

Consistent Domain Structure Learning and Domain Alignment for 2D Image-Based 3D Objects Retrieval

Yuting Su, Yuqian Li, Dan Song*, Weizhi Nie, Wenhui Li and An-An Liu*

School of Electrical and Information Engineering, Tianjin University, China

{dan.song, liuanan}@tju.edu.cn

Abstract

2D image-based 3D objects retrieval is a new topic for 3D objects retrieval which can be used to manage 3D data with 2D images. The goal is to search some related 3D objects when given a 2D image. The task is challenging due to the large domain gap between 2D images and 3D objects. Therefore, it is essential to consider domain adaptation problems to reduce domain discrepancy. However, most of the existing domain adaptation methods only utilize the semantic information from the source domain to predict labels in the target domain and neglect the intrinsic structure of the target domain. In this paper, we propose a domain alignment framework with consistent domain structure learning to reduce the large gap between 2D images and 3D objects. The domain structure learning module makes use of both the semantic information from the source domain and the intrinsic structure of the target domain, which provides more reliable predicted labels to the domain alignment module to better align the conditional distribution. We conducted experiments on two public datasets, MI3DOR and MI3DOR-2, and the experimental results demonstrate the proposed method outperforms the state-of-the-art methods.

1 Introduction

The 3D objects retrieval, especially image-based 3D objects retrieval, is a very important task for data management. Most previous works aim to classify 3D objects or retrieve them within the same modality [Li *et al.*, 2019b; He *et al.*, 2018; Liu *et al.*, 2018a]. For example, Li *et al.* [Li *et al.*, 2019b] proposed an angular triplet-center loss for 3D retrieval and conducted retrieval experiments on the ModelNet dataset [Wu and *et al.*, 2015] and ShapeNet dataset [Chang and *et al.*, 2015], respectively.

However, there are less works about the 2D image-based 3D object retrieval with two-fold reasons: 1) the query and gallery are from different modalities and domains, and their

features have different distributions, and 2) lack of related annotated datasets. Since it is of great value in application (e.g., 3D printing, entertainment) and research, some researchers and academic organizations pay attention to image-based 3D objects retrieval (e.g., [Lee *et al.*, 2018; Li *et al.*, 2019a]) and use domain adaptation to reduce domain gap [Su *et al.*, 2019]. They often assume one domain as source domain with labels, and another one is target domain without labels. In this paper, we also adopt domain adaptation idea to reduce domain gap and learn a latent space for 2D images and 3D objects. We set 2D images as the source domain and 3D objects as the target domain.

1.1 Motivations

2D image-based 3D objects retrieval aims to find related 3D objects in the gallery when given 2D images. There are two critical problems to solve for this promising but challenging task:

How to build a unified feature space for 2D images and 3D objects. 2D images and 3D objects are from different modalities. To reduce the gap between 3D objects and 2D images in visual level, we can capture a set of rendered images for each 3D object, e.g. MVCNN [Su *et al.*, 2015]. In this case, we can extract the visual representations of 3D objects and 2D images by MVCNN and CNN (e.g., AlexNet [Krizhevsky *et al.*, 2012]), respectively, and compute their similarity. However, the retrieval performances in such a naive way are poor with two-fold reasons: 1) their features have different distributions and 2) the high-dimensional features contain noisy information. Therefore, it is mandatory to explore a unified feature space with low-dimensional structure for 2D images and 3D objects.

How to fully exploit intrinsic structure of target domain for domain alignment. Domain alignment aims to reduce the discrepancy in marginal distributions (global level) and conditional distributions (local level). In unsupervised domain adaptation, the conditional distributions alignment at the local level plays a more important role than the marginal distributions alignment at the global level [Zhang *et al.*, 2019]. To achieve the conditional distributions alignment, the most of existing methods just rely on the labeled samples in the source domain to train a classifier and predict the labels of samples from the target domain (e.g., [Su *et al.*, 2019]). However, these methods neglect the intrinsic structure in the target

*Corresponding authors: Dan Song & An-An Liu

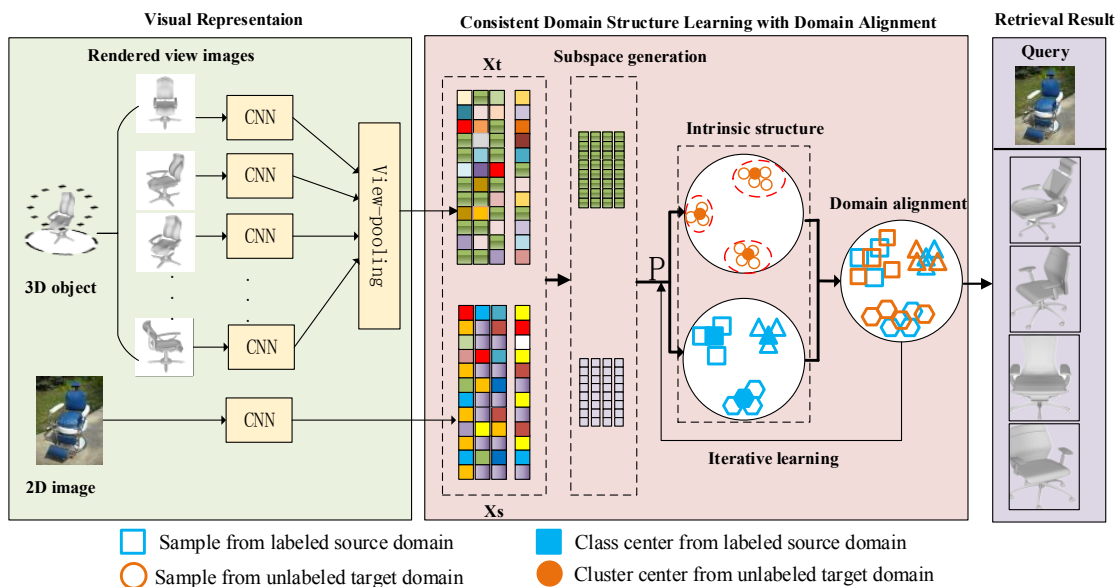


Figure 1: Architecture of the proposed method for 2D image-based 3D objects retrieval. It contains five modules: visual representation, subspace generation, consistent domain structure learning, domain alignment and retrieval. \mathbf{P} is a projection matrix.

domain and consequently result in poor quality of pseudo-labels. Therefore, it is necessary to fully exploit semantic information from the source domain and the intrinsic structure in the target domain to better align the conditional distributions.

To handle the above issues, we propose a consistent domain structure learning and domain alignment framework to reduce the large domain gap for 2D images and 3D objects. As shown in Fig. 1, the framework contains five modules: visual feature extraction, subspace generation, consistent domain structure learning, domain alignment and retrieval. The visual feature extraction module aims to extract visual representation for 2D images and 3D objects. After this step, we use subspace generation module to remove noisy information in the original visual features and obtain more robust features. To learn a latent space for 2D images and 3D objects, we introduce a consistent domain structure learning (CDSL) module that can predict the class probability of the target domain, which further provides more reliable labels to the domain alignment module. At last, the output of the CDSL module, i.e. the class probability, is treated as the invariant feature for 2D images and 3D objects and similarity is measured by L2-distance.

1.2 Contributions

In this paper, the main contributions are as follows:

- We propose a novel visual domain adaptation method for 2D image-based 3D objects retrieval, which can learn a common latent subspace with domain alignment.
- We explore the domain structure consistency between 2D images and 3D objects to promote the prediction accuracy for the labels of the 3D objects domain, which can better guide the conditional distribution alignment.

- We conduct extensive experiments on two popular public datasets, MI3DOR and MI3DOR-2. The experimental results show the superiority of the proposed method.

2 Related Work

2.1 3D Objects Retrieval

3D objects retrieval has attracted more attentions than ever [Liu *et al.*, 2016; 2018a; 2018b; Lu *et al.*, 2019a; 2019b; Cheng *et al.*, 2018; Hong *et al.*, 2016]. Generally, we can divide existing 3D objects retrieval method into two categories: model-based methods and view-based methods.

Model-based methods [Xie *et al.*, 2017; Shen *et al.*, 2018] deal with the 3D structure of original objects, such as voxel grid [Ghadai *et al.*, 2019] and point cloud [Qi *et al.*, 2017]. For example, Ghadai *et al.* [Ghadai *et al.*, 2019] learned features by a multi-level voxel representation of spatial data.

View-based methods represent each 3D object by a set of rendered images. For example, Su *et al.* [Su *et al.*, 2015] proposed Multi-view Convolutional Neural Network (MVCNN) to represent 3D objects by aggregating the features of a set of rendered images in an end-to-end way. Feng *et al.* [Feng *et al.*, 2018] modeled the hierarchical correlation to learn 3D shape description.

Currently, the image-based 3D objects retrieval framework is a new promising branch for 3D objects retrieval [Wang *et al.*, 2015; Dai *et al.*, 2017]. Dai *et al.* [Dai *et al.*, 2017] reduced the discrepancy between sketch and 3D shape domain by a deep correlated metric learning method. However, the method focuses on sketch-based 3D objects retrieval. It is different from our task since natural images contain more complex backgrounds and more noisy objects when compared sketch. Recently, Lee *et al.* [Lee *et al.*, 2018] learned an embedding space for images and 3D shapes in an end-to-end method. However, we can not annotate all the images and 3D

objects in practice. Therefore, it is necessary to reduce the cross domain gap by an unsupervised method.

2.2 Domain Adaptation

Generally, most of domain adaptation methods reduce the domain discrepancy in an unsupervised way, where given two related domains (sharing common classes), one labeled (called source domain) and another one unlabeled (called target domain).

At early time, unsupervised domain adaption methods focus on the marginal distribution alignment, e.g., Gopalan et al. [Gopalan *et al.*, 2011] generated an intermediate representation between two domains on the Grassmann manifold. Later, Gong et al. [Gong *et al.*, 2012] exploited low-dimensional structure with the GFK [Gopalan *et al.*, 2011].

Although two domains are aligned in the marginal distributions, it can not guarantee to produce better classification performance. The main reason for this problem is the lack of labels of samples in the target domain and can not provide labels to align the conditional distributions. For example, Wang et al. [Wang *et al.*, 2018] proposed an adaptive method to align the marginal and conditional distribution with pseudo-label learning. However, this method neglects the intrinsic structure of the target domain. Facing the large gap for our task, we adopt domain adaptation ideas to reduce domain discrepancy and learn invariant features for the query (2D images) and gallery (3D objects).

3 Proposed Method

3.1 Overview

In this work, we propose a consistent domain structure learning and alignment framework. It can learn a projection matrix \mathbf{P} that maps the visual representation of labeled source domain (2D images) D_s and unlabeled target domain (3D objects) D_t to a latent space of a lower dimensional Z . In the new learned common latent space Z , the domain structure consistency means: 1) nearby samples are likely from the same class; and 2) samples on the same structure (e.g., a cluster) are likely from the same class. Based on this assumption, we explore domain structure consistency to obtain high-quality pseudo labels for the target domain which benefits the conditional distribution alignment.

The proposed framework mainly contains four key modules (as shown in Fig. 1):

Visual representation. Visual representation module aims to obtain the representation of 3D objects or 2D images. For 3D objects, we can use a set of rendered images captured by cameras around the objects under different viewpoints. In this paper, we used Phong model [Phong, 1975] to capture the multi-view images for 3D objects and used MVCNN [Su *et al.*, 2015] to obtain the visual representation of 3D objects. For 2D images, since 2D image can be as only one single view, we also used the same network to obtain the visual representation of 2D images.

Subspace generation. As we know, the high dimensional features always contain noisy information or redundant information. Therefore, we reduce the dimension of the fea-

ture first by Principle Component Analysis (PCA) and obtain more robust feature compared with the original feature data.

Domain alignment. Domain alignment is a key step for the domain adaptation. Previous works always rely on the distance between the feature points, while we suggest to use the label to constrain the distance based on whether the labels are the same (as shown in Eq.(3)).

Consistent domain structure learning. The consistent domain structure learning module aims to predict the label of the samples from target domain. It also provides more reliable likelihood information for the domain alignment. It is different from traditional pseudo-label learning methods since it joints the intrinsic structure of the target and source domains to promote the pseudo-labels prediction accuracy on the target domain.

Before introducing the details, we first define the notations. The 2D images belong to the source domain $D_s = \{(\mathbf{x}_i^s, y_i^s)\} (i = 1, 2, 3, \dots, n_s)$ and 3D objects belong to target domain $D_t = \{(\mathbf{x}_i^t)\} (i = 1, 2, 3, \dots, n_t)$. $\mathbf{x}_i \in R^d$ represents the visual representation of the i^{th} sample from 2D images or 3D objects, d means the dimension of the original feature, and $y_i^s \in Y^s$ denotes the label of the i^{th} sample in the source domain. Y means the label space. In this paper, we assume that Y^t is equal to Y^s .

3.2 Subspace Generation

The high dimensional feature space always contain noise information or redundant information. To address this, we adopt PCA to reduce the dimension of the feature first and obtain more robust feature compared with original feature data as in [Wang and Breckon, 2019]. We concatenated the features from the visual representation of 2D images and 3D objects as a matrix $\mathbf{X} = [x_1^s, x_2^s, \dots, x_{n_s}^s, x_1^t, x_2^t, \dots, x_{n_t}^t] \in R^{d \times n}$, where $n = n_s + n_t$. The objective of PCA is:

$$\max_{\mathbf{V}^T \mathbf{V} = \mathbf{I}} \text{tr}(\mathbf{V}^T \mathbf{X} \mathbf{H} \mathbf{X}^T \mathbf{V}) \quad (1)$$

where $\mathbf{H} = \mathbf{I} - \frac{1}{n} \mathbf{1} \mathbf{1}^T$ is the center matrix, $\mathbf{1}$ denotes an all one $n \times n$ matrix, $\text{tr}(\cdot)$ means the trace of matrix. To solve Eq.(1), we transform Eq.(1) to the following eigenvalue problem:

$$\mathbf{X} \mathbf{H} \mathbf{X}^T \mathbf{v} = \phi \mathbf{v} \quad (2)$$

Then by solving the Eq.(2), we can obtain the projection matrix $\mathbf{V} = [v_1, \dots, v_{d_1}] \in R^{d \times d_1}$ and the lower robust feature can be as $\bar{\mathbf{X}} = \mathbf{V}^T \mathbf{X}$, where $\bar{\mathbf{X}} \in R^{d_1 \times n}$. After applying PCA, the feature dimensionality is d_1 and $d_1 \leq d$.

3.3 Domain Alignment

Subspace generation module aims to reduce noisy information in the original feature space and generate a new feature space $\bar{\mathbf{X}}$ in an unsupervised manner. However, it does not consider the domain discrepancy. Therefore, we hope we can find a projection matrix \mathbf{P} , which can map $\bar{\mathbf{X}}$ to a more lower dimensional and discriminative space Z . In spired by MMD [Gretton *et al.*, 2012] and locality preserving projection [He and Niyogi, 2003], we define the objection function as follows:

$$\min_{\mathbf{P}} \sum_{i,j} \|\mathbf{P}^T \mathbf{x}_i - \mathbf{P}^T \mathbf{x}_j\|_2^2 \mathbf{B}_{ij} \quad (3)$$

where $\mathbf{P} \in R^{d_1 \times d_2}$ and $d_2 \leq d_1$ is the learned feature space dimensionality. The term $\mathbf{B} \in R^{(n_s+n_t) \times (n_s+n_t)}$, where $B_{ij} = 1$ if $y_i = y_j$, otherwise $B_{ij} = 0$, aims to choose good mapping for each class. The matrix \mathbf{B} is a simple version of MMD. To align two domains and obtain the optimal \mathbf{P} , we need to know the pseudo labels for the target objects. The pseudo label learning will be described in the next subsection.

3.4 Consistent Domain Structure Learning

The consistent domain structures learning module aims to predict the label of the samples from target domain in the common latent space Z and provides reliable likelihood information for domain alignment. We first define a probability annotation matrix \mathbf{M} . Formally, let $c \in \{1, \dots, C\}$ represents the class label. The probability annotation matrix $\mathbf{M} \in R^{C \times n_t}$, where its entry value \mathbf{M}_{c_j} denotes the probability of \mathbf{x}_j^t belonging to class c . The source domain data $\bar{\mathbf{x}}^s$ and the target domain data $\bar{\mathbf{x}}^t$ can be reformulated $\mathbf{z}^s = \mathbf{P}^T \bar{\mathbf{x}}^s$ and $\mathbf{z}^t = \mathbf{P}^T \bar{\mathbf{x}}^t$, respectively.

Intrinsic Structure of the Source Domain

We use the class prototype of source domain as the intrinsic structure of the source domain. Each class center \mathbf{h}_c^s of the source domain can be defined by the mean feature vector of the source domain samples whose label is c , which can be written as:

$$\mathbf{h}_c^s = \frac{1}{|D_s^{(c)}|} \sum_i^{n_s} \mathbf{z}_i^s \mathbf{I}(y_i^s = c) \quad (4)$$

where $\mathbf{I}(\cdot)$ denotes an indicator function. Then we can the conditional probability of a given target sample \mathbf{x}^t belonging to class c through the following equation:

$$p_1(c|\mathbf{x}^t) = \frac{\exp(-\|\mathbf{z}^t - \mathbf{h}_c^s\|)}{\sum_{c=1}^C \exp(-\|\mathbf{z}^t - \mathbf{h}_c^s\|)} \quad (5)$$

Intrinsic Structure of the Target Domain

After finishing the above step, we can obtain the labels of samples in the target domain. However, the above method does not consider the intrinsic structure of the target domain. To explore such structure information, we use K-means to generate $|C|$ clusters in the projection feature space Z .

We first initialize the cluster center of the target domain by Eq.(4). Then, to make the cluster centroid of the target domain close to the class prototype of the source domain, we define the optimization problem as in [Wang and Breckon, 2019]:

$$\begin{aligned} \min_{\mathbf{A}} \sum_j^{|C|} \sum_j^{|C|} \mathbf{A}_{ij} d(\mathbf{h}_i^t, \mathbf{h}_j^s) \\ \text{s.t. } \forall i, \sum_j \mathbf{A}_{ij} = 1; \forall j, \sum_i \mathbf{A}_{ij} = 1 \end{aligned} \quad (6)$$

where $\mathbf{A} \in \{0, 1\}^{|C| \times |C|}$ represents a matching matrix where $\mathbf{A}_{ij} = 1$ means the i^{th} center of the target cluster matches with the j^{th} the class center of source domain. \mathbf{h}_i^t denotes the i^{th} cluster center in the target domain. The Eq.(6) can be solved according to [Wang and Chen, 2017].

Let \mathbf{h}_c^t denotes the cluster center corresponding to the class c , similar to Eq. (5), we can calculate the conditional probability of a given target sample \mathbf{x}^t belonging to class c :

$$p_2(c|\mathbf{x}^t) = \frac{\exp(-\|\mathbf{z}^t - \mathbf{h}_c^t\|)}{\sum_{c=1}^{|C|} \exp(-\|\mathbf{z}^t - \mathbf{h}_c^t\|)} \quad (7)$$

Consistent Structure Learning

Although the above two methods can label the samples from the target domain by iterative learning strategy, they are intrinsically different. The high probability $p_1(c|\mathbf{x}^t)$ means the corresponding samples are close to the class center of the source domain, whilst the high probability $p_2(c|\mathbf{x}^t)$ means the corresponding samples are close to the cluster center in the target domain regardless of the distance to the source domain. To make use of the these complementary information, we combine Eq.(5) and Eq.(7) as follows:

$$p(c|\mathbf{x}^t) = (1 - \mu)p_1(c|\mathbf{x}^t) + \mu p_2(c|\mathbf{x}^t) \quad (8)$$

where μ is a trade-off parameter. As a result, the pseudo-label of a given target sample \mathbf{x}_t can be predicted by: $\hat{y}^t = \arg \max_{c \in Y} p(c|\mathbf{x}^t)$. Now, we obtain all the pseudo-labels of samples in the target domain and compute the optimum \mathbf{P} .

3.5 Optimization

The objective Eq.(3) can be rewritten as:

$$\max_{\mathbf{P}} \frac{\text{tr}(\mathbf{P}^T \hat{\mathbf{X}} \mathbf{D} \hat{\mathbf{X}}^T \mathbf{P})}{\text{tr}(\mathbf{P}^T (\hat{\mathbf{X}} \mathbf{L} \hat{\mathbf{X}}^T) \mathbf{P})} \quad (9)$$

where $\mathbf{L} = \mathbf{D} - \mathbf{B}$ is the laplacian matrix, and \mathbf{D} denotes a diagonal matrix whose element $D_{ii} = \sum_j B_{ij}$. $\hat{\mathbf{X}} \in R^{d_1 \times (n_s+n_t)}$ is the combination matrix for all data. For Eq.(9), we add a regularization term $\text{Tr}(\mathbf{P}^T \mathbf{P})$ to obtain a better projection matrix \mathbf{P} . Therefore, the Eq.(9) can be rewritten as:

$$\max_{\mathbf{P}} \frac{\text{tr}(\mathbf{P}^T \hat{\mathbf{X}} \mathbf{D} \hat{\mathbf{X}}^T \mathbf{P})}{\text{tr}(\mathbf{P}^T (\hat{\mathbf{X}} \mathbf{L} \hat{\mathbf{X}}^T + \mathbf{I}) \mathbf{P})} \quad (10)$$

Eq.(10) is equivalent to the following generalized eigenvalue problem:

$$\hat{\mathbf{X}} \mathbf{D} \hat{\mathbf{X}}^T \mathbf{p} = \lambda (\hat{\mathbf{X}} \mathbf{L} \hat{\mathbf{X}}^T + \mathbf{I}) \mathbf{p} \quad (11)$$

To solve Eq.(11), we set $\mathbf{P} = [p_1, p_2, \dots, p_{d_2}]$, where p_1, p_2, \dots, p_{d_2} are eigenvectors in terms of the largest d_2 eigenvalues.

4 Experiment

4.1 Datasets

M3DOR. The dataset [Li *et al.*, 2019a] has 21 categories and contains 21,000 images and 7690 3D objects. The 2D training set is about 10500 images and the rest images are be used for testing. The 3D objects training set contains 3,842 objects and the rest 3,748 objects are used for testing.

M3DOR-2. The dataset [Su *et al.*, 2019] has 40 categories and also contains 19,694 images and 3,982 objects. The 2D training set is about 19,294 images and the rest images are used to test. For the 3D objects, 3,182 objects are used for training and the rest 800 objects are used for testing.

Dataset	Method	NN	FT	ST	F-measure	DCG	ANMRR
MI3DOR	1NN	0.252	0.149	0.237	0.043	0.162	0.846
	CORAL[Sun <i>et al.</i> , 2016]	0.362	0.174	0.256	0.060	0.199	0.816
	GFK[Gong <i>et al.</i> , 2012]	0.323	0.309	0.338	0.065	0.314	0.688
	JGSA[Zhang and <i>et al.</i> , 2017]	0.528	0.391	0.523	0.099	0.416	0.596
	JAN[Long <i>et al.</i> , 2017]	0.435	0.341	0.494	0.085	0.362	0.649
	LSR[Xie <i>et al.</i> , 2018]	0.615	0.448	0.603	0.116	0.479	0.536
	JMMD[Li <i>et al.</i> , 2019a]	0.446	0.343	0.495	0.085	0.364	0.647
	MVML[Li <i>et al.</i> , 2019a]	0.612	0.443	0.556	0.116	0.474	0.541
	JHFL[Su <i>et al.</i> , 2019]	0.717	0.645	0.659	0.138	0.662	0.346
Ours	0.812	0.652	0.791	0.152	0.683	0.331	
MI3DOR-2	1NN	0.518	0.355	0.488	0.355	0.383	0.629
	CORAL[Sun <i>et al.</i> , 2016]	0.538	0.369	0.497	0.369	0.399	0.614
	GFK[Gong <i>et al.</i> , 2012]	0.513	0.471	0.495	0.471	0.484	0.527
	JGSA[Zhang and <i>et al.</i> , 2017]	0.584	0.475	0.598	0.475	0.501	0.511
	JAN[Long <i>et al.</i> , 2017]	0.533	0.416	0.545	0.416	0.445	0.568
	LSR[Xie <i>et al.</i> , 2018]	0.538	0.395	0.529	0.395	0.434	0.587
	JHFL[Su <i>et al.</i> , 2019]	0.645	0.593	0.601	0.593	0.612	0.395
	Ours	0.660	0.539	0.678	0.539	0.563	0.446

Table 1: Performance comparison on MI3DOR and MI3DOR-2.

4.2 Evaluation Criteria & Implementation Details

We adopted NN, FT, ST, DCG, F-Measure and ANMRR as the evaluation criteria, as described in [Liu *et al.*, 2018a]. These criteria range from 0 to 1, with the exception of ANMRR, the higher the better.

To obtain the rendered images for 3D objects, we set the cameras every 30 degrees around each 3D object and then obtain 12 view images for it. In our experiments, we used the AlexNet as CNN architecture in the MVCNN. The dimension of the original visual feature of 2D images and 3D objects is $d = 4096$. The dimension of subspace obtained by *PCA* is $d_1 = 512$, and the dimension of the latent space obtained by *P* is $d_2 = 128$. We set the number of iterations $T = 6$.

4.3 Comparison with State-of-the-Art Methods

We compared the proposed method with several representative methods: 1) 1NN means using the original visual representation of 2D images and 3D objects before the subspace generation to measure their similarity; 2) CORAL [Sun *et al.*, 2016] aligns the second-order statistical characteristics of the source and target domains; 3) GFK [Gong *et al.*, 2012] integrates unlimited number of subspaces to reduce domain shift; 4) JGSA [Zhang and *et al.*, 2017] aligns distributions by the label propagation; 5) JAN [Long *et al.*, 2017] aligns the distributions in a transfer network; 6) LSR [Xie *et al.*, 2018] aligns the centroids of labeled source domain and the centroids of pseudo-labeled target domain to learn semantic representations of the samples from the target domain; 7) JHFL [Su *et al.*, 2019] can adaptively align the distributions for 2D images and 3D objects.

The comparison is shown in Table 1. From the results, we can obtain four key observations.

1) Our method is better than representative methods since it contains consistent domain structure learning, which benefits the domain alignment. On MI3DOR, the proposed method can achieve the gain of 14.0%-222.1% in NN, 1.1%-337.8% in FT, 20.0%-233.7% in ST, 10.2%-253.6% in F-measure and 3.2%-321.6% in DCG, and decline 4.3%-60.9% in ANMRR.

On MI3DOR-2, the proposed method is better than the representative methods except JHFL. When compared with JHFL, the proposed method achieves better in NN and ST.

2) JGSA and JHFL are significantly better than CORAL and GFK. This reveals that aligning both the marginal distributions and conditional distributions is more important than only aligning the marginal distributions.

3) All methods except for 1NN are better than 1NN and it demonstrates there exists a large gap between 2D and 3D visual representations.

4) The proposed method is better than JAN and LSR. JAN and LSR require large scale data and tuning a large number of parameters to obtain the best results. In contrast, the proposed method can set fewer parameters with only cross-validation and a small amount of data.

4.4 Ablation Study

For the proposed method, the consistent domain structure learning module is a very important module. To evaluate the importance of the class centers from the source domain (marked, CCS) or the intrinsic structure of the target domain (marked as IST) in the consistent domain structure learning module, we design and conduct three comparison experiments: CCS, IST, CCS with IST (marked as CCS+IST).

In Table 2, we obtain two key observations: 1) The semantic information from the source domain plays a key role to predict soft label. Especially, CCS can significantly improve the performances when compared with ‘1NN’ in Table 1. 2) The IST can be as a complementary information for the CCS and benefits the retrieval performances.

4.5 Effects of Hyper-parameters

There are four hyper-parameters in the proposed method: the trade-off parameter μ for the intrinsic structure learning module, the dimensionality of subspace d_1 after PCA, the latent space d_2 obtained by the projection *P* and the number of iterations *T*. We evaluate them on MI3DOR and MI3DOR-2. The results are shown in Fig. 2.

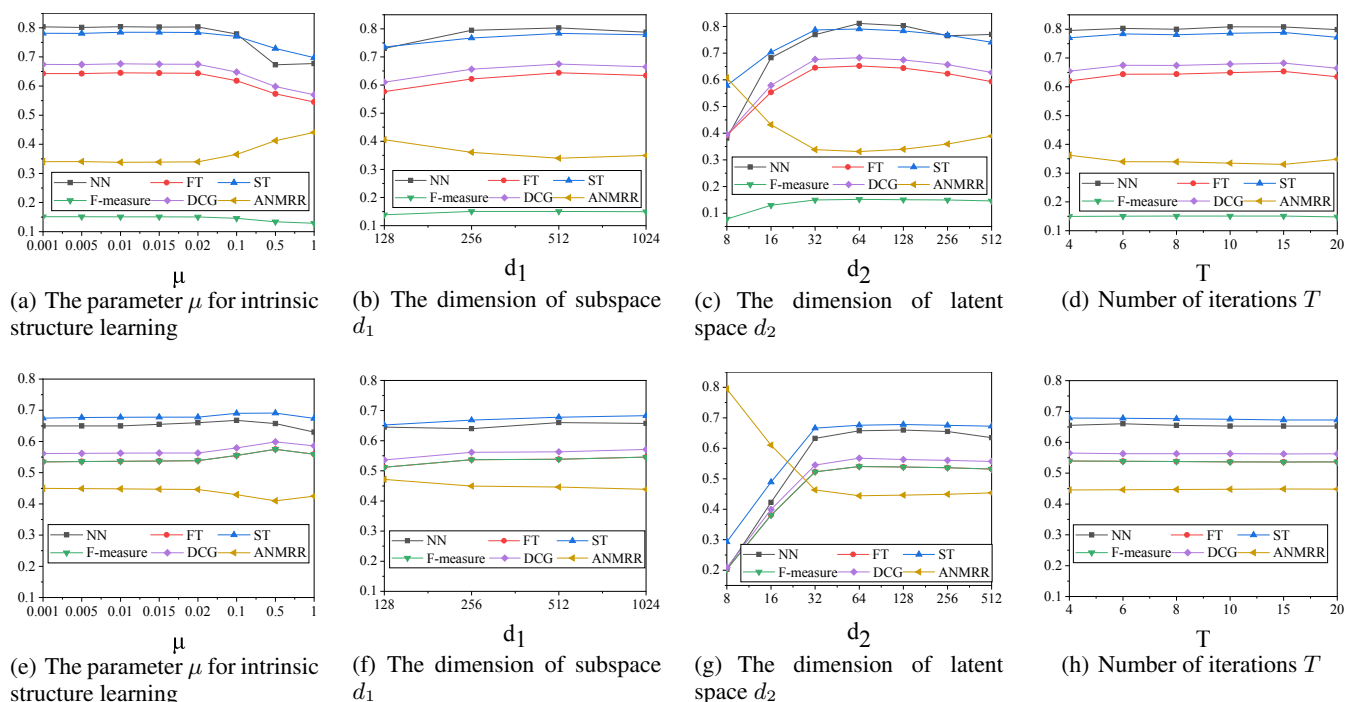


Figure 2: Sensitivity analysis on MI3DOR (a-d) and MI3DOR-2 (e-h).

Method	MI3DOR						MI3DOR-2					
	NN	FT	ST	F	DCG	ANMRR	NN	FT	ST	F	DCG	ANMRR
IST	0.657	0.533	0.680	0.127	0.558	0.453	0.548	0.491	0.597	0.491	0.516	0.495
CCS	0.806	0.649	0.786	0.152	0.681	0.334	0.650	0.534	0.675	0.534	0.561	0.450
CCS+IST	0.812	0.652	0.791	0.152	0.683	0.331	0.660	0.539	0.678	0.539	0.563	0.446

Table 2: Ablation study on MI3DOR and MI3DOR-2. F means F-measure.

μ is a very important parameter to the consistent domain structure learning module. As shown in Fig. 2 (a) and (e), the optimal μ is 0.01 for MI3DOR and 0.1 for MI3DOR-2. The d_1 and d_2 are also very important parameters for our task. On MI3DOR, as shown in Fig. 2 (b) and (c), the optimal d_1 and d_2 is 512 and 64, respectively. On MI3DOR-2, as shown in Fig. 2 (f) and (g), the optimal d_1 and d_2 is 512 and 128, respectively. In Fig. 2 (d) and (h), the performance gains slowly with iterations growing. The optimal T is 15 and 6 on MI3DOR and MI3DOR-2, respectively. To summarize, the proposed method is robust to these hyper-parameters except d_2 , which should be larger than the number of classes.

5 Conclusion

In this paper, we propose a consistent domain structure learning and domain alignment framework for 2D image-based 3D objects retrieval. It can reduce domain discrepancy and learn a latent feature space of a lower dimension. The consistent domain structure learning module explores both the semantic information of the source domain and the intrinsic structure of the target domain. It can provide reliable labels for the unlabeled target domain to domain alignment. The extensive experiments show that the proposed method achieve remark-

able results on two public datasets.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (61772359, 61872267, 61902277, 61572356), the grant of Tianjin New Generation Artificial Intelligence Major Program(19ZXZNGX00110, 18ZXZNGX00150, 17ZXRGGX00180), the Open Project Program of the State Key Lab of CAD & CG, Zhejiang University (Grant No. A2005, A2012).

References

- [Chang and et al., 2015] Angel X. Chang and et al. Shapenet: An information-rich 3d model repository. *CoRR*, abs/1512.03012, 2015.
- [Cheng et al., 2018] Zhiyong Cheng, Ying Ding, and et al. A³ncf: An adaptive aspect attention model for rating prediction. In *IJCAI*, pages 3748–3754, 2018.
- [Dai et al., 2017] Guoxian Dai, Jin Xie, Fan Zhu, and Yi Fang. Deep correlated metric learning for sketch-based 3d shape retrieval. In *AAAI*, pages 4002–4008, 2017.

- [Feng *et al.*, 2018] Yifan Feng, Zizhao Zhang, Xibin Zhao, Rongrong Ji, and Yue Gao. GVCNN: group-view convolutional neural networks for 3d shape recognition. In *CVPR*, pages 264–272, 2018.
- [Ghadai *et al.*, 2019] Sambit Ghadai, Xian Yeow Lee, Aditya Balu, Soumik Sarkar, and Adarsh Krishnamurthy. Multi-level 3d CNN for learning multi-scale spatial features. In *CVPR Workshops 2019*, page 0, 2019.
- [Gong *et al.*, 2012] Boqing Gong, Yuan Shi, and Fei Sha *et al.* Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, pages 2066–2073, 2012.
- [Gopalan *et al.*, 2011] Raghuraman Gopalan, Ruonan Li, and *et al.* Domain adaptation for object recognition: An unsupervised approach. In *ICCV*, pages 999–1006, 2011.
- [Gretton *et al.*, 2012] Arthur Gretton, Karsten M. Borgwardt, and *et al.* A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, 2012.
- [He and Niyogi, 2003] Xiaofei He and Partha Niyogi. Locality preserving projections. In *NIPS*, pages 153–160, 2003.
- [He *et al.*, 2018] Xinwei He, Yang Zhou, and *et al.* Triplet-center loss for multi-view 3d object retrieval. In *CVPR*, pages 1945–1954, 2018.
- [Hong *et al.*, 2016] Richang Hong, Zhenzhen Hu, and *et al.* Multi-view object retrieval via multi-scale topic models. *IEEE Trans. Image Processing*, 25(12):5814–5827, 2016.
- [Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and *et al.* Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1106–1114, 2012.
- [Lee *et al.*, 2018] Tang Lee, Yen-Liang Lin, and *et al.* Cross-domain image-based 3d shape retrieval by view sequence learning. In *2018 International Conference on 3D Vision*, pages 258–266, 2018.
- [Li *et al.*, 2019a] Wenhui Li, Anan Liu, and Weizhi Nie *et al.* Monocular image based 3d model retrieval. In *3DOR*, pages 103–110, 2019.
- [Li *et al.*, 2019b] Zhaoqun Li, Cheng Xu, and Biao Leng. Angular triplet-center loss for multi-view 3d shape retrieval. In *AAAI*, pages 8682–8689, 2019.
- [Liu *et al.*, 2016] Anan Liu, Weizhi Nie, Yue Gao, and Yuting Su. Multi-modal clique-graph matching for view-based 3d model retrieval. *IEEE Trans. Image Processing*, 25(5):2103–2116, 2016.
- [Liu *et al.*, 2018a] An-An Liu, Weizhi Nie, Yue Gao, and Yuting Su. View-based 3-d model retrieval: A benchmark. *IEEE Trans. Cybernetics*, 48(3):916–928, 2018.
- [Liu *et al.*, 2018b] Anan Liu, Shu Xiang, Wenhui Li, Weizhi Nie, and Yuting Su. Cross-domain 3d model retrieval via visual domain adaptation. In *IJCAI*, pages 828–834, 2018.
- [Long *et al.*, 2017] Mingsheng Long, Han Zhu, and *et al.* Deep transfer learning with joint adaptation networks. In *ICML*, pages 2208–2217, 2017.
- [Lu *et al.*, 2019a] Xu Lu, Lei Zhu, and *et al.* Flexible online multi-modal hashing for large-scale multimedia retrieval. In *ACMMM 2019*, pages 1129–1137, 2019.
- [Lu *et al.*, 2019b] Xu Lu, Lei Zhu, and *et al.* Online multi-modal hashing with dynamic query-adaption. In *ACM SIGIR*, pages 715–724, 2019.
- [Phong, 1975] Bui Tuong Phong. Illumination for computer generated pictures. *Commun. ACM*, 18(6):311–317, 1975.
- [Qi *et al.*, 2017] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NIPS*, pages 5099–5108, 2017.
- [Shen *et al.*, 2018] Yiru Shen, Chen Feng, Yaoqing Yang, and Dong Tian. Mining point cloud local structures by kernel correlation and graph pooling. In *CVPR*, pages 4548–4557, 2018.
- [Su *et al.*, 2015] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik G. Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *ICCV*, pages 945–953, 2015.
- [Su *et al.*, 2019] Yu-Ting Su, Yu-Qian Li, and *et al.* Joint heterogeneous feature learning and distribution alignment for 2d image-based 3d object retrieval. *IEEE Trans. Circuits Syst. Video Techn.*, 2019.
- [Sun *et al.*, 2016] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *AAAI*, pages 2058–2065, 2016.
- [Wang and Breckon, 2019] Qian Wang and Toby P. Breckon. Unsupervised domain adaptation via structured prediction based selective pseudo-labeling. *CoRR*, abs/1911.07982, 2019.
- [Wang and Chen, 2017] Qian Wang and Ke Chen. Zero-shot visual recognition via bidirectional latent embedding. *IJCV*, 124(3):356–383, 2017.
- [Wang *et al.*, 2015] Fang Wang, Le Kang, and Yi Li. Sketch-based 3d shape retrieval using convolutional neural networks. In *CVPR*, pages 1875–1883, 2015.
- [Wang *et al.*, 2018] Jindong Wang, Wenjie Feng, and Yiqiang Chen *et al.* Visual domain adaptation with manifold embedded distribution alignment. In *MM*, pages 402–410, 2018.
- [Wu and *et al.*, 2015] Zhirong Wu and Shuran Song *et al.* 3d shapenets: A deep representation for volumetric shapes. In *CVPR*, pages 1912–1920, 2015.
- [Xie *et al.*, 2017] Jin Xie, Guoxian Dai, Fan Zhu, Edward K. Wong, and Yi Fang. Deepshape: Deep-learned shape descriptor for 3d shape retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(7):1335–1345, 2017.
- [Xie *et al.*, 2018] Shaoan Xie, Zibin Zheng, and *et al.* Learning semantic representations for unsupervised domain adaptation. In *ICML*, pages 5419–5428, 2018.
- [Zhang and *et al.*, 2017] Jing Zhang and *et al.* Joint geometrical and statistical alignment for visual domain adaptation. In *CVPR*, pages 5150–5158, 2017.
- [Zhang *et al.*, 2019] Lei Zhang, Shanshan Wang, and *et al.* Manifold criterion guided transfer learning via intermediate domain generation. *CoRR*, abs/1903.10211, 2019.