J. Vis. Commun. Image R. 83 (2022) 103426

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Full length article

Collaborative Distribution Alignment for 2D image-based 3D shape retrieval☆

Nian Hu [a,b], Heyu Zhou [a], An-An Liu [a,b,*], Xiangdong Huang [a], Shenyuan Zhang [c], Guoqing Jin [d], Junbo Guo [d], Xuanya Li [e,**]

[a] School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China
[b] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230088, China
[c] People's Daily, Beijing 100733, China
[d] State Key Laboratory of Communication Content Cognition, People's Daily Online, Beijing 100733, China
[e] Baidu Inc., Beijing 100105, China

ARTICLE INFO

MSC:
41A05
41A10
65D05
65D17

Keywords:
3D shape retrieval
Cross-domain retrieval
Multi-view learning

ABSTRACT

Retrieving 3D shapes with 2D images has become a popular research area nowadays, and a great deal of work has been devoted to reducing the discrepancy between 3D shapes and 2D images to improve retrieval performance. However, most approaches ignore the semantic information and decision boundaries of the two domains, and cannot achieve both domain alignment and category alignment in one module. In this paper, a novel Collaborative Distribution Alignment (CDA) model is developed to address the above existing challenges. Specifically, we first adopt a dual-stream CNN, following a similarity guided constraint module, to generate discriminative embeddings for input 2D images and 3D shapes (described as multiple views). Subsequently, we explicitly introduce a joint domain-class alignment module to dynamically learn a class-discriminative and domain-agnostic feature space, which can narrow the distance between 2D image and 3D shape instances of the same underlying category, while pushing apart the instances from different categories. Furthermore, we apply a decision boundary refinement module to avoid generating class-ambiguity embeddings by dynamically adjusting inconsistencies between two discriminators. Extensive experiments and evaluations on two challenging benchmarks, MI3DOR and MI3DOR-2, demonstrate the superiority of the proposed CDA method for 2D image-based 3D shape retrieval task.

1. Introduction

3D shape retrieval has drawn increasing attention due to its wide applications in 3D digital manufacturing, 3D printing, virtual reality and medical imaging [1,2]. Although great achievements have been made in the homogeneous 3D shape retrieval community, the following problems still exist. Most methods are trained in a supervised manner based on manually labeled 3D shape datasets [3,4], which are expensive to collect in practice. Meanwhile, these methods are limited in the case where only unlabeled 3D shape data is available. Compared to methods that use 3D shapes as queries, the 2D image-based 3D shape retrieval is extremely attractive since 2D images provide a flexible and user-friendly way for input, users can take pictures of the objects they are interested in through mobile phones or cameras, or search images directly from the Internet. Despite the advantages of 2D image-based 3D shape retrieval, it is a quite challenge due to the huge distribution bias between two domains. We may not train an out-of-the-box deep network since we need to address the following issues.

How to improve the intra-class compactness and inter-class separability of each domain? As shown in Fig. 1, 2D images from the same category usually have distinct disparities, resulting in significant intra-class variations, and this may create difficulties in feature learning for saliency. As for 3D shapes, due to the lack of ground truth labels, the learned representations may not be optimally discriminative, which in turn raises the possibility that different categories of representations are mixed together in the deep feature space. Therefore, it is necessary to take steps to minimize the intra-class differences and maximize the inter-class discrepancy in the feature learning process. By doing so, we can preserve the semantic structure for both domains and enhance the discrimination of feature representations.

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author at: School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China.
** Corresponding author.
E-mail addresses: anan0422@gmail.com (A.-A. Liu), lixuanya@baidu.com (X. Li).

https://doi.org/10.1016/j.jvcir.2021.103426
Received 23 May 2021; Received in revised form 22 September 2021; Accepted 22 December 2021
Available online 22 January 2022

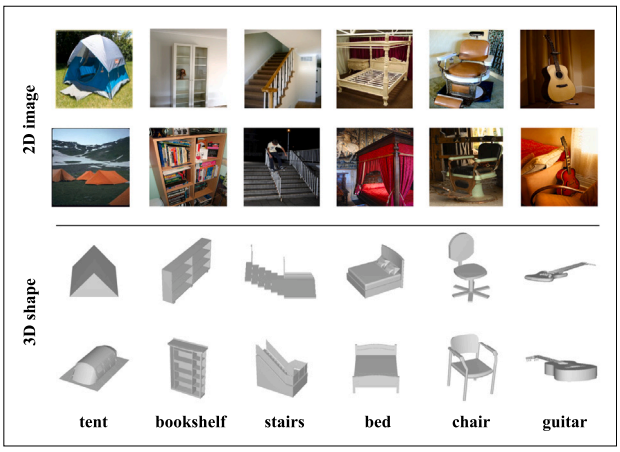


Fig. 1. Examples of 2D images and 3D shapes on MI3DOR.

How to achieve cross-domain distribution alignment and consequently improve retrieval accuracy? Fig. 1 presents a few samples of 2D images and their corresponding 3D shapes. We can immediately observe the huge differences between the two domains in terms of background, illumination, viewpoint, etc., which leads to nonnegligible cross-domain discrepancy. To alleviate this discrepancy, a variety of approaches have been proposed to align the feature distribution across domains [5–9]. One of the classic approaches is adversarial-based methods [9], which jointly train a feature extractor and a domain discriminator in an adversarial training paradigm. Unfortunately, this practice does not guarantee a successful transfer, as it can only guarantee the global domain alignment, not the local class alignment. Furthermore, another strategy is reducing the distance between two distributions, *i.e.*, moments or maximum mean discrepancy [6,8]. However, these distance-based approaches may fail to extract domain-invariant features because they do not consider the relationship between feature representations and category decision boundaries when aligning distributions. These strategies focus on reducing the distance between domains and tend to ignore class decision boundaries, thereby producing features that are close to the decision boundaries and are not discriminative enough to be categorized into specific class. Therefore, aligning different domain distributions while considering the relationship between feature representations and category decision boundaries is the challenge we must address.

To tackle the above problems, we propose a novel model, namely Collaborative Distribution Alignment (CDA) to diminish the domain discrepancy for 2D image-based 3D shape retrieval. Fig. 2 shows the framework of our proposed model. Directly matching 3D shapes to 2D images suffers great discrepancy between 2D and 3D representations. Therefore, we project the 3D shapes to multiple 2D rendered views to further bridge the domain gap. In the training process, we first utilize a dual-stream CNNs to generate embeddings of both input 2D images and 3D shapes (multi-view). In addition to minimizing the classification error of labeled 2D images, we construct a similarity guided constraint to preserve the semantic structure for both domains, this constraint is able to explicitly model the intra-class compactness and the inter-class separability via triplet loss optimization. Subsequently, in order to diminish the huge cross-domain divergence between learned embeddings of 2D images and 3D shapes, we develop a joint domain-class alignment network, to perform both domain-aware and class-aware alignments with a 2N-way adversarial loss. Here 2N represents the total number of categories in 2D image and 3D shape domain. The first N represents the 2D image categories, the second N represents the 3D shape categories. There are two 2N-way discriminators and a feature extractor in our work. The 2N-way discriminator is trained to simultaneously identify the domain and class label of embeddings, while the feature extractor is trained to confuse it [10]. Compared to previous methods design a binary or N-way discriminator [9,11,12], our 2N-way discriminator can better produce a domain-agnostic and category-discriminative feature representation. Although we have tried to align the joint distribution of domain and class, the learned decision boundaries may generalize poorly on the 3D shape domain. Since we have not taken class boundaries into account during the adaptation process, there may exist many suboptimal solutions near the decision boundaries. Therefore, we adopt a novel mechanism to train two discriminators in an adversarial manner, we first maximize the disagreement between two discriminators' outputs to detect 3D shape features that are close to the decision boundaries. Then we update the feature extractor to generate 3D shape features far from the decision boundaries by minimizing the disagreement. By doing so, the class-ambiguous 3D shape embeddings are converted to more discriminative ones.

To summarize, the key technical contributions of this work are three-fold:

- We apply the similarity guided constraint module to 2D image and 3D shape embeddings to explicitly model the intra-class compactness and the inter-class separability.
- We introduce a novel adversarial learning-based framework for 2D image-based 3D shape retrieval. The framework realizes both global domain alignment and local category alignment in one module, which greatly alleviates the problem of cross-domain feature mismatch.
- Our method achieves competitive performance compared to the state-of-the-art on MI3DOR and MI3DOR-2 benchmarks.

The rest of this paper is organized as follows. Related work on 3D shape retrieval and domain adaptation is described in Section 2, and the proposed method is presented in Section 3. Section 4 illustrates the experimental results and corresponding analysis, and Section 5 concludes the paper.

2. Related work

The work done in this paper is related to previous work in 3D shape retrieval, which has been extensively studied for decades. This work also deals with domain adaptation, which is a very insightful and promising research on transferring the knowledge from labeled data to unlabeled data. Below, we briefly review the related works.

2.1. 3D shape retrieval

The existing 3D shape retrieval approaches can be broadly grouped into two categories: model-based approaches [13–15] and view-based approaches [3,16,17]. Model-based approaches typically describe 3D shapes as polygon meshes [13], point clouds [14], and voxels [15], *etc.* For instance, Bronstein et al. [13] extended the CNN operations to the weighted undirected graphs defined by irregular triangle meshes. Qi et al. [14] presented an outstanding work PointNet, which takes full advantage of the symmetric function to learn 3D features on unordered point cloud. Wang et al. [15] constructed an octree of the input 3D shapes and packed the required information into it. With the help of this octree structure, all 3D CNN operations can be efficiently executed on the GPU.

Compared to model-based approaches that work directly on 3D data, view-based approaches generally convert 3D shapes into 2D projections, and utilize established 2D CNNs to process these rendered views to obtain multi-view features [3,4,16,18]. By mimicking the human visual perception of 3D shapes with 2D observations, the view-based methods have obtained very competitive performance on the 3D shape retrieval task. Among them, Su et al. [3] proposed the landmark work, Multi-view Convolutional Neural Networks, which designs a

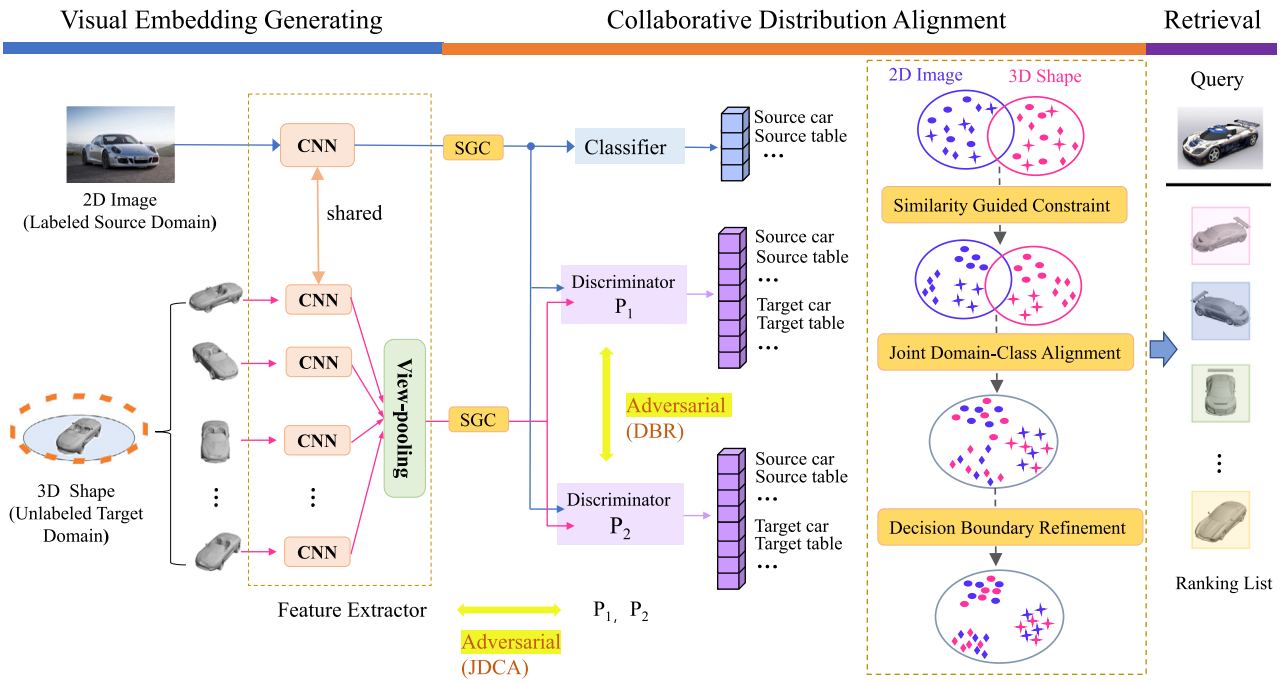


Fig. 2. The pipeline of our model for 2D image-based 3D shape retrieval. We first utilize a feature extractor to produce embeddings of both input 2D images and 3D shapes (described as multiple views). Specifically, the embeddings of 2D images are marked in blue, while the embeddings of 3D shapes are marked in pink. Then we apply a similarity guided constraint (SGC) on both embeddings to compact the semantically similar instances and separate the semantically dissimilar instances. In the joint domain-class alignment module, all embeddings are shared by two distinct 2N-way discriminators P_1, P_2. The 2N-way discriminator is trained to simultaneously identify the domain and class label of the embeddings, while the feature extractor is trained to fool it to achieve both domain and class alignment. Furthermore, we train two discriminators in an adversarial manner to push the class-ambiguity embeddings far away from the decision boundaries in the decision boundary refinement module. Finally, we adopt the L2-distance as the metric to measure the visual distance of 2D image embeddings and 3D shape embeddings to complete the retrieval task.

multi-view CNN operating on multiple 2D projections that rendered from different angles of the 3D shape, and employs a view-wise pooling strategy to assemble multi-view features into a unified and compact 3D shape descriptor. This work was later extended to many other networks to further explore the associations between rendered views [4,16,18]. Nie et al. [17] adopted panoramic views to represent each 3D shape, in their work, an innovative multi-scale CNN is used to generate visual features for each panoramic view. Then a new multi-channel network is presented to generate a representation of each 3D shape by combining panoramic features from different axes.

All the above 3D shape retrieval methods are designed on closed dataset, while the emerging cross-domain 3D shape retrieval is currently a more challenging task with a wider range of applications. Recently, two benchmarks, SHREC14'SBR and SHREC19'MI3DOR, have been proposed to evaluate sketch-based 3D shape retrieval methods and 2D image-based 3D shape retrieval methods, respectively. For the former, Chen et al. introduced a cross-modal transformation network to compensate for the discrepancy between the two modalities of sketch and 3D shape [1]. For the latter, Lee et al. proposed a cross-domain triplet network to bridge the gap between 2D images and 3D shapes, which contains an adaptation layer to better align features from different domains [19]. However, these methods target supervised cross-domain 3D shape retrieval, and we are unable to label all 3D shapes in practice. Therefore, it is indispensable and essential to mitigate the cross-domain discrepancy in the absence of 3D shape labels.

2.2. Domain adaptation

2D image-based 3D shape retrieval is closely related to domain adaptation. Early methods for domain adaptation are based on shallow learning [20–25], where the feature representation is fixed during training, and then a geometric transformation is applied to the features to match the source and target domains. That is, the feature learning and domain adaptation are not trained together. These methods usually perform poorly on large datasets. With the advancement of deep learning, a growing number of researchers tend to focus on deep domain adaptation.

Deep domain adaptation work combines domain adaptation with deep learning, and several losses have been designed to tackle the domain shift issue, such as discrepancy loss [26–29] and adversarial loss [30–33]. Among them, discrepancy-based approaches explicitly measure the differences between source and target domains, such as multi-kernel variants of maximum mean discrepancy [26], moment matching [27,28] and correlation alignment [29], *etc*. Adversarial-based methods generally combine domain classifier with feature extractor to produce spurious source or target data to obfuscate discriminator, typically based on GANs [34] and their variants such as CycleGAN [30], CoGAN [31,32] and SimGAN [33]. Although in recent years, with the rapid development of computer vision technology, domain adaptation has made great progress, but domain adaptation can not be directly used in 2D image-based 3D shape retrieval tasks, because (1) they cannot process 3D data and (2) these methods do not take into account the setting of the retrieval task.

3. The proposed algorithm

3.1. Overview

Given a labeled 2D image domain (source domain) $D_s = \{x_s^i, y_s^i\}_{i=1}^{n_s}$ and an unlabeled 3D shape domain (target domain) $D_t = \{x_t^i\}_{i=1}^{n_t}$. The goal of our model is to utilize labeled 2D images and unlabeled 3D shapes to jointly decrease the cross-domain divergence for 2D image-based 3D shape retrieval. Note that the intrinsic distribution $\mathbb{P}$ and $\mathbb{Q}$ corresponding to 2D image domain and 3D shape domain are different in the shared label space.

As depicted in Fig. 2, our model contains three major modules: visual embedding generating, collaborative distribution alignment and

retrieval. The visual embedding generating module aims to generate visual features for both domains. After that, we fed those feature representations into the collaborative distribution alignment module. There are three sub-modules in it: similarity guided constraint, joint domain-class alignment and decision boundary refinement. The first sub-module is designed to preserve the semantic information of both domains by decreasing the intra-class distance and increasing the inter-class distance, and to further provide more reliable annotations for 3D shape instances. The second sub-module is designed to learn a class-discriminative and domain-agnostic feature space, whereas the third sub-module is designed to push the embeddings far away from the decision boundaries to further facilitate the generation of discriminative embeddings. In the retrieval step, we take the outputs of visual embedding generating module on the test set as the feature representations, then measure the visual distance between them by L2-distance, and finally rank the visual distance to complete the retrieval. We will detail the proposed architecture in the following subsections.

3.2. Visual embedding generating

As shown in Fig. 2, the first step of our method is to obtain the visual features of 2D images and 3D shapes. In order to alleviate the modality divergence between 2D images and 3D shapes, we utilize the multi-view representation to characterize 3D shapes. Specifically, the Phong reflection model [35] is employed as the rendering engine to generate 2D projections of 3D shapes. Following [3], we place 12 virtual cameras at intervals of 30° evenly around the 3D shape, all pointing to the centroid of 3D shape, thus generating 12 different 2D projections. We adopt a dual-stream CNNs architecture to manipulate the 2D images and the 2D projections separately. It is worth noting that both streams share the same parameters. After visual embedding generating, we get 12 view features for each 3D shape. Then, we merge them into a unified and compact 3D shape descriptor by max-pooling operation [3].

3.3. Collaborative distribution alignment

3.3.1. Similarity guided constraint

Given a 2D image x_s, the feature extractor $\Phi(\cdot)$ outputs a feature map $\Phi(x_s)$, which is fed into the classifier $C_\theta(\cdot)$ to generate a category prediction:

$$\mathbf{f} = C_\theta(\Phi(x_s)) \in \mathbb{R}^N \tag{1}$$

where N represents the total number of classes in 2D image domain, and the number of 3D shape classes is the same as that of 2D image classes. The feature extractor $\Phi(\cdot)$ and the source-supervised classifier $C_\theta(\cdot)$ are trained together on the labeled 2D image domain to minimize the following classification objective:

$$\mathcal{L}_{cls} = E_{(x,y)\sim\mathbb{P}}\, \ell_{CE}\left(C_\theta(\Phi(x)), y\right) \tag{2}$$

where $\ell_{CE}(\cdot,\cdot)$ is the cross-entropy loss.

For the 3D shape data, their ground-truth labels are absent in the training process, thus we adopt the source-supervised classifier $C_\theta(\cdot)$ to assign pseudo labels to them:

$$n = arg\,max_n\; C_\theta(\Phi(x_t))[n] \tag{3}$$

In order to ensure the accuracy and reliability of pseudo-labeling, we re-assign pseudo labels to the 3D shape data every few iterations in the training process. With the assistance of the alignment module, the source-supervised classifier will progressively improve its accuracy on the 3D shape instances. In this way, our model allows for a relatively accurate estimation of the underlying label hypothesis of 3D shape instances.

After that, in order to make the learned representations intra-class compact and inter-class separable, we impose similarity guided constraint (SGC) on the visual feature representation learning for 2D images and 3D shapes, respectively, with SGC takes the form of weighted regularization triplet loss [36]:

$$\mathcal{L}_{sgc} = \log\left(1 + \exp\left(\sum_p w_{a,p} D_{a,p} - \sum_n w_{a,n} D_{a,n}\right)\right) \tag{4}$$

$$w_{a,p} = \frac{\exp\left(D_{a,p}\right)}{\sum_{D_{a,p}\in\mathcal{P}_a}\exp\left(D_{a,p}\right)}, w_{a,n} = \frac{\exp\left(-D_{a,n}\right)}{\sum_{D_{a,n}\in\mathcal{N}_a}\exp\left(-D_{a,n}\right)} \tag{5}$$

where (a, p, n) represents a hard triplet within each training batch. For anchor a, $\mathcal{P}_a$ is the corresponding positive set which has the same category as a, and $\mathcal{N}_a$ is the negative set which category is different from a. $D_{a,p}$ represents the squared Euclidean distance of a positive pair and $D_{a,n}$ represents the squared Euclidean distance of a negative pair. The formula for calculating the squared Euclidean distance is as follows:

$$D_{i,j} = \left\|\Phi\left(x_i\right) - \Phi\left(x_j\right)\right\|_2^2 \tag{6}$$

By minimizing the weighted regularization triplet loss for each domain separately, SGC compacts the semantically similar instances and separates the semantically dissimilar ones. This constraint helps to mitigate the semantic misalignment phenomenon and enhance the discrimination of our model. Furthermore, the weighted regularization not only inherits the merit of relative distance optimization between positive and negative pairs, but also avoids introducing any additional margin parameters or hyper-parameters in the training process.

3.3.2. Joint domain-class alignment

To reduce the cross-domain divergence, previous domain adaptation approaches typically impose a binary adversarial loss on the discriminator to identify which domain the feature comes from [9,11], in fact, such methods only consider the cross-domain global distribution alignment, but ignore the conditional (class) distribution alignment. As a result, instances of different categories may be aligned in the embedding space incorrectly, which in turn leads to suboptimal performance. To resolve this problem, we design a 2N-way adversarial loss on the discriminators to further improve the transferable ability of the model. The first N represents the 2D image categories, and the second N represents the 3D shape categories. Such an architecture allows learning the common variables of both domain and class, thus enabling dual alignment within a single component. Up to this point, for an input x, the output of two distinct discriminators P_1, P_2 can be represented separately as:

$$\begin{aligned}\mathbf{d}_1 &= P_1(\Phi(x)) \in \mathbb{R}^{2N}\\ \mathbf{d}_2 &= P_2(\Phi(x)) \in \mathbb{R}^{2N}\end{aligned} \tag{7}$$

For the 2D image samples, we train the two discriminators P_1 and P_2 with the same prediction loss as Eq. (2):

$$\begin{aligned}\mathcal{L}_s(P_1) &= E_{(x,y)\sim\mathbb{P}}\, \ell_{CE}\left(P_1(\Phi(x)), [\mathbf{y}, \mathbf{0}]\right)\\ \mathcal{L}_s(P_2) &= E_{(x,y)\sim\mathbb{P}}\, \ell_{CE}\left(P_2(\Phi(x)), [\mathbf{y}, \mathbf{0}]\right)\end{aligned} \tag{8}$$

Where $\mathbf{0}$ is the zero vector of length N, which is utilized to let the last N probability values be zero for the labeled 2D image instances. This time, only two discriminators are updated, while the feature extractor $\Phi(x)$ is fixed.

Although the 2D images and 3D shapes differ significantly at visual level, they overlap at semantic level. Therefore, we can impose the same prediction loss on the unlabeled 3D shapes.

$$\begin{aligned}\mathcal{L}_t(P_1) &= E_{x\sim\mathbb{Q}}\, \ell_{CE}\left(P_1(\Phi(x)), [\mathbf{0}, \hat{\mathbf{y}}]\right)\\ \mathcal{L}_t(P_2) &= E_{x\sim\mathbb{Q}}\, \ell_{CE}\left(P_2(\Phi(x)), [\mathbf{0}, \hat{\mathbf{y}}]\right)\end{aligned} \tag{9}$$

Where $\hat{\mathbf{y}} = \mathbf{e}_n$, $n = arg\,max_n\; C_\theta(\Phi(x_t))[n]$. $\mathbf{e}_n$ is an identity of size N whose nth element is 1.

For the labeled 2D image samples, we train the feature extractor to fool the 2N-way discriminator by altering the label from $[\mathbf{y}, \mathbf{0}]$ to $[\mathbf{0}, \mathbf{y}]$:

$$\begin{aligned}\mathcal{L}_s(\boldsymbol{\Phi}) = & E_{(x,y)\sim\mathbb{P}}\,\ell_{CE}\left(P_1(\boldsymbol{\Phi}(x)), [\mathbf{0}, \mathbf{y}]\right) \\ & + E_{(x,y)\sim\mathbb{P}}\,\ell_{CE}\left(P_2(\boldsymbol{\Phi}(x)), [\mathbf{0}, \mathbf{y}]\right)\end{aligned} \tag{10}$$

For the 3D shape samples, we alter the pseudo label from $[\mathbf{0}, \hat{\mathbf{y}}]$ to $[\hat{\mathbf{y}}, \mathbf{0}]$:

$$\begin{aligned}\mathcal{L}_t(\boldsymbol{\Phi}) = & E_{x\sim\mathbb{Q}}\,\ell_{CE}\left(P_1(\boldsymbol{\Phi}(x)), [\hat{\mathbf{y}}, \mathbf{0}]\right) \\ & + E_{x\sim\mathbb{Q}}\,\ell_{CE}\left(P_2(\boldsymbol{\Phi}(x)), [\hat{\mathbf{y}}, \mathbf{0}]\right)\end{aligned} \tag{11}$$

Note that when minimizing $\mathcal{L}_s(\boldsymbol{\Phi})$ and $\mathcal{L}_t(\boldsymbol{\Phi})$, only the feature extractor $\boldsymbol{\Phi}(\cdot)$ needs to be updated.

3.3.3. Decision boundary refinement

Although we have attempted to align the joint distribution of domain and class, the learned decision boundaries are not guaranteed to generalize well over the 3D shape domain, since the feature extractor may produce semantically ambiguous features close to the decision boundaries. To avoid that, we enforce the two discriminators to be diverse in the learned parameters via consensus regularization [37,38]. Specifically, we maximize the prediction disagreement between two discriminators to detect the class-ambiguity 3D shape features:

$$\max_{P_1, P_2} \mathcal{L}_d \tag{12}$$

$$\mathcal{L}_d = E_{x\sim\mathbb{Q}} \frac{1}{N} \sum_{n=1}^{N} \left|\mathbf{d}_1[n] - \mathbf{d}_2[n]\right| \tag{13}$$

where the $\mathbf{d}_1[n]$ and $\mathbf{d}_2[n]$ represent probability outputs of $\mathbf{d}_1$ and $\mathbf{d}_2$ for category n respectively. Then, we minimize the disagreement by training the feature extractor to produce features far from the decision boundaries:

$$\min_{\boldsymbol{\Phi}} \mathcal{L}_d \tag{14}$$

After that, the class-ambiguity features are retrained to become more discriminative ones.

Overall Objective

We combine the above sub-objectives and divide our training process into three steps:

Step 1 We train the feature extractor $\boldsymbol{\Phi}(\cdot)$ to minimize the weighted regularization triplet loss, and we minimize the 2D image classification loss of the source-supervised classifier $C_\theta(\cdot)$ and discriminators P_1, P_2.

$$\min_{\boldsymbol{\Phi}, C_\theta, P_1, P_2} \mathcal{L}_{cls} + \lambda_{sgc}\mathcal{L}_{sgc} + \lambda_s\mathcal{L}_s(P_1) + \lambda_s\mathcal{L}_s(P_2) \tag{15}$$

Step 2 We fix the feature extractor, and update the source-supervised classifier and discriminators. We train the model by minimizing the 2D image and 3D shape classification loss of the discriminators, and the 2D image classification loss of the source-supervised classifier. In the meantime, we maximize the disagreement between discriminators.

$$\min_{C_\theta, P_1, P_2} \mathcal{L}_{cls} + \lambda_s\mathcal{L}_s(P_1) + \lambda_s\mathcal{L}_s(P_2) + \lambda_t\mathcal{L}_t(P_1) + \lambda_t\mathcal{L}_t(P_2) - \lambda_d\mathcal{L}_d \tag{16}$$

Step 3 We fix the source-supervised classifier and discriminators, and update feature extractor. We minimize the 2D image and 3D shape alignment loss of the discriminators, the weighted regularization triplet loss, and the disagreement between discriminators.

$$\min_{\boldsymbol{\Phi}} \lambda_{sa}\mathcal{L}_s(\boldsymbol{\Phi}) + \lambda_{ta}\mathcal{L}_t(\boldsymbol{\Phi}) + \lambda_{sgc}\mathcal{L}_{sgc} + \lambda_d\mathcal{L}_d \tag{17}$$

In our work, Step2 and Step3 are iteratively optimized.

During testing, we can obtain the 2D image and 3D shape feature representations via same architecture. Then the visual distance between them are ranked to retrieval the corresponding 3D shapes based on the query 2D image.

4. Experiment

4.1. Experimental settings

4.1.1. Datasets

We choose two widely adopted datasets, *i.e.,* MI3DOR [5] and MI3DOR-2 [5], for experimental evaluations.

- The MI3DOR includes a total of 21,000 2D images and 7,690 3D shapes from 21 categories. The dataset is partitioned into a training set and a test set. Specifically, 10,500 2D images and 3,842 3D shapes are used for training and the remaining 10,500 2D images and 3,848 3D shapes are used for testing.
- The MI3DOR-2 consists of 19,694 2D images and 3,982 3D shapes from 40 categories. 19,294 2D images and 3,182 3D shapes are used for training while the remaining 400 2D images and 800 3D shapes are used for testing.

4.1.2. Evaluation criteria

In our experiment, we measure the retrieval performance using the following criteria [39]:

(1) Nearest neighbor (NN), which measures the top 1 retrieval accuracy.
(2) First tier (FT), which represents the recall of the top τ retrievals, τ denotes the number of 3D shapes in the database that belong to the same category as the query.
(3) Second tier (ST), which records the recall of the top 2τ retrievals.
(4) F-Measure (F), which calculates the harmonic mean of the precision and recall for the top retrievals.
(5) Discounted cumulated gain (DCG), which assigns higher weights to the top-ranking 3D shapes.
(6) Average Normalized Modified Retrieval Rank (ANMRR), which is a ranking-based measure that considers the ranking position of related 3D shapes.

Notably, a smaller ANMRR value means a better retrieval result, while for the other five criteria, a larger value means a better retrieval result.

4.1.3. Implementation detail

We utilize PyTorch framework for implementation. We use the AlexNet [40] pretrained on ImageNe [41] as the feature extractor to produce embeddings of 2D image and 3D shape (multi-view), and we aggregate multi-view embeddings to generate a compact 3D shape descriptor by a view-pooling layer [3]. We train our model along with the Stochastic Gradient Descent (SGD) algorithm with the momentum of 0.9. The initial learning rate of feature extractor is set to 3.5e−3. Two discriminators P_1 and P_2 are built in the same architecture including two fully-connected (FC) layers with a width of 1024, and the classifier is based on two FC layers with a subsequent softmax operation. The learning rate of these layers is 10 times that of the feature extractor. The learning rate decayed by a polynomial decay policy, where the initial learning rate is multiplied by $\left(1 - \frac{iter}{max\_iter}\right)^{power}$ with power = 0.9. The maximum iteration is set to 50,000 and the batch size is set to 16. Empirically, we reassign pseudo labels to 3D shapes every 500 iterations to continuously improve the reliability of the pseudo labels. We search the optimal hyper-parameters over $\lambda_{sgc}, \lambda_d, \lambda_s, \lambda_t, \lambda_{sa}, \lambda_{ta} \in \{0.1, 0.5, 1.0\}$. The optimal hyper-parameters on MI3DOR and MI3DOR-2 are shown in Table 1.

4.1.4. Compared methods

We compare CDA with the following competitive approaches:

- AlexNet [40]: It is a popular CNN framework that can be directly utilized to produce embeddings of 3D shapes (multi-view) and 2D images for retrieval.

Table 1
Optimal hyper-parameters on MI3DOR and MI3DOR-2 datasets.

	λ_{sgc}	λ_d	λ_s	λ_t	λ_{sa}	λ_{ta}
MI3DOR	1.0	1.0	1.0	1.0	0.1	0.1
MI3DOR-2	1.0	1.0	0.5	1.0	0.1	0.1

Table 2
Retrieval performance on MI3DOR dataset.

	NN	FT	ST	F	DCG	ANMRR
AlexNet [40]	0.424	0.323	0.469	0.099	0.345	0.667
CORAL [42]	0.362	0.174	0.256	0.060	0.199	0.816
GFK [25]	0.323	0.309	0.338	0.065	0.314	0.688
MEDA [43]	0.430	0.344	0.501	0.046	0.361	0.646
JGSA [44]	0.612	0.443	0.599	0.116	0.473	0.541
JAN [45]	0.446	0.343	0.495	0.085	0.364	0.647
RevGard [34]	0.650	0.505	0.643	0.112	0.542	0.474
DLEA [5]	0.764	0.558	0.716	0.143	0.597	0.421
SC-IFA [8]	0.721	0.584	0.721	0.163	0.637	0.363
HIFA [6]	0.778	0.618	0.768	0.151	0.654	0.362
Ours	**0.823**	**0.638**	**0.795**	**0.170**	**0.701**	**0.327**

- Correlation Alignment (CORAL) [42]: This approach diminish the domain shift by aligning the covariance matrices of the source and target domains.
- Geodesic Flow Kernel (GFK) [25]: It is a kernel-based domain adaptation method, aims to find a domain-invariant subspace on the geodesic flow from the source data to the target data.
- Manifold Embedded Distribution Alignment (MEDA) [43]: This approach aligns the cross-domain distribution in the manifold to handle the degenerated feature transformation problem.
- Joint Geometrical and Statistical Alignment (JGSA) [44]: This method finds two coupled projections for both domains to retain the discriminative information in source domain while maximize the variance in target domain.
- Joint Adaptation Networks (JAN) [45]: This method utilizes an adversarial training strategy to reduce the domain bias by maximizing the joint maximum mean discrepancy.
- RevGard [34]: This method obtains domain-invariant features by maximizing the loss of the domain discriminator and minimizing the loss of the classifier.
- Dual-level Embedding Alignment (DLEA) [5]: This method performs a domain adversarial loss and a centroid alignment loss on the source and target embeddings to achieve domain and class dual alignment.
- Semantic Consistency Guided Instance Feature Alignment (SC-IFA) [8]: This method reduces the domain gap via minimizing the maximum margin disparity discrepancy between the source and target distributions.
- Hierarchical Instance Feature Alignment (HIFA) [6]: This method achieves domain alignment via minimizing the maximum mean feature norm discrepancy between the source and target distributions.

4.2. Comparisons to the state-of-the-arts

As reported in Tables 2 and 3, we compare our method with three types of methods on the MI3DOR and MI3DOR-2 datasets. (1) source-only method without domain adaptation: AlexNet [40], (2) shallow domain adaptation methods: CORAL [42], GFK [25], MEDA [43], JGSA [44], (3) deep domain adaptation methods: JAN [45], RevGard [34], DLEA [5], SC-IFA [8], HIFA [6].

From Tables 2 and 3, we can observe that our method can achieve significantly better results than other works w.r.t all evaluation criteria. Specifically, on MI3DOR dataset, our method surpasses previous methods with a gain of 5.8%–154.8%, 3.2%–266.7%, 3.5%–210.5%, 4.3%–269.6%, 7.2%–252.3% in terms of NN, FT, ST, F, DCG, and a drop of 9.7%–59.9% on ANMRR. On MI3DOR-2 dataset, our method outperforms previous methods with gains of 10.3%–59.9%, 1.2%–82.8%, 1.5%–53.5%, 1.3%–77.7%, 0%-69.2% for NN, FT, ST, F, DCG, and a drop of 10.2%–41.0% on the part of ANMRR. This suggests the overall performance advantages of the proposed CDA for 2D image-based 3D shape retrieval. Moreover, the following observations can be drawn:

Table 3
Retrieval performance on MI3DOR-2 dataset.

	NN	FT	ST	F	DCG	ANMRR
AlexNet [40]	0.518	0.355	0.488	0.355	0.383	0.629
CORAL [42]	0.538	0.369	0.497	0.369	0.399	0.614
GFK [25]	0.513	0.471	0.495	0.471	0.484	0.527
MEDA [43]	0.570	0.392	0.523	0.392	0.425	0.590
JGSA [44]	0.585	0.405	0.533	0.405	0.433	0.577
JAN [45]	0.608	0.501	0.646	0.501	0.527	0.484
RevGard [34]	0.623	0.467	0.614	0.467	0.503	0.514
DLEA [5]	0.700	0.555	0.681	0.555	0.593	0.424
SC-IFA [8]	0.713	0.641	0.738	0.623	0.648	0.415
HIFA [6]	0.725	0.570	0.710	0.570	0.598	0.413
Ours	**0.800**	**0.649**	**0.749**	**0.631**	**0.648**	**0.371**

- The AlexNet model has the worst retrieval performance. This is expected and reasonable since this model is trained with only labeled 2D image data without any cross-domain transfer learning techniques, thus our method yields notable improvement over it.
- We notice that our method also shows promising retrieval results compared to shallow domain adaptation approaches, *e.g.* CORAL, GFK, MEDA, JGSA. These approaches usually apply mapping functions on the fixed embeddings, or align geometric characteristics of embeddings to reduce the domain discrepancy, thus the feature learning process and the domain adaptation process are separated during training. In contrast, our approach is able to train both the feature extractor and discriminators simultaneously, making the learned embeddings continuously updated. As a consequence, we can obtain more transferable embeddings and achieve better retrieval results.
- Moreover, our method achieves better performance than deep domain adaptation approaches, *e.g.* JAN, RevGrad, DLEA, SC-IFA, HIFA. This suggests that our method is more capable to exploit domain-invariant and semantic-discriminative features for 2D-to-3D tasks than other methods. The possible reason is that our method leverages the 2N-way adversarial learning, which can simultaneously reduce cross-domain discrepancies and enhance the semantic correlations between two distributions. Specifically, our method is superior to the category-agnostic methods (JAN, RevGrad), which only align the 2D image and 3D shape distributions globally and lead to the semantic misalignment issue, *i.e.*, features of different categories are mixed together in the feature space. Instead, our approach considers not only global alignment but also category distribution alignment. Furthermore, the retrieval results of CDA are better than the second-best method HIFA [6], which separates the category alignment from the domain alignment. In contrast, our method implements the global domain-wise and local category-wise distribution alignment in a single module.

We present some retrieval examples in Fig. 3. As shown in the figure, the first column represents 8 queries in different categories, and each row represents the top 10 retrievals placed by their ranking scores. The incorrect retrievals are marked as cyan, while the correct retrievals are marked as gray. Our method performs extremely well in distinctive categories such as car, airplane and curtain. We also observe that some categories are hard to distinguish that they yield relatively poor retrieval results. For example, xbox and radio, which differ only in a small part of their shape, are difficult to distinguish even for humans.

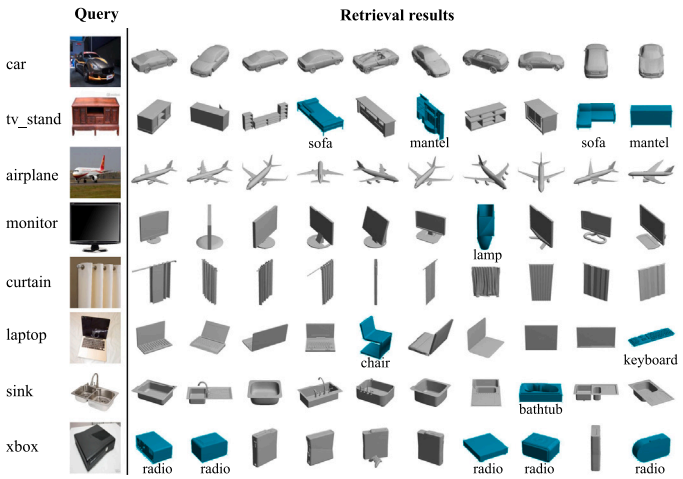


Fig. 3. Retrieval examples of MI3DOR-2 dataset. Cyan denotes the incorrect retrievals.

Table 4
Retrieval results of different modules on the MI3DOR and MI3DOR-2 datasets.

Method	MI3DOR						MI3DOR-2					
	NN	FT	ST	F	DCG	ANMRR	NN	FT	ST	F	DCG	ANMRR
CNN	0.424	0.323	0.469	0.099	0.345	0.667	0.518	0.355	0.488	0.355	0.383	0.629
CNN+JDCA	0.787	0.622	0.780	0.158	0.675	0.354	0.768	0.604	0.730	0.604	0.640	0.390
CNN+JDCA+SGC	0.809	0.628	0.786	0.164	0.690	0.340	0.781	0.633	0.739	0.621	0.645	0.381
CNN+JDCA+DBR	0.810	0.631	0.788	0.166	0.691	0.336	0.785	0.635	0.742	0.622	0.645	0.378
CNN+JDCA+SGC+DBR	**0.823**	**0.638**	**0.795**	**0.170**	**0.701**	**0.327**	**0.800**	**0.649**	**0.749**	**0.631**	**0.648**	**0.371**

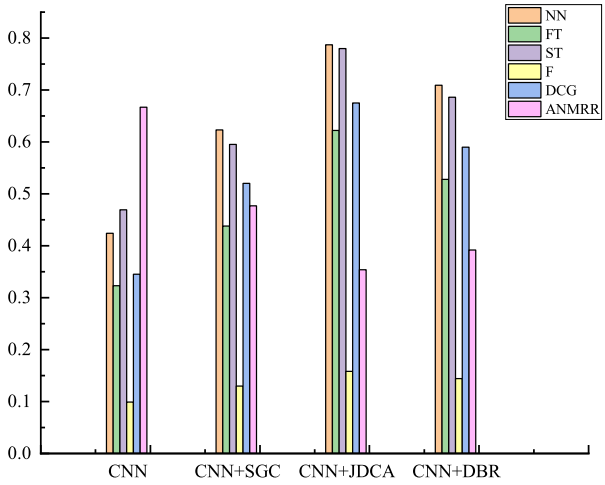


Fig. 4. Intuitive illustration of the effect of each component on MI3DOR dataset.

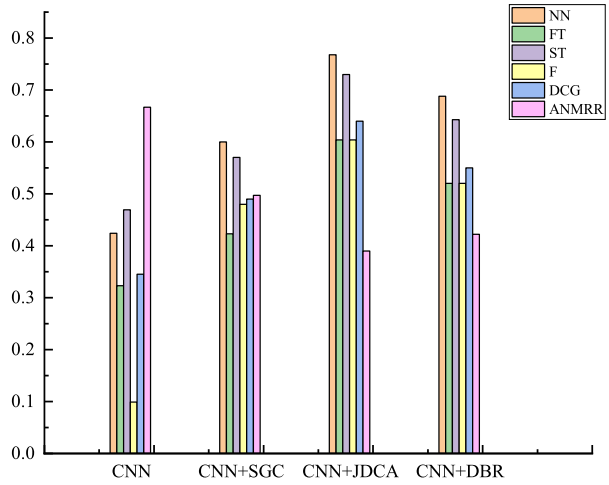


Fig. 5. Intuitive illustration of the effect of each component on MI3DOR-2 dataset.

4.3. Ablation study

To investigate the relative contribution of key modules in CDA, we performed extensive ablation analysis in Table 4, and the intuitive illustration of the effect of each component are shown in Fig. 4 and Fig. 5. Meanwhile, the corresponding PR curves are reported in Fig. 6 and Fig. 7.

Effect of Joint Domain-Class Alignment (JDCA). The baseline method, denoted by "CNN", merely learns visual embeddings without considering the cross-domain adaptation problems. From Table 4, we can observe that the baseline method yields a rather poor performance, due to its weakness in dealing with cross-domain divergence. Meanwhile, it can be seen that "CNN + JDCA" achieves higher retrieval results compared to the standard CNN, which suggests that JDCA is effective in minimizing the distribution discrepancy. Specifically, on the MI3DOR and MI3DOR-2 datasets, the network with JDCA module is superior to the baseline (standard CNN) with a rise of 85.61%/48.26%, 92.57%/70.14%, 66.31%/49.59%, 59.60%/70.14%, 95.65%/67.10% on the part of NN, FT, ST, F, DCG, and a drop of 46.93%/38.00% on ANMRR, respectively. By introducing the JDCA module, "CNN + JDCA" significantly boots the performance of the baseline, implying that the 2N-way adversarial can effectively alleviate the distribution disparity problem and enhance the correlation between different domain distributions.

Effect of Similarity Guided Constraint (SGC). "CNN + JDCA" can jointly align domain-level and class-level distributions, but it ignores the intrinsic semantic characteristic of both domains. By introducing the SGC module, "CNN + JDCA + SGC" can compact the embeddings of instances within a class and further provide more reliable labels

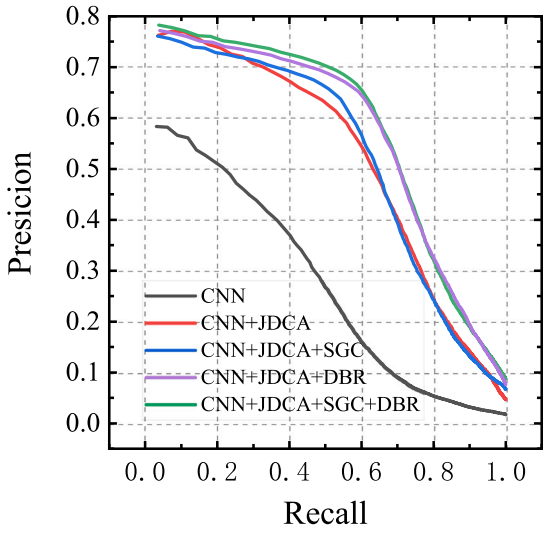


Fig. 6. PR curves with different components on MI3DOR dataset.

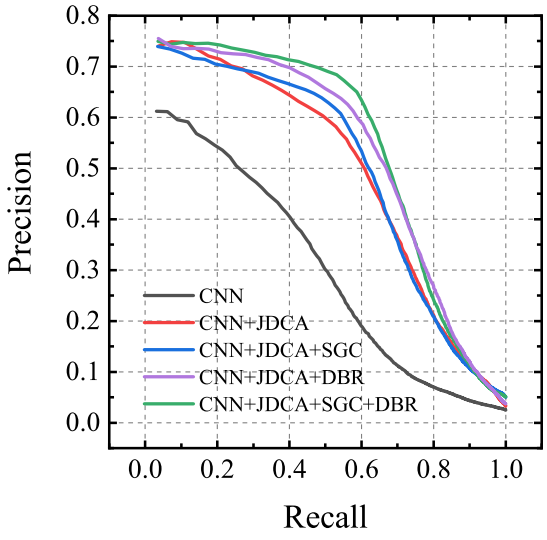


Fig. 7. PR curves with different components on MI3DOR-2 dataset.

for unlabeled 3D shapes. Specifically, on the MI3DOR and MI3DOR-2 datasets, "CNN + JDCA + SGC" is better than "CNN + JDCA" with a rise of 2.80%/1.69%, 0.96%/4.80%, 0.77%/1.23%, 3.80%/2.81%, 2.22%/0.78% with respect to NN, FT, ST, F, DCG, and a drop of 3.95%/2.30% on the part of ANMRR, respectively. This indicates that the proposed constraint is able to preserve the cross-domain similarities and suppress the detrimental influence brought by the false pseudo-labels. In the absence of the SGC module, there is a weaker constraint for feature extractor to learn domain-invariant embeddings.

Effect of Decision Boundary Refinement (DBR). "CNN + JDCA + SGC" can learn a domain-agnostic and category-discriminative feature space, but it neglects the category-ambiguity features. The DBR module can detect category-ambiguity embeddings and push them away from the decision boundaries, thus greatly facilitating domain alignment. Specifically, on the MI3DOR and MI3DOR-2 datasets, "CNN + JDCA + SGC + DBR" outperforms "CNN + JDCA + SGC" with a rise of 1.73%/2.43%, 1.59%/2.53%, 1.15%/1.35%, 3.66%/1.61%, 1.59%/0.47% with respect to NN, FT, ST, F, DCG, and a drop of 3.82%/2.62% on the part of ANMRR, respectively. This result further validates the model design advantage of DBR in reducing the number of category-ambiguity features. Moreover, we can observe a consistent improvement of "CNN + JDCA + DBR" on two benchmarks compared to "CNN + JDCA". With the collaborative supervision of JDCA, SGC and DBR, our model achieves the best retrieval results.

4.4. Qualitative evaluation

In order to understand the effect of each module more intuitively, we utilize the t-SNE [46] visualization technique to observe the feature distribution in 2D-space. Fig. 8 and Fig. 9 show the visualization of embeddings learned on the test set of MI3DOR dataset and MI3DOR-2 dataset, respectively. The blue dots denote the 3D shape embeddings and the red dots represent the 2D image embeddings.

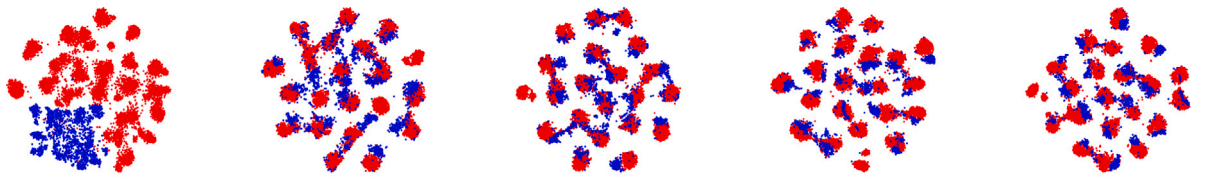
According to Fig. 8(a) and Fig. 9(a), there is a small overlap between 2D images (red) and 3D shapes (blue), indicating cross-domain distributions exist large discrepancy. Through the domain adaptation phase with joint domain-class alignment module (JDCA), abundant overlaps between two domains are shown in Fig. 8(b) and Fig. 9(b). It suggests that JDCA is effective to narrow the cross-domain distribution gap. This shares a similar observation and underlying reason as in Table 4, validating the benefit of JDCA in joint align domain-level and category-level distributions.

Furthermore, we found that JDCA can roughly align distributions at category-level, but it fails to handle some features well. We think this is caused by some incorrect pseudo-labeled 3D shapes. Compared to "CNN + JDCA", we can observe that the 3D shape embeddings learned by "CNN + JDCA + SGC" displayed smaller intra-class distance as well as larger inter-class margin. This indicates that the similarity guided constraint (SGC) module can generate more discriminative features and suppress the detrimental influence brought by the false pseudo-labels.

Besides that, the feature distributions of "CNN + JDCA + DBR" are much more separated than "CNN + JDCA", which demonstrates that our decision boundary refinement (DBR) module is able to reduce the generation of class-ambiguity embeddings to some extent. Benefiting from the SGC, JDCA and DBR modules, 2D image and 3D shape embeddings of the same category are co-located as shown in Fig. 8(e) and Fig. 9(e). This further indicates that our model can preserve cross-domain similarities and reduce the local semantic divergences when aligning distributions.

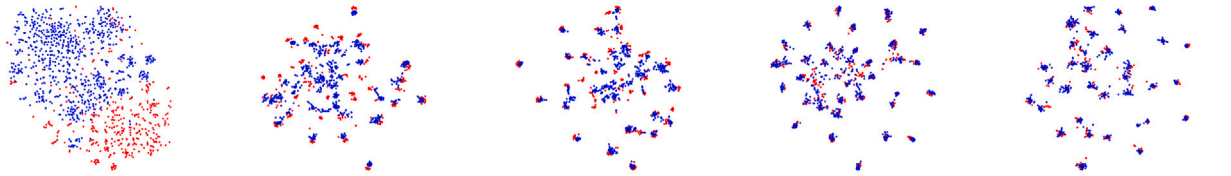
5. Conclusion

In this paper, we developed a novel collaborative distribution alignment network for 2D image-based 3D shape retrieval. We first learned domain-specific discriminative embeddings by employing the similarity guided constraint to preserve semantic structures in both domains. Then we proposed a joint domain-class alignment module, aiming to address the problem of semantic inconsistency incurred by pure global feature alignment. Finally, we introduced a decision boundary refinement mechanism to further generating embeddings far from the decision boundaries. Experimental results validate the effectiveness of CDA, which yields very competitive retrieval performance compared to state-of-the-art methods. In the future work, we will explore 3D shapes in the real world for experiments. The 3D shapes involved in this paper are all CAD shapes, which have clear background and clear object contour, so they are easy to be recognized. In contrast, 3D shapes in the real world have more complex background and richer visual information. Hence, it is necessary to take some measures to enhance the representation ability of features to better handle more challenging 3D shapes. In view of the successful application of spatial attention mechanism in Baidu's PaddlePaddle deep learning platform [47,48], we will design a spatial attention module in the future to further explore the visual content of 3D shapes.



(a) CNN (b) CNN+JDCA (c) CNN+JDCA+SGC (d) CNN+JDCA+DBR (e) CNN+JDCA+SGC+DBR

Fig. 8. The feature visualization of different modules via T-SNE on MI3DOR dataset.



(a) CNN (b) CNN+JDCA (c) CNN+JDCA+SGC (d) CNN+JDCA+DBR (e) CNN+JDCA+SGC+DBR

Fig. 9. The feature visualization of different modules via T-SNE on MI3DOR-2 dataset.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Nature Science Foundation of China (U21B2024, 61772359), and the grant of Tianjin New Generation Artificial Intelligence Major Program, China (19ZXZNGX00110).

References

[1] J. Chen, Y. Fang, Deep cross-modality adaptation via semantics preserving adversarial learning for sketch-based 3D shape retrieval, in: ECCV, Vol. 11217, 2018, pp. 624–640.
[2] G. Dai, J. Xie, Y. Fang, Deep correlated holistic metric learning for sketch-based 3D shape retrieval, IEEE Trans. Image Process. 27 (7) (2018) 3374–3386.
[3] H. Su, S. Maji, E. Kalogerakis, E.G. Learned-Miller, Multi-view convolutional neural networks for 3D shape recognition, in: ICCV, 2015, pp. 945–953.
[4] J. Jiang, D. Bao, Z. Chen, X. Zhao, Y. Gao, MLVCNN: multi-loop-view convolutional neural network for 3D shape retrieval, in: AAAI, 2019, pp. 8513–8520.
[5] H. Zhou, A. Liu, W. Nie, Dual-level embedding alignment network for 2D image-based 3D object retrieval, in: ACM MM, 2019, pp. 1667–1675.
[6] H. Zhou, W. Nie, W. Li, D. Song, A. Liu, Hierarchical instance feature alignment for 2D image-based 3D shape retrieval, in: IJCAI, 2020, pp. 839–845.
[7] X. Lu, L. Liu, L. Nie, X. Chang, H. Zhang, Semantic-driven interpretable deep multi-modal hashing for large-scale multimedia retrieval, IEEE Trans. Multim. 23 (2021) 4541–4554.
[8] H. Zhou, W. Nie, D. Song, N. Hu, X. Li, A. Liu, Semantic consistency guided instance feature alignment for 2D image-based 3D shape retrieval, in: ACM MM, 2020, pp. 925–933.
[9] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V.S. Lempitsky, Domain-adversarial training of neural networks, J. Mach. Learn. Res. 17 (2016) 59:1–59:35.
[10] S. Cicek, S. Soatto, Unsupervised domain adaptation via regularized conditional alignment, in: ICCV, 2019, pp. 1416–1425.
[11] E. Tzeng, J. Hoffman, K. Saenko, T. Darrell, Adversarial discriminative domain adaptation, in: CVPR, 2017, pp. 2962–2971.
[12] Z. Pei, Z. Cao, M. Long, J. Wang, Multi-adversarial domain adaptation, in: AAAI, 2018, pp. 3934–3941.
[13] M.M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, P. Vandergheynst, Geometric deep learning: Going beyond euclidean data, IEEE Signal Process. Mag. 34 (4) (2017) 18–42.
[14] C.R. Qi, H. Su, K. Mo, L.J. Guibas, PointNet: Deep learning on point sets for 3D classification and segmentation, in: CVPR, 2017, pp. 77–85.
[15] P. Wang, Y. Liu, Y. Guo, C. Sun, X. Tong, O-CNN: octree-based convolutional neural networks for 3D shape analysis, ACM Trans. Graph. 36 (4) (2017) 72:1–72:11.
[16] J. Chen, J. Qin, Y. Shen, L. Liu, F. Zhu, L. Shao, Learning attentive and hierarchical representations for 3D shape recognition, in: ECCV, Vol. 12360, 2020, pp. 105–122.
[17] W. Nie, K. Wang, Q. Liang, R. He, Panorama based on multi-channel-attention CNN for 3D model recognition, Multim. Syst. 25 (6) (2019) 655–662.
[18] G. Dai, J. Xie, Y. Fang, Siamese CNN-BiLSTM architecture for 3D shape representation learning, in: IJCAI, 2018, pp. 670–676.
[19] T. Lee, Y. Lin, H. Chiang, M. Chiu, W.H. Hsu, P. Huang, Cross-domain image-based 3D shape retrieval by view sequence learning, in: 3DV, 2018, pp. 258–266.
[20] J. Huang, A.J. Smola, A. Gretton, K.M. Borgwardt, B. Schölkopf, Correcting sample selection bias by unlabeled data, in: NeurIPS, 2006, pp. 601–608.
[21] J. Blitzer, R.T. McDonald, F. Pereira, Domain adaptation with structural correspondence learning, in: EMNLP, 2006, pp. 120–128.
[22] C. Wang, S. Mahadevan, Manifold alignment without correspondence, in: IJCAI, 2009, pp. 1273–1278.
[23] S.J. Pan, I.W. Tsang, J.T. Kwok, Q. Yang, Domain adaptation via transfer component analysis, in: IJCAI, 2009, pp. 1187–1192.
[24] K. Saenko, B. Kulis, M. Fritz, T. Darrell, Adapting visual category models to new domains, in: ECCV, Vol. 6314, 2010, pp. 213–226.
[25] B. Gong, Y. Shi, F. Sha, K. Grauman, Geodesic flow kernel for unsupervised domain adaptation, in: CVPR, 2012, pp. 2066–2073.
[26] H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu, W. Zuo, Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation, in: CVPR, 2017, pp. 945–954.
[27] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, B. Wang, Moment matching for multi-source domain adaptation, in: ICCV, 2019, pp. 1406–1415.
[28] C. Chen, Z. Fu, Z. Chen, S. Jin, Z. Cheng, X. Jin, X. Hua, HoMM: Higher-order moment matching for unsupervised domain adaptation, in: AAAI, 2020, pp. 3422–3429.
[29] B. Sun, K. Saenko, Deep CORAL: correlation alignment for deep domain adaptation, in: ECCV, Vol. 9915, 2016, pp. 443–450.
[30] X. Xie, J. Chen, Y. Li, L. Shen, K. Ma, Y. Zheng, Self-supervised cyclegan for object-preserving image-to-image domain adaptation, in: ECCV, Vol. 12365, 2020, pp. 498–513.
[31] M. Liu, O. Tuzel, Coupled generative adversarial networks, in: NeurIPS, 2016, pp. 469–477.
[32] J. Wang, J. Jiang, Conditional coupled generative adversarial networks for zero-shot domain adaptation, in: ICCV, 2019, pp. 3374–3383.
[33] Y. Jiang, T. Zhang, D. Ho, Y. Bai, C.K. Liu, S. Levine, J. Tan, SimGAN: Hybrid simulator identification for domain adaptation via adversarial reinforcement learning, 2021, CoRR abs/2101.06005.
[34] Y. Ganin, V.S. Lempitsky, Unsupervised domain adaptation by backpropagation, in: ICML, Vol. 37, 2015, pp. 1180–1189.
[35] B.T. Phong, Illumination for computer generated pictures, Commun. ACM 18 (1975) 311–317.
[36] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, S.C.H. Hoi, Deep learning for person re-identification: A survey and outlook, 2020, CoRR abs/2001.04193.
[37] K. Saito, K. Watanabe, Y. Ushiku, T. Harada, Maximum classifier discrepancy for unsupervised domain adaptation, in: CVPR, 2018, pp. 3723–3732.
[38] Y. Du, Z. Tan, Q. Chen, X. Zhang, Y. Yao, C. Wang, Dual adversarial domain adaptation, 2020, CoRR abs/2001.00153.
[39] A. Liu, W. Nie, Y. Gao, Y. Su, View-based 3-D model retrieval: A benchmark, IEEE Trans. Cybern. 48 (3) (2018) 916–928.

[40] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: NeurIPS, 2012, pp. 1106–1114.
[41] J. Deng, W. Dong, R. Socher, L. Li, K. Li, F. Li, ImageNet: A large-scale hierarchical image database, in: CVPR, 2009, pp. 248–255.
[42] B. Sun, J. Feng, K. Saenko, Return of frustratingly easy domain adaptation, in: AAAI, 2016, pp. 2058–2065.
[43] J. Wang, W. Feng, Y. Chen, H. Yu, M. Huang, P.S. Yu, Visual domain adaptation with manifold embedded distribution alignment, in: ACM MM, 2018, pp. 402–410.
[44] J. Zhang, W. Li, P. Ogunbona, Joint geometrical and statistical alignment for visual domain adaptation, in: CVPR, 2017, pp. 5150–5158.
[45] M. Long, H. Zhu, J. Wang, M.I. Jordan, Deep transfer learning with joint adaptation networks, in: ICML, Vol. 70, 2017, pp. 2208–2217.
[46] V.D.M. Laurens, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2605) (2008) 2579–2605.
[47] Y. Ma, D. Yu, T. Wu, H. Wang, PaddlePaddle: An open-source deep learning platform from industrial practice, Front. Data Comput. 1 (1) (2019).
[48] Paddlepaddle, Paddlepaddle: An easy-to-use, easy-to-learn deep learning platform, 2020, http://www.paddlepaddle.org/.