove $\exp(v_i^q \cdot v_{ii}^r / \tau)$ from the denominator and the numerator in Eq. 5. The experiment result reported in Table 4 shows that taking ground-truth shape has a better performance. We think that the lack of constraint on ground-truth label in the category level may cause the proportion of ground-truth shape to decrease in all shapes, which may drag down the performance of the instance loss.

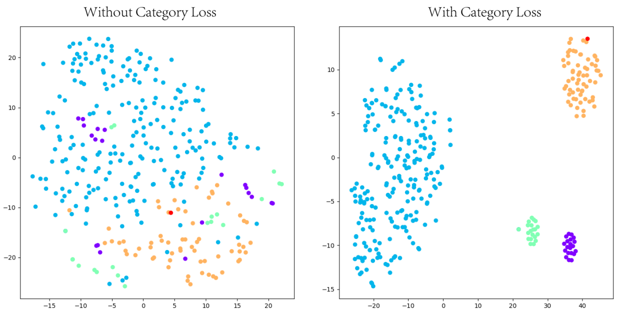

Figure 4. **Visualization of Category Loss.** We visualized all embedding spaces for a query image in Pix3D by t-SNE. The purple, green, orange, blue and red represent bed, sofa, table, chair and the query image. The query image belongs to table category. The left is the result after without category loss, and the right is the result with category loss. In this example, although the query image can find the correct category in both the left and right methods (red point covering orange point), we can see that the same category points will gather closer and different categories will be far away from each other in the entire space when category loss is used.

Color Augmentation: We adjust the method of color augmentation as shown in Table 2 to further demonstrate the importance of it. All color augmentations are combined with traditional augmentation, which includes affine transformation, crop and flip. HSV augmentation is performed with 0.5 probability to change hue, saturation and value of the input image randomly. Results with the color transfer mechanism is on average 3% better than that with HSV across all datsets. The reasons for this results may lie in that color transfer mechanism applies the colors of one image to another to decouple objects and color in 2D images while HSV does not use other images for augmentation. What's more, the L$\alpha\beta$ space of color transfer mechanism helps de-correlate color space, where 3 channels are more independent than that in HSV space. From the view of visualization, the image transformed by the color transfer mechanism is more natural. Compared to only with common data augmentation methods, our methods get 3%, 10% and 2% retrieval accuracy gain. We notice that the gain in Comp Cars is much larger than other 2 datsets. The reason is that there are 2,700, 3,800 and 8,000 images in Pix3D, Comp Cars and Stanford Cars datasets. The image amount is not enough in Comp Cars. Despite of the fact that Pix3D only has 2700, it contains 4 categories with more diversity. We also notice that the HSV augmentation in Stanford Cars has a 0.2% negative gain. As a result, we believe that when the number of training set increases to a certain extent, a simple HSV augmentation cannot improve the performance of the network while the color transfer mechanism still has an obvious effect with reference to other query images. The above experiments prove that color transfer mechanism plays an important role in our proposed method.

5. Conclusion

In this work, we present a novel approach with a cross-modal Instance-Category loss, which is based on contrastive learning from instance and category levels, for image-based 3D shape retrieval. We introduce the color transfer mechanism as a strong data augmentation for contrastive learning and decouple objects and color in 2D images. Experimental results show that our proposed method significantly improves the performance of previous SOTA with the same experiment settings. In future work, we will implement our proposed approach in Jittor [24], which is a fully just-in-time (JIT) compiled deep learning framework with higher performance.

Acknowledgement This work was supported by the National Natural Science Foundation of China (No. 61872440 and No. 62061136007), the Royal Society Newton Advanced Fellowship (No. NAF\R2\192151), the Youth Innovation Promotion Association CAS, the Alibaba Innovative Research (AIR) Program and the Open Research Projects of Zhejiang Lab (No. 2021KE0AB06) .

References

[1] Mathieu Aubry and Bryan C. Russell. Understanding deep features with computer-generated imagery. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015. 6

[2] Kyungjune Baek, Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Hyunjung Shim. Rethinking the truly unsupervised image-to-image translation, 2021. 3

[3] Song Bai, Xiang Bai, Zhichao Zhou, Zhaoxiang Zhang, and Longin Jan Latecki. GIFT: A real-time and scalable 3D shape search engine. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3

[4] Aayush Bansal, Bryan Russell, and Abhinav Gupta. Marr revisited: 2D-3D alignment via surface normal prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[5] Eric Brachmann, Frank Michel, Alexander Krull, Michael Ying Yang, Stefan Gumhold, and carsten Rother. Uncertainty-driven 6D pose estimation of objects and scenes from a single rgb image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[6] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12546–12558. Curran Associates, Inc., 2020. 3

[7] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015. 1, 5, 6

[8] Ding-Yun Chen, Xiao-Pei Tian, Yu-Te Shen, and Ming Ouhyoung. On visual similarity based 3D model retrieval. *Computer Graphics Forum*, 22(3):223–232, 2003. 2

[9] John Chen, Samarth Sinha, and Anastasios Kyrillidis. Imclr: Implicit contrastive learning for image classification. *arXiv preprint arXiv:2011.12618*, 2020. 3

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020. 2, 3

[11] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3D imitative-contrastive learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[12] Huan Fu, Shunming Li, Rongfei Jia, Mingming Gong, Binqiang Zhao, and Dacheng Tao. Hard example generation by texture synthesis for cross-domain shape similarity learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 14675–14687. Curran Associates, Inc., 2020. 1, 2, 3, 5, 6

[13] Weifeng Ge. Deep metric learning with hierarchical triplet loss. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 3

[14] Alexander Grabner, Peter M. Roth, and Vincent Lepetit. 3D pose estimation and 3D model retrieval for objects in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 6

[15] Alexander Grabner, Peter M. Roth, and Vincent Lepetit. Location field descriptors: Single image 3D model retrieval in the wild. In *2019 International Conference on 3D Vision (3DV)*, pages 583–593, 2019. 1, 3, 5, 6

[16] Li Han, Jingyu Piao, Yuning Tong, Bing Yu, and Pengyan Lan. Deep learning for non-rigid 3D shape classification based on informative images. *Multimedia Tools and Applications*, 80(1):973–992, 2021. 3

[17] Zhizhong Han, Honglei Lu, Zhenbao Liu, Chi-Man Vong, Yu-Shen Liu, Matthias Zwicker, Junwei Han, and C. L. Philip Chen. 3D2SeqViews: Aggregating sequential views for 3D global feature learning by cnn with hierarchical attention aggregation. *IEEE Transactions on Image Processing*, 28(8):3986–3999, 2019. 2

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[19] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 4, 5

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 5

[21] Xinwei He, Tengteng Huang, Song Bai, and Xiang Bai. View n-gram network for 3D object retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 3

[22] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4182–4192. PMLR, 13–18 Jul 2020. 3

[23] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2019. 3

[24] Shi-Min Hu, Dun Liang, Guo-Ye Yang, Guo-Wei Yang, and Wen-Yang Zhou. Jittor: a novel deep learning framework with meta-operators and unified graph execution. *Science China Information Sciences*, 63(12):1–21, 2020. 8

[25] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey

on contrastive self-supervised learning. *Technologies*, 9(1), 2021. 3

[26] Jianwen Jiang, Di Bao, Ziqiang Chen, Xibin Zhao, and Yue Gao. MLVCNN: Multi-loop-view convolutional neural network for 3D shape retrieval. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8513–8520, Jul. 2019. 2

[27] Longlong Jing, Yucheng Chen, Ling Zhang, Mingyi He, and Yingli Tian. Self-supervised modal and view invariant feature learning. *CoRR*, abs/2005.14169, 2020. 2

[28] Longlong Jing, Elahe Vahdani, Jiaxing Tan, and Yingli Tian. Cross-modal center loss for 3D cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3142–3151, June 2021. 2

[29] Minguk Kang and Jaesik Park. Contragan: Contrastive learning for conditional image generation. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21357–21369. Curran Associates, Inc., 2020. 3

[30] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc., 2020. 1, 2, 3, 5, 8

[31] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, December 2013. 2

[32] Kwot Sin Lee, Ngoc-Trung Tran, and Ngai-Man Cheung. Infomax-gan: Improved adversarial image generation via information maximization and contrastive learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3942–3952, January 2021. 3

[33] Tang Lee, Yen-Liang Lin, Hungyueh Chiang, Ming-Wei Chiu, Winston Hsu, and Polly Huang. Cross-domain image-based 3D shape retrieval by view sequence learning. In *2018 International Conference on 3D Vision (3DV)*, pages 258–266, 2018. 3

[34] Yangyan Li, Hao Su, Charles Ruizhongtai Qi, Noa Fish, Daniel Cohen-Or, and Leonidas J. Guibas. Joint embeddings of shapes and images via cnn image purification. *ACM Trans. Graph.*, 34(6), Oct. 2015. 3

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 5

[36] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 507–516. JMLR.org, 2016. 3

[37] Sindy Löwe, Peter O' Connor, and Bastiaan Veeling. Putting an end to end-to-end: Gradient-isolated learning of representations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 3

[38] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[39] Yair Movshovitz-Attias, Alexander Toshev, Thomas K. Leung, Sergey Ioffe, and Saurabh Singh. No fuss distance metric learning using proxies. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 3

[40] Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 319–345, Cham, 2020. Springer International Publishing. 3

[41] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 5

[42] Charles R. Qi, Hao Su, Matthias Niessner, Angela Dai, Mengyuan Yan, and Leonidas J. Guibas. Volumetric and multi-view cnns for object classification on 3D data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3

[43] Qi Qian, Lei Shang, Baigui Sun, Juhua Hu, Hao Li, and Rong Jin. SoftTriple loss: Deep metric learning without triplet sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 3

[44] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer graphics and applications*, 21(5):34–41, 2001. 2, 3

[45] Oren Rippel, Manohar Paluri, Piotr Dollar, and Lubomir Bourdev. Metric learning with adaptive density discrimination, 2016. 3

[46] Daniel L Ruderman, Thomas W Cronin, and Chuan-Chin Chiao. Statistics of cone responses to natural images: implications for visual coding. *JOSA A*, 15(8):2036–2045, 1998. 3

[47] Aditya Sanghi. Info3D: Representation learning on 3D objects using mutual information maximization and contrastive learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 626–642, Cham, 2020. Springer International Publishing. 3

[48] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 1, 3, 7

[49] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3D shape recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015. 3

[50] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B. Tenenbaum, and William T. Freeman. Pix3D: Dataset and methods for single-image 3D shape modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 5

[51] J.W.H. Tangelder and R.C. Veltkamp. A survey of content based 3D shape retrieval methods. In *Proceedings Shape Modeling Applications, 2004.*, pages 145–156, 2004. 2

[52] Flora Ponjou Tasse and Neil Dodgson. Shape2Vec: Semantic-based descriptors for 3D shapes, sketches and images. *ACM Trans. Graph.*, 35(6), Nov. 2016. 3

[53] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding, 2020. 3

[54] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. 3

[55] Yaming Wang, Xiao Tan, Yi Yang, Xiao Liu, Errui Ding, Feng Zhou, and Larry S. Davis. 3D pose estimation for fine-grained object categories. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2018. 2, 5

[56] Kilian Q Weinberger, John Blitzer, and Lawrence Saul. Distance metric learning for large margin nearest neighbor classification. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2006. 1

[57] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 499–515, Cham, 2016. Springer International Publishing. 3

[58] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3

[59] Yun-Peng Xiao, Yu-Kun Lai, Fang-Lue Zhang, Chunpeng Li, and Lin Gao. A survey on deep geometry learning: From a representation perspective. *Computational Visual Media*, 6(2):113–133, 2020. 2

[60] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection, 2021. 3

[61] Saining Xie, Jiatao Gu, Demi Guo, Charles R. Qi, Leonidas Guibas, and Or Litany. PointContrast: Unsupervised pretraining for 3D point cloud understanding. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 574–591, Cham, 2020. Springer International Publishing. 3

[62] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 173–190, Cham, 2020. Springer International Publishing. 4, 5

[63] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 833–842, June 2021. 3