Dual-level Embedding Alignment Network for 2D Image-Based 3D Object Retrieval KEYWORDS 3D Object Retrieval; Domain Adaptation; Multi-View Learning * Corresponding

Article · December 2019

CITATIONS
0

READS
43

3 authors, including:

Heyu Zhou
Tinajin University
5 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Weizhi Nie
Tianjin University
77 PUBLICATIONS **886** CITATIONS
SEE PROFILE

Dual-level Embedding Alignment Network for 2D Image-Based 3D Object Retrieval

Heyu Zhou
School of Electrical and Information Engineering, Tianjin University
zhy455@tju.edu.cn

An-An Liu*
School of Electrical and Information Engineering, Tianjin University
liuanan@tju.edu.cn

Weizhi Nie*
School of Electrical and Information Engineering, Tianjin University
weizhinie@tju.edu.cn

ABSTRACT

Recent advances in 3D modeling software and 3D capture devices contribute to the availability of large-scale 3D objects. However, manually labelled large-scale 3D object dataset is still too expensive to build in practice. An intuitive idea is to transfer the knowledge from label-rich 2D images (source domain) to unlabelled 3D objects (target domain) to facilitate 3D big data management. In this paper, we propose an unsupervised dual-level embedding alignment (DLEA) network for a new task, 2D image-based 3D object retrieval. It mainly consists of two modules, visual feature learning and cross-domain feature adaptation, for jointly optimizing. The first module transforms individual 3D object into a set of multi-view images and utilizes 2D CNNs to extract visual features of both multi-view image sets and the source 2D images. For multi-view fusion by reducing the distribution divergence between both domains, we propose a cross-domain view-wise attention mechanism to adaptively compute the weights of individual views and aggregate them into a compact descriptor to narrow the gap between source and target domains. With the visual representation of both domains, the module of cross-domain feature adaptation aims to enforce the domain-level and class-level embedding alignment of cross-domain feature spaces. For domain-level embedding alignment, we train a discriminator to align the global distribution statistics of both spaces. For class-level embedding alignment, we map the features in the same class but from different domains nearby through aligning the centroid of each class from both domains. To our knowledge, this is the first unsupervised work to jointly realize cross-domain feature learning and distribution alignment in an end-to-end manner for this new task. Moreover, we constructed two new datasets, MI3DOR and MI3DOR-2, to advocate the research on this topic. Extensive comparison experiments can demonstrate the superiority of DLEA against the state-of-art methods.

KEYWORDS

3D Object Retrieval; Domain Adaptation; Multi-View Learning

*Corresponding Author: An-An Liu and Weizhi Nie

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3351011

ACM Reference Format:
Heyu Zhou, An-An Liu, and Weizhi Nie. 2019. Dual-level Embedding Alignment Network for 2D Image-Based 3D Object Retrieval. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3343031.3351011

1 INTRODUCTION

3D techniques have gained prominence in various applications, such as medical diagnosis, 3D printing and digital entertainment [18], which has led to huge deluge of 3D content. It's mandatory to develop scalable and effective 3D object retrieval methods for 3D data management [3, 5, 15, 33, 40]. Most of previous literatures focus on searching the relevant 3D objects from the base datasets given a query 3D object. Compared to the query 3D objects, 2D images are much easier to access to under most circumstances. Thus, there is a strong requirement to design an effective 2D image-based 3D object retrieval method for real applications.

1.1 Motivation

2D image-based 3D object retrieval aims to find the relevant 3D objects from the assigned dataset given a query 2D image. Although fruitful line of works have been done in 3D object retrieval [38], there still exist three critical problems for this novel task:

1). ***How to enforce visual domain adaptation between 2D images and 3D objects without strong dependence on large-scale labelled data.*** It is an intuitive way to align the samples in the same class but from different domains for domain adaptation. However, it is highly dependent on large-scale labelled 2D images and 3D objects. Considering the heavy work to label millions of 2D images for ImageNet and so on, manually labelled large-scale 3D object dataset is too expensive to build in practice. Based on the theory on unsupervised domain adaptation, there is a strong motivation to tackle this problem by transferring the knowledge from the label-rich 2D images (source domain) to the unlabelled 3D objects (target domain).

2). ***How to learn the domain-invariant yet discriminative features for cross-domain data.*** Although much work has been done on visual domain adaptation, few studies have been given to a special case when the source sample is 2D images while the target sample is the rendered multi-view image set of 3D objects. It is intuitive that even with the optimal classifier on source images, it's still difficult to recognize target 3D objects because the multi-view image sets have significantly different characteristics comparing against the source 2D images, such as texture, viewpoint change, and background. Moreover, prior domain adaptation methods, which usually aim to exploit shared feature space in two

domains, will fail for this novel task. Due to the large discrepancy between the 2D and 3D feature space, there may not exit such a common space where the distributions of 2D and 3D features are the same and the data characteristics are also maximally preserved in the mean time. Therefore, it is mandatory to design the domain-invariant yet discriminative feature learning method for this task.

3). ***How to fully exploit the semantic information of each category for domain adaptation.*** Most existing domain adaptation methods [6] usually align the global distribution statistics of two domains based on generative adversarial networks (GAN) [8]. Specifically, they jointly trained a feature extractor and a domain discriminator. The domain discriminator aims to distinguish whether the sampled feature comes from the source domain or target domain while the feature extractor is trained to fool it. However, they ignore the crucial semantic information of each category. Even with ideal confusion alignment, it is difficult to guarantee that the samples from different domains but with the identical label will be mapped nearby in the feature space. For example, features of 3D airplanes may be mapped near features of the 2D desks. The lack of semantic alignment can lead to performance reduction [11, 23, 26]. Consequently, it's necessary to enforce the class-level embedding alignment by exploiting the semantic information.

To address the aforementioned problems, we propose a dual-level embedding alignment network (DLEA). As shown in Fig. 1, DLEA mainly consists of two key modules, visual feature learning and cross-domain feature adaptation. In the first module, we first render each 3D object from preset viewpoints to generate the corresponding multi-view image set. Both 2D images and multi-view image sets are encoded by 2D CNNs for visual feature extraction. With the visual representation of 2D and 3D data, we propose a cross-domain view-wise attention mechanism to automatically aggregate multiple view features into a compact 3D object descriptor. In the second module, we enforce the embedding alignment in both domain and class levels. Domain-level embedding alignment employs a discriminator to align the global distribution statistics of both 2D and 3D feature space. For the class-level embedding alignment, we firstly employ the classifier trained in the source domain to assign the pseudo labels for the target domain, which can guide the class-level embedding alignment. In return, the semantic information can further improve the performance of the classifier to recognize the 3D object. This cycle will iteratively enhance the classification accuracy on target samples. Obviously, there may exist false labels in pseudo labels, and we expect that the correct-pseudo-labelled samples can minimize the negative effect caused by the falsely-pseudo-labelled samples. Therefore, we align the centroid of each class in two domains rather than treat the pseudo labels as ground truth.

1.2 Contribution

The main contributions of the proposed method can be summarized as follows:

- This paper proposes a novel method for 2D-image based 3D object retrieval. To our knowledge, it is the first unsupervised work to jointly realize cross-domain feature learning and distribution alignment in an end-to-end manner for this new task.
- This paper proposes a cross-domain view-wise attention mechanism to adaptively compute the weights of individual view features and fuse them into a compact descriptor to narrow the gap between the source and target domains.
- Different from most of domain adaptation methods, which only enforce the alignment of global domain statics, DLEA can simultaneously enforce the domain-level and class-level embedding alignment.
- We built two new datasets, MI3DOR and MI3DOR-2, for the evaluation on 2D image-based 3D object retrieval.

The rest of this paper is structured as follows. We review the related works in Section. 2. Section. 3 details the proposed method. In Section. 4, we illustrate the experimental settings and discuss the experimental results. Finally, we conclude this paper in Section. 5.

2 RELATED WORKS

In this section, we will briefly review the recent progress on both 3D object retrieval and domain adaptation.

2.1 3D Object Retrieval

Generally, the current 3D object retrieval methods can be classified into two types, model-based methods and view-based methods.

Model-based methods opt to generate 3D descriptors directly from 3D data, such as meshes, point clouds, voxel [25, 29]. Wu *et al.* proposed 3DShapeNets [37] by applying a convolutional deep belief network to 3D voxel grid for visual representation. Li *et al.* proposed Self-Organizing Network (SO-Net) [16]. This method can build a Self-Organizing Map to model the spatial distribution of point cloud. Then, SO-Net represents the point cloud by performing hierarchical feature on individual points and SOM nodes. Wu *et al.* proposed 3D Generative Adversarial Network (3DGAN) [36]. This method combines both the volumetric convolutional networks and the generative adversarial nets to generate 3D objects from a probabilistic space. However, the model-based methods have not been widely adopted in real applications since 1) The data sparsity caused by voxelized 3D objects and the costly computation of 3D convolution limit the practical application of volume-based methods; 2) Point-based methods are constrained by the resolution of dealing with disordered points and discovering the correlation of different points.

View-based methods tend to describe individual 3D object with a set of multi-view images [17]. Su *et al.* proposed Multi-View Convolutional Neural Network (MVCNN) [32]. This method captures multiple views of a 3D object, then employs the CNN to extract individual view features and at last max-pools them into a compact 3D descriptor. Feng *et al.* proposed Group-View Convolutional Neural Network (GVCNN) [5]. Based on MVCNN, this method first groups the multiple view features into several clusters based on the importance of each view, then max-pools the features in each set and at last fuse the multi-group features to represent the 3D object. This method can achieve better performance than the original MVCNN. Han *et al.* proposed SeqViews2SeqLabels [9]. This method consists of an encoder-RNN and a decoder-RNN. The encoder aims to learn the global features of sequential views while

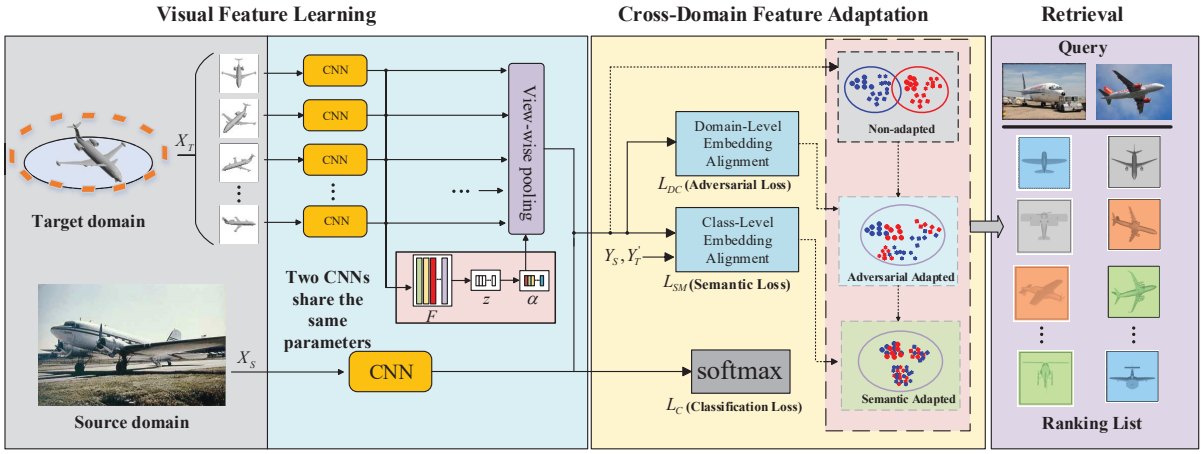


Figure 1: DLEA for 2D image-based 3D object retrieval. DLEA mainly consists of two key modules, visual feature learning and cross-domain feature adaptation. In the first module, we render the 3D object from preset viewpoints to generate multi-view image set. Both the 2D image and the multi-view image set are encoded by the CNN to extract the visual features. We design a cross-domain view-wise attention mechanism to automatically aggregate multiple view features into a compact 3D object descriptor. In the second module, we enforce the embedding alignment in both domain and class levels. Domain level embedding alignment employs a discriminator to generate domain-invariant features. Then, we align the centroid of each class in two domains to enforce the alignment in class level. At last, we can extract visual features of both 2D and 3D data for retrieval. Blue/Red points denote source/target samples, respectively.

the decoder-RNN predicts sequential labels step by step to perform more accurate classification. Kanezaki1 *et al.* proposed RotationNet [13]. This method can take multiple view images of an object as input and predict its category and viewpoint label. He *et al.* studied variants of deep metric learning losses for 3D object retrieval [10]. This method enforces the distance between the sample and its corresponding category center closer than those from different classes by introducing triplet-loss.

2.2 Domain Adaptation

Generally, current domain adaptation methods can be divided into two classes, traditional transfer learning methods and deep transfer learning methods.

Traditional transfer learning methods generalize a learner across different domains of different distributions by either matching the conditional distributions [4] or the marginal distributions [34]. Gong *et al.* proposed Conditional Transferable Components (CTC) [7]. This method proposes to learn conditional transferable components under the circumstance where the distribution of the covariate and the conditional distribution of the target given the covariate change across domains for domain adaptation. Pan *et al.* proposed Transfer Component Analysis (TCA)[27]. This method can learn a set of components which can reduce the distances between source and target samples in a Reproducing Kernel Hilbert Space (RKHS) by using Maximum Mean Discrepancy (MMD) strategy. Long *et al.* proposed Joint Distribution Adaptation (JDA) [21]. This method can construct effective feature representation for substantial distribution difference by jointly adapting the marginal distribution and conditional distribution.

Deep transfer learning methods tend to study how to utilize knowledge from other fields by deep neural networks. Hong *et al.* proposed Conditional Domain Adversarial Networks (CDAN) [20]. This method can make full use of the discriminative information conveyed in the classifier predictions to improve the performance of adversarial adaptation. Saito *et al.* proposed to enforce the distribution alignment by exploiting the task-specific decision boundaries [12]. Specifically, this method proposes to detect target samples far from the support of the source by maximizing the discrepancy between two classifiers' outputs. Shen *et al.* proposed Wasserstein Distance Guided Representation Learning (WDGRL) [30]. This method utilizes an adversarial network to compute the empirical Wasserstein distance between source and target samples and then minimizes the estimated Wasserstein distance in an adversarial manner by optimizing the feature extractor. Long *et al.* proposed Joint Adaptation Networks (JAN) [21]. This method employs multiple domain-specific layers across source and target domains to align the joint distributions based on a joint maximum mean discrepancy criterion.

3 PROPOSED METHOD

In this section, we first present the method of the dual-level embedding alignment (DLEA). Then, we will throw light on the key modules in details. For unsupervised domain adaptation, we have access to n_s labelled samples $\{(x_S^{(i)}, y_S^{(i)})\}_{i=1}^{n_s}$ from the source domain D_S, where $\{x_S^{(i)} \in X_S, y_S^{(i)} \in Y_S\}$ and n_t unlabelled samples $\{(x_T^{(i)})\}_{i=1}^{n_t}$ from the target domain D_T, where $x_T^{(i)} \in X_T$. $Y_T^{'}$ means the pseudo labels of target samples. In our work, 2D labelled image is selected as the source domain and the 3D objects without label requirement for the retrieval task is considered as the target domain.

3.1 Overview

As shown in Fig. 1, DLEA mainly contains two key modules.

1 **Visual Feature Learning:** This module first extracts the visual features of both multi-view image sets of 3D objects and 2D images with 2D CNNs. Then, we employ a cross-domain view-wise attention mechanism to fuse multiple view features into a compact 3D object descriptor.
2 **Cross-domain Feature Adaptation:** This module aims to align the distributions of cross-domain features in both domain and class levels. We design a domain discriminator to enforce the alignment of global domain statics. For class-level embedding alignment, we map the features in the same class but different domains nearby through aligning the centroid of each class in two domains.

3.2 Visual Feature Learning

Given 3D objects, we can generate multi-view image sets to represent individual 3D objects by the Phong reflection mode [28]. We set the virtual camera array similar to [18]. Then the multi-view image set, $S = \{s_i\}_{i=1}^{N}$ (N is the number of views), are passed through 2D CNNs to generate visual features, $F = \{f_i\}_{i=1}^{i=N}, f_i \in R^{1*D}$.

We design a cross-domain view-wise attention mechanism to automatically fuse multiple view features into a compact descriptor. In order to tackle the issue of exploiting view dependencies, we first employ global average pooling operation to generate view-wise statistics. Formally, a statistic $z \in R^{N*1}$ is generated by individual view features, such that the i-th element of z is calculated by:

$$z_i = \frac{1}{D}\sum_{j=1}^{D} f_i(1,j) \tag{1}$$

To make full use of the information aggregated in z and capture the view-wise dependencies, the function should observe two criteria: 1) it can learning a nonlinear interaction to exploit the intrinsic hierarchical correlation and discriminability among views; 2) it must learn a non-mutually-exclusive relationship to ensure the multiple views can be emphasized otherwise the local saliency in individual views will be neglected. To meet the criteria, we opt to employ a gating mechanism with a sigmoid activation:

$$\alpha = \sigma\left(W_2\delta\left(W_1 z\right)\right) \tag{2}$$

where σ, δ refer to sigmoid and ReLU function, respectively, $W_1 \in R^{1\times N}$ and $W_2 \in R^{N\times 1}$ and $\alpha = \{\alpha_i\}_{i=1}^{i=N}, \alpha_i \in R^1$. α can be seen as a context view vector of a fixed query, "what is the informative view". Then, we can describe the 3D object based on α, as following:

$$\varphi(F,\alpha) = \sum_{i=1}^{N} \alpha_i f_i \tag{3}$$

For 2D image, we can also employ the identical 2D CNNs to extract visual features in the source domain.

3.3 Cross-Domain Feature Adaptation

This module mainly consists of two parts, domain-level embedding alignment and class-level embedding alignment.

3.3.1 ***Domain-Level Embedding Alignment.*** As shown in Fig. 1, the features of source domain are far from those of target domain when no alignment is enforced on both 2D and 3D features. To generate domain-invariant features on two domains, we train a visual classifier by minimizing the source classification error and the difference between two domains. Specifically,

$$L = \underbrace{\mathbb{E}_{(x,y)\sim D_S}\left[J\left(f\left(x\right),y\right)\right]}_{L_C(X_S,Y_S)} + \lambda \underbrace{d\left(X_S,X_T\right)}_{L_{DC}(X_S,X_T)} \tag{4}$$

where $J(\cdot)$ represents the classification cross entropy loss, D_S represents source domain, λ is the hyperparameter to balance the classification loss L_C and domain confusion loss L_{DC} and d(·) represents the divergence between two domains. Here, we select the domain adversarial similarity loss to measure the difference. Specifically, we utilize an additional domain discriminator **D** to distinguish whether the features from source domain or target domain while the feature extractor (2D CNNs) is trained to fool the **D**. The goal of this two-player game is to reach an equilibrium where the features arising from 2D CNNs can be domain-invariant. Formally,

$$\begin{aligned} d\left(X_S,X_T\right) =& \mathbb{E}_{x\sim D_S}\left[log\left(1-D(G\left(x\right))\right)\right] + \\ & \mathbb{E}_{x\sim D_T}\left[log\left(D(G\left(x\right))\right)\right] \end{aligned} \tag{5}$$

where G represents the feature extractor 2D CNNs, D_T represents the target domain.

3.3.2 ***Class-Level Embedding Alignment.*** As shown in Fig. 1, domain-level embedding alignment can contribute to invariance of the global distribution and consequently the source and target features can be mixed up. However, domain-invariance does not mean discrimination. Features of 3D airplanes may be mapped near features of 2D desks while satisfying the condition of domain-invariant.

To resolve this problem, we design a centroid alignment operation to semantically align the embedding by explicitly restricting the distance between centroids in the same class but different domains. However, we don't have the label information of the target samples for the unsupervised domain adaptation task. Naturally, we resort to pseudo labels to meet the requirement of distribution matching in the class level. Obviously, there must be some false labels and they have a negative influence on the domain adaptation. Centroid alignment operation can not only enforce the class-level alignment semantically but also suppress the noisy signals in those false pseudo-labelled samples. Specifically, we compute the centroid for each class by using the pseudo-labelled (true or wrong) samples together and the negative influences caused by the false pseudo-labelled samples are expected to be neutralized by correct pseudo labels. Formally,

$$L_{SM}(X_S,Y_S,X_T) = \underbrace{\sum_{k=1}^{K}\Phi\left(C_S^k, C_T^k\right)}_{L_{SM}(X_S,Y_S,X_T)} \tag{6}$$

where C_S^k, C_T^k are centroid for the k^{th} class in both 2D and 3D feature space, K is the number of categories, $\Phi(\cdot)$ is a distance measure function and L_{SM} represents semantic loss. We employ the squared Euclidean distance in our experiments and get $2K$

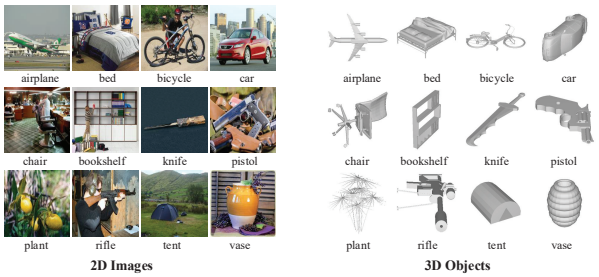


Figure 2: Samples of MI3DOR.

Figure 3: Samples of MI3DOR-2.

centroids. The centroid alignment operation can ensure the features in the same class but different domains mapped nearby by explicitly restricting its distance.

Formally, the total objective loss can be written as follows:

$$L(X_S, Y_S, X_T) = L_C(X_S, Y_S) + \lambda L_{DC}(X_S, X_T) + \gamma L_{SM}(X_S, Y_S, X_T) \quad (7)$$

where λ and γ are hyperparameters to balance the classification loss, the domain confusion loss and the semantic loss.

DLEA attempts to achieve semantic transfer for unsupervised domain adaptation by aligning the centroids in the same class but different domains. We employ the classifier optimized in the source domain to assign the pseudo labels of target samples, which further guide the semantic alignment for visual feature learning. The acquired semantic information can improve the ability of the classifier to recognize the 3D objects in return. This cycle will interactively enhance the performance in the target domain.

4 EXPERIMENTS

4.1 Experimental Settings

4.1.1 Dataset and Evaluation Criteria. To evaluate the proposed method, we employ several popular criteria, including Nearest Neighbor (NN), First Tier (FT), Second Tier (ST), Discounted Cumulative Gain (DCG), F Measure, Average Normalized Modified Retrieval Rank (ANMRR) and Precision-Recall curves (PR) as [19]. For most criteria, the higher value means the better performance other than ANMRR.

We built two datasets [1] for this new task since there is no available public dataset for evaluation.

[1]Both datasets have been used for one international contest. Due to anonymous review, we will add the contest name and data link if the paper can be accepted.

MI3DOR: We selected the 3D objects from the common categories of the popular 3D datasets, including ShapeNetCore55 [1], ModelNet40 [37], PSB [31], NTU [2]. The 2D images from the corresponding categories were collected from ImageNet. MI3DOR contains 21 categories with 7,690 3D objects and 21,000 images. 3,842 objects and 10,500 images are used as the training set. The remaining 3,848 objects and 10,500 images are used for test. The specific number of each category is shown Table. 1 and some samples of 2D images and 3D objects are shown in Fig. 2.

MI3DOR-2: The 3D object set is a subset of ModelNet40 [37]. Based on the categories of the 3D objects, we collected the 2D images from Google. MI3DOR-2 consists of 40 categories with 3,982 objects and 19,694 images. 3,182 objects and 19,294 images are used as the training set while 800 objects and 400 images are used as the test set. The specific number of each category is shown Table. 2 and some samples of 2D real images and 3D objects are shown in Fig. 3.

Comparison: We provide these two datasets in two difficulty levels: 1) Comparing against the 3D objects in MI3DOR-2, the 3D objects of each category in MI3DOR are collected from multiple 3D datasets, which will lead to the huge inner-class difference. Therefore, this will highlight the importance of class-level alignment. 2) Comparing against the 2D images in MI3DOR-2, the 2D images in MI3DOR have more complicated background, which will lead to much difficulty to explore the data distribution. Therefore, this will highlight the importance of domain-level alignment. In conclusion, MI3DOR is more challenging and suitable for large-scale evaluation while MI3DOR-2 with much more categories is suitable for algorithm development and validation.

4.1.2 Implementation Details. In our experiments, we employed the AlexNet as the backbone CNN architecture, which consists of five convolutional layers $conv_1 \sim conv_5$ and three fully connected layers $fc_6 \sim fc_8$. For safer transfer representation learning, we added a bottleneck layer f_b with 256 unites after fc_7 layer. The output of fc_b was used to represent visual feature. For the discriminator D, we utilized the identical architecture with RevGard [6], $x \rightarrow 1024 \rightarrow 1024 \rightarrow 1$. We set the batch size as 128. The entire framework was trained with the rate decay strategy (the origin learning rate μ_0 was 0.01) in an end-to-end manner by SGD (0.9 momentum). Specifically, the learning rate was annealed by $\mu_p = \frac{\mu_0}{(1+\alpha,p)^\beta}$, where $\alpha = 10$, $\beta = 0.75$ and p is training progress varying from 0 to 1 like [6]. For the weight balance parameters λ, γ, we set $\gamma = \lambda$ and $\lambda = \frac{2}{1+exp(-\epsilon,p)} - 1$, where $\epsilon = 10$. It was optimized by [6], which can bring in two advantages in early training phase: 1) suppress the noisy signal from the discriminator; 2) suppress the noisy information brought by false labels. The view number N was tuned and empirically set with 12 when the optimal performance can be achieved.

4.1.3 Competing Methods. For fair comparison on these new datasets, we re-implemented the following representative methods for evaluation: **1) Standard deep learning method.** Since the backbone CNN architecture of DLEA is AlexNet, we utilized AlexNet to extract the visual features of 2D images and multi-view images of 3D objects and directly computed the Euclidean distance in-between for retrieval. This is regarded as the baseline method. **2) Traditional**

Table 1: Statistics of MI3DOR.

Category	2D	3D	Category	2D	3D
airplane	1000	500	monitor	1000	500
bed	1000	500	motorcycle	1000	285
bicycle	1000	62	pistol	1000	245
bookshelf	1000	500	plant	1000	477
camera	1000	90	radio	498	24
car	1000	500	rifle	1000	500
chair	1000	500	stairs	1000	143
flower pot	1000	500	tent	1000	192
guitar	1000	500	vase	1000	500
keyboard	1000	217	wardrobe	1000	500
knife	1000	355	**Total**	**21000**	**7690**

Table 2: Statistics of MI3DOR-2. Tr: Training set; Te: Test set.

Category	2D(Tr/Te)	3D(Tr/Te)	Class	2D(Tr/Te)	3D(Tr/Te)
airplane	492/10	80/20	laptop	416/10	80/20
bathtub	385/10	80/20	mantel	468/10	80/20
bed	356/10	80/20	monitor	641/10	80/20
night stand	392/10	80/20	bench	482/10	80/20
bookshelf	450/10	80/20	person	554/10	79/20
bottle	393/10	80/20	piano	617/10	80/20
bowl	517/10	64/20	plant	435/10	80/20
car	530/10	80/20	radio	498/10	80/20
range hood	510/10	80/20	chair	743/10	80/20
cone	352/10	80/20	sink	632/10	80/20
cup	460/10	79/20	sofa	596/10	80/20
curtain	611/10	80/20	stairs	537/10	80/20
desk	384/10	80/20	stool	454/10	80/20
door	556/10	80/20	table	479/10	80/20
dresser	421/10	80/20	tent	404/10	80/20
flower pot	450/10	80/20	toilet	451/10	80/20
glass box	363/10	80/20	tv stand	377/10	80/20
guitar	332/10	80/20	vase	564/10	80/20
keyboard	486/10	80/20	wardrobe	458/10	80/20
lamp	560/10	80/20	xbox	488/10	80/20

transfer learning methods: a. MEDA [35] can jointly learn the domain-invariant classifier in Grassmann manifold and perform dynamic distribution alignment to analyze the relative importance of marginal and conditional distributions. b. JGSA [39] tends to project the source and target features into common low-dimensional subspace to narrow the geometrical shift and distribution shift. The visual feature representation of 2D image and multiple views is the same with (1). **3) Deep transfer learning methods**: a. JAN [22] employs multiple domain-specific layers across source and target domains to align the joint distributions based on a joint maximum mean discrepancy criterion; b. RevGard [6] employs a discriminator to enforce the embedding alignment in domain level. The relevant experiment settings are identical to DLEA.

Table 3: Performance on MI3DOR.

	NN	FT	ST	F	DCG	ANMRR
AlexNet [14]	0.424	0.323	0.469	0.099	0.345	0.667
MEDA [35]	0.430	0.344	0.501	0.046	0.361	0.646
JGSA [39]	0.612	0.443	0.599	0.116	0.473	0.541
JAN [22]	0.446	0.343	0.495	0.085	0.364	0.647
RevGard [6]	0.650	0.505	0.643	0.112	0.542	0.474
Ours	**0.764**	**0.558**	**0.716**	**0.143**	**0.597**	**0.421**

Table 4: Performance on MI3DOR-2.

	NN	FT	ST	F	DCG	ANMRR
AlexNet [14]	0.518	0.355	0.488	0.355	0.383	0.629
MEDA [35]	0.570	0.392	0.523	0.392	0.425	0.590
JGSA [39]	0.585	0.405	0.533	0.405	0.433	0.577
JAN [22]	0.608	0.501	0.646	0.501	0.527	0.484
RevGard [6]	0.623	0.467	0.614	0.467	0.503	0.514
Ours	**0.700**	**0.555**	**0.681**	**0.555**	**0.593**	**0.424**

4.2 Comparison Against the State of the Arts

We compared DLEA against the representative methods for domain adaptation. As shown in Table.3 & 4, it is obvious that DLEA can achieve the best performances on both datasets. On MI3DOR, DLEA can outperform the others with the gain of 17.5%-80.2%, 10.5%-72.8%, 11.4%-52.7%, 27.7%-44.4%, 10.1%-73.0% in terms of NN, FT, ST, F_measure, DCG, respectively, and achieve the decline of 11.2%-36.9% in terms of ANMRR. On MI3DOR-2, DLEA can outperform the others with the gain of 12.4%-35.1%, 18.8%-56.3%, 10.9%-39.5%, 18.8%-56.3%, 17.9%-54.8%in terms of NN, FT, ST, F_measure, DCG, respectively, and achieve the decline of 17.5%-32.6% in terms of ANMRR. Besides, we observed three key observations:

- DLEA can outperform the representative deep transfer learning methods [6, 22]. Most of deep transfer learning methods can only enforce the alignment of global domain statistics, while ignoring the semantic information of each category. Comparatively, DLEA enforces both domain-level and class-level alignment. The class-level alignment can further benefit domain adaptation.
- DLEA is superior to the representative traditional transfer learning methods [35, 39]. The traditional transfer learning methods usually perform the visual feature learning and the domain alignment separately. Thus, it is impossible to guarantee that the extracted visual features can fit the domain alignment. Comparatively, DLEA can jointly realize the cross-domain feature learning and distribution alignment in an end-to-end manner.
- DLEA outperforms the standard deep learning method [14]. It demonstrates that standard deep networks can not eliminate distribution divergence of two domains although it has been trained on large-scale dataset. As expected, the standard deep learning method works worst.

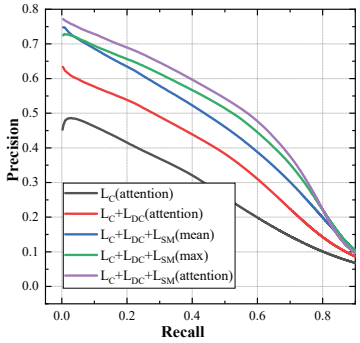


Figure 4: PR curves of different settings on MI3DOR.

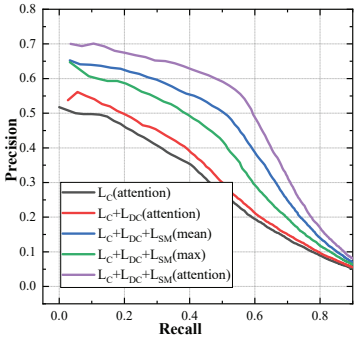


Figure 5: PR curves of different settings on MI3DOR-2.

4.3 Ablation Study

In this section, we evaluate the effect by two key components, multi-view pooling strategy for visual feature learning and different loss functions for cross-domain adaptation.

The Effectiveness of Different Loss Functions. As discussed in Section. 3, the loss function is composed of three parts, the classification loss (L_C), the domain confusion loss ($L_{(DC)}$), and the semantic loss ($L_{(SM)}$). We systematically analyze the effectiveness of different losses.

The Effectiveness of View-Pooling Strategy. For multi-view representation, we need to fuse multiple view features into a compact and discriminative 3D descriptor. This paper propose a cross-domain view-wise attention mechanism to automatically aggregate multiple view features based on their significance for this task. To validate its effectiveness, we compare it against the popular pooling strategy, max pooling and mean pooling.

Table. 5 & 6 show 2D image-based 3D object retrieval results on MI3DOR and MI3DOR-2 by different experimental settings. PR curves are provided in Fig. 4 & 5. We have following observations:

- The network with completed loss ($L_C + L_{DC} + L_{SM}$) can consistently outperform the others ($L_C, L_C + L_{DC}$) since it can enforce the alignment of cross-domain features in dual levels. Specifically, domain confusion loss can guarantee to generate domain-invariant features and semantic loss can further algin distributions from the same class but different domains.
- Aggregating multiple view features with cross-domain attention mechanism can achieve better performance against other pooling strategies (max pooling and mean pooling). The mean and max operations treat individual view features equally while the proposed attention mechanism can adaptively compute the contribution of individual views to narrow the gap between the source and target domains.

4.4 Qualitative Evaluation on Distribution Adaptation

In this section, we perform qualitative evaluation on the distribution adaptation to understand how DLEA enforces dual level embedding alignment. As shown in Fig.6 & 7, we can visualize the representation learned with different losses on MI3DOR and MI3DOR-2 using t-distributed stochastic neighbor embedding (t-SNE) [24]. To guarantee the visualization quality, we randomly sampled 10 classes from the base datasets. Blue points denote 2D images (source domain) and red points denotes 3D objects (target domain). Fig.6 & 7 (a) show the distribution without any adaptation. Fig.6 & 7 (b) show the distribution by adversarial domain adaptation only with domain-level alignment. Fig.6 & 7 (c) show the distribution by DLEA with both domain-level and class-level alignment. We can obtain the following observations:

- The standard deep learning methods without any adaptation can't eliminate the domain shift in two distributions and lead to the significant gap between the source and target domains. Comparatively, the adversarial domain methods can train a domain classifier to distinguish whether the features come from the source or target domain while the feature extractor is trained to fool the discriminator. When this two-player game reaches an equilibrium, we can achieve domain-invariant features. Therefore, the adversarial domain methods can benefit domain adaptation as shown in Fig.6 & 7 (a & b).
- DLEA with dual-level alignment can align features in the same class from different domains while separating features in different classes. Comparatively, the domain adversarial methods ignore the semantic information and might generate ambiguous features near class boundary, which is challenging for retrieval task. Therefore, DLEA can further benefit domain adaptation as shown in Fig.6 & 7 (b & c).

5 CONCLUSION

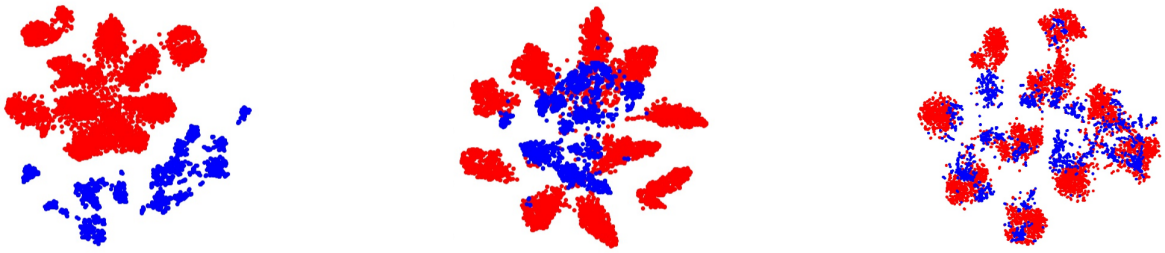
In this paper, we propose an unsupervised dual-level embedding alignment for 2D image-based 3D object retrieval. The proposed method can jointly realize visual feature learning and cross-domain feature adaptation. On one hand, we design a cross-domain view-wise attention mechanism to aggregate the multi-view image set of individual 3D object into a compact 3D object descriptor. One the other hand, cross-domain feature adaptation is designed to align the distributions of cross-domain features in both domain and class levels. Moreover, we built two new datasets, MI3DOR and MI3DOR-2, to evaluate this new task. We compare the proposed method against the state-of-the-art methods. We further explore the influence caused by different loss functions and multi-view pooling strategies. The qualitative analysis on domain adaptation

Table 5: Performances on MI3DOR by different settings.

	NN	FT	ST	F	DCG	ANMRR
L_C(attention)	0.453	0.358	0.499	0.095	0.384	0.628
$L_C + L_{DC}$(attention)	0.634	0.439	0.593	0.116	0.471	0.545
$L_C + L_{DC} + L_{SM}$(mean)	0.717	0.503	0.693	0.139	0.569	0.450
$L_C + L_{DC} + L_{SM}$(max)	0.725	0.542	0.705	0.140	0.578	0.438
$L_C + L_{DC} + L_{SM}$(attention)	**0.764**	**0.558**	**0.716**	**0.143**	**0.597**	**0.421**

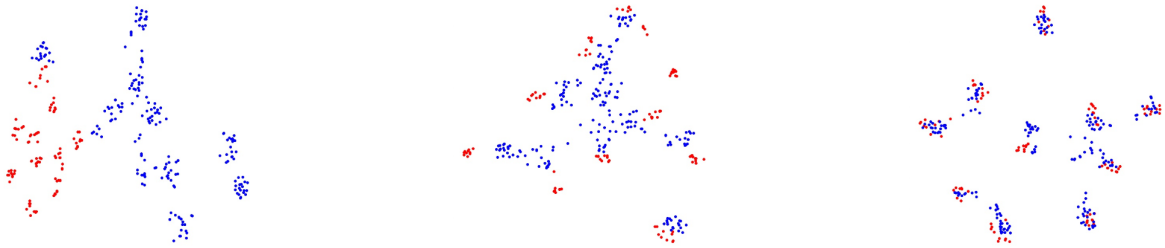
Table 6: Performances on MI3DOR-2 by different settings.

	NN	FT	ST	F	DCG	ANMRR
L_C(attention)	0.518	0.369	0.510	0.369	0.402	0.614
$L_C + L_{DC}$(attention)	0.538	0.395	0.529	0.395	0.434	0.586
$L_C + L_{DC} + L_{SM}$(mean)	0.648	0.456	0.592	0.456	0.491	0.523
$L_C + L_{DC} + L_{SM}$(max)	0.652	0.503	0.643	0.503	0.539	0.478
$L_C + L_{DC} + L_{SM}$(attention)	**0.700**	**0.555**	**0.681**	**0.555**	**0.593**	**0.424**



(a) Non-adapted (b) Adversarial Adapted (c) Semantic Adapted

Figure 6: Domain adaptation by different methods. Blue points are source samples (2D image) and red points are target samples (3D object).



(a) Non-adapted (b) Adversarial Adapted (c) Semantic Adapted

Figure 7: Domain adaptation by different methods on MI3DOR-2. Blue points are source samples (2D image) and red points are target samples (3D object).

can intuitively illustrate how DLEA realizes dual-level embedding alignment between the source and target domains. Extensive comparisons demonstrated the superiority of the proposed method.

6 ACKNOWLEDGEMENT

This work was supported in part by the National Natural Science Foundation of China (61772359, 61572356, 61872267), the grant of 2019 Tianjin New Generation Artificial Intelligence Major Program, the grant of 2018 Tianjin New Generation Artificial Intelligence Major Program (18ZXZNGX00150), the Open Project Program of the State Key Lab of CAD & CG (Grant No.A1907), Zhejiang University, the grant of Elite Scholar Program of Tianjin University (2019XRX-0035).

REFERENCES

[1] Angel X. Chang, Thomas A. Funkhouser, Leonidas J. Guibas, Pat Hanrahan, Qi-Xing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. 2015. ShapeNet: An Information-Rich 3D Model Repository. *CoRR* abs/1512.03012 (2015).

[2] Ding-Yun Chen, Xiao-Pei Tian, Yu-Te Shen, and Ming Ouhyoung. 2003. On Visual Similarity Based 3D Model Retrieval. *Comput. Graph. Forum* 22, 3 (2003), 223–232.

[3] Zhiyong Cheng, Xiaojun Chang, Lei Zhu, Rose Catherine Kanjirathinkal, and Mohan S. Kankanhalli. 2019. MMALFM: Explainable Recommendation by Leveraging Reviews and Images. *ACM Trans. Inf. Syst.* 37, 2 (2019), 16:1–16:28.

[4] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. 2017. Joint distribution optimal transportation for domain adaptation. In *Advances in Neural Information Processing Systems*. 3733–3742.

[5] Yifan Feng, Zizhao Zhang, Xibin Zhao, Rongrong Ji, and Yue Gao. 2018. GVCNN: Group-View Convolutional Neural Networks for 3D Shape Recognition. In *2018 IEEE Conference on Computer Vision and Pattern Recognition*. 264–272.

[6] Yaroslav Ganin and Victor S. Lempitsky. 2015. Unsupervised Domain Adaptation by Backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*. 1180–1189.

[7] Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. 2016. Domain Adaptation with Conditional Transferable Components. In *Proceedings of the 33nd International Conference on Machine Learning*. 2839–2848.

[8] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*. 2672–2680.

[9] Zhizhong Han, Mingyang Shang, Zhenbao Liu, Chi-Man Vong, Yu-Shen Liu, Matthias Zwicker, Junwei Han, and C. L. Philip Chen. 2019. SeqViews2SeqLabels: Learning 3D Global Features via Aggregating Sequential Views by RNN With Attention. *IEEE Trans. Image Processing* 28, 2 (2019), 658–672.

[10] Xinwei He, Yang Zhou, Zhichao Zhou, Song Bai, and Xiang Bai. 2018. Triplet-Center Loss for Multi-View 3D Object Retrieval. In *2018 IEEE Conference on Computer Vision and Pattern Recognition*. 1945–1954.

[11] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. 2018. CyCADA: Cycle-Consistent Adversarial Domain Adaptation. In *Proceedings of the 35th International Conference on Machine Learning*. 1994–2003.

[12] Weixiang Hong, Zhenzhen Wang, Ming Yang, and Junsong Yuan. 2018. Conditional Generative Adversarial Network for Structured Domain Adaptation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition*. 1335–1344.

[13] Asako Kanezaki, Yasuyuki Matsushita, and Yoshifumi Nishida. 2018. RotationNet: Joint Object Categorization and Pose Estimation Using Multiviews From Unsupervised Viewpoints. In *2018 IEEE Conference on Computer Vision and Pattern Recognition*. 5010–5019.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*. 1106–1114.

[15] Tang Lee, Yen-Liang Lin, HungYueh Chiang, Ming-Wei Chiu, Winston Hsu, and Polly Huang. 2018. Cross-Domain Image-Based 3D Shape Retrieval by View Sequence Learning. In *2018 International Conference on 3D Vision*. 258–266.

[16] Jiaxin Li, Ben M. Chen, and Gim Hee Lee. 2018. SO-Net: Self-Organizing Network for Point Cloud Analysis. In *2018 IEEE Conference on Computer Vision and Pattern Recognition*. 9397–9406.

[17] An-An Liu, Weizhi Nie, Yue Gao, and Yuting Su. 2018. View-Based 3-D Model Retrieval: A Benchmark. *IEEE Trans. Cybernetics* 48, 3 (2018), 916–928.

[18] Anan Liu, Shu Xiang, Wenhui Li, Weizhi Nie, and Yuting Su. 2018. Cross-Domain 3D Model Retrieval via Visual Domain Adaption. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. 828–834.

[19] A. A. Liu, W. Z. Nie, Y. Gao, and Y. T. Su. 2017. View-Based 3-D Model Retrieval: A Benchmark. *IEEE Trans. Cybernetics* 48, 3 (2017), 916–928.

[20] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. 2018. Conditional Adversarial Domain Adaptation. In *Advances in Neural Information Processing Systems*. 1647–1657.

[21] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S. Yu. 2013. Transfer Feature Learning with Joint Distribution Adaptation. In *IEEE International Conference on Computer Vision*. 2200–2207.

[22] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. 2017. Deep Transfer Learning with Joint Adaptation Networks. In *Proceedings of the 34th International Conference on Machine Learning*. 2208–2217.

[23] Zelun Luo, Yuliang Zou, Judy Hoffman, and Fei-Fei Li. 2017. Label Efficient Learning of Transferable Representations acrosss Domains and Tasks. In *Advances in Neural Information Processing Systems*. 164–176.

[24] Laurens Van Der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Commun. ACM* 9 (2008), 2579–2605.

[25] Daniel Maturana and Sebastian Scherer. 2015. VoxNet: A 3D Convolutional Neural Network for real-time object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 922–928.

[26] Saeid Motiian, Quinn Jones, Seyed Mehdi Iranmanesh, and Gianfranco Doretto. 2017. Few-Shot Adversarial Domain Adaptation. In *Advances in Neural Information Processing Systems*. 6673–6683.

[27] Sinno Jialin Pan, Ivor W. Tsang, James T. Kwok, and Qiang Yang. 2011. Domain Adaptation via Transfer Component Analysis. *IEEE Trans. Neural Networks* 22, 2 (2011), 199–210.

[28] Bui Tuong Phong. 1975. Illumination for Computer Generated Pictures. *Commun. ACM* 18, 6 (1975), 311–317.

[29] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. 2017. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*. 77–85.

[30] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. 2018. Wasserstein Distance Guided Representation Learning for Domain Adaptation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*. 4058–4065.

[31] Philip Shilane, Patrick Min, Michael M. Kazhdan, and Thomas A. Funkhouser. 2004. The Princeton Shape Benchmark. In *2004 International Conference on Shape Modeling and Applications*. 167–178.

[32] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik G. Learned-Miller. 2015. Multi-view Convolutional Neural Networks for 3D Shape Recognition. In *2015 IEEE International Conference on Computer Vision*. 945–953.

[33] Yuting Su, Wenhui Li, Anan Liu, and Weizhi Nie. 2018. Hierarchical Graph Structure Learning for Multi-View 3D Model Retrieval. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. 913–919.

[34] Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. 2007. Direct Importance Estimation with Model Selection and Its Application to Covariate Shift Adaptation. In *Advances in Neural Information Processing Systems*. 1433–1440.

[35] Jindong Wang, Wenjie Feng, Yiqiang Chen, Han Yu, Meiyu Huang, and Philip S. Yu. 2018. Visual Domain Adaptation with Manifold Embedded Distribution Alignment. In *2018 ACM Multimedia Conference on Multimedia Conference*. 402–410.

[36] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. 2016. Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling. In *Advances in Neural Information Processing Systems*. 82–90.

[37] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 2015. 3D ShapeNets: A deep representation for volumetric shapes. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1912–1920.

[38] Haoxuan You, Yifan Feng, Rongrong Ji, and Yue Gao. 2018. PVNet: A Joint Convolutional Network of Point Cloud and Multi-View for 3D Shape Recognition. In *2018 ACM Multimedia Conference on Multimedia Conference*. 1310–1318.

[39] Jing Zhang, Wanqing Li, and Philip Ogunbona. 2017. Joint Geometrical and Statistical Alignment for Visual Domain Adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*. 5150–5158.

[40] Lei Zhu, Zi Huang, Xiaojun Chang, Jingkuan Song, and Heng Tao Shen. 2017. Exploring Consistent Preferences: Discrete Hashing with Pair-Exemplar for Scalable Landmark Search. In *Proceedings of the 2017 ACM on Multimedia Conference (MM)*. 726–734.