Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)

# Cross-Domain 3D Model Retrieval via Visual Domain Adaptation

**Anan Liu, Shu Xiang, Wenhui Li*, Weizhi Nie** and **Yuting Su**
School of Electrical and Information Engineering, Tianjin University, China
liwenhui@tju.edu.cn

## Abstract

Recent advances in 3D capturing devices and 3D modeling software have led to extensive and diverse 3D datasets, which usually have different distributions. Cross-domain 3D model retrieval is becoming an important but challenging task. However, existing works mainly focus on 3D model retrieval in a closed dataset, which seriously constrain their implementation for real applications. To address this problem, we propose a novel cross-domain 3D model retrieval method by visual domain adaptation. This method can inherit the advantage of deep learning to learn multi-view visual features in the data-driven manner for 3D model representation. Moreover, it can reduce the domain divergence by exploiting both domain-shared and domain-specific features of different domains. Consequently, it can augment the discrimination of visual descriptors for cross-domain similarity measure. Extensive experiments on two popular datasets, under three designed cross-domain scenarios, demonstrate the superiority and effectiveness of the proposed method by comparing against the state-of-the-art methods. Especially, the proposed method can significantly outperform the most recent method for cross-domain 3D model retrieval and the champion of Shrec'16 Large-Scale 3D Shape Retrieval from ShapeNet Core55.

## 1 Introduction

The rapid development of 3D techniques for modeling, reconstruction, printing has led to huge deluge of 3D content. 3D model retrieval is becoming mandatory in diverse domains, such as e-business, digital entertainment, medical diagnosis and education[Liu *et al.*, 2017; Cheng *et al.*, 2017; Nie *et al.*, 2016; Tang *et al.*, 2017; He *et al.*, 2017]. Especially, effective methods for cross-domain 3D model retrieval play an important role on real applications in virtual and augmented reality, shape completion and scene synthesis. It has become one of the hot research topics in both computer vision and machine learning.

* Corresponding author

### 1.1 Motivations

3D model retrieval aims to search the relevant candidates from the assigned dataset given a query 3D model. Although much work has been done for 3D model retrieval, there still exist two critical problems:

**1) How to make good use of the current small-scale 3D model datasets to augment the generalization of algorithms.** Compared with millions of 2D image datasets, e.g. ImageNet and MSCOCO, the current 3D model datasets only contain limited samples. The most recent 3D datasets, such as ModelNet40 [Wu *et al.*, 2015] and ShapeNetCore55 [Chang *et al.*, 2015], only contain 12311 and 51300 models, respectively. Although the deep learning methods, e.g. Multi-View Convolutional Neural Network (MVCNN) [Su *et al.*, 2015] which won the first prize of Shrec'16 Large-Scale 3D Shape Retrieval from ShapeNet Core55, achieved significant improvement for the task under the identical-domain scenario, they cannot work well for the real applications when the source and target come from different domains. Theoretically, deep learning is highly dependent on big data. In essence, the current deep learning methods can be regarded as overfitting with respect to individual datasets. Therefore, it is necessary to develop sophisticated methods of visual domain adaptation to integrate these small-scale datasets and improve the generalization of 3D model retrieval methods.

**2) How to retrieve 3D models from different datasets with diverse data distributions.** In the past few years, multiple 3D modeling devices have been widely applied in human life. Diverse 3D datasets have been released for the research on 3D understanding. For example, there are multiple RGB-D data captured in real world by depth sensors, e.g. Microsoft Kinect, Intel RealSense. Meanwhile, there are many new 3D datasets, such as ShapeNet and 3D warehouse, which consist of 3D CAD models. For the real applications, the target and source 3D models usually come from different datasets, even different modalities. 3D models may have different visual and structural information even though they belong to the same category. However, there are limited works to address the challenging task of cross-domain 3D model retrieval.

To handle the problems mentioned above, we propose a novel cross-domain 3D model retrieval method via visual domain adaptation as shown in Fig. 1. First, MVCNN is utilized to extract the visual features for the multi-view images of each 3D model. Then, visual domain adaptation is imple-

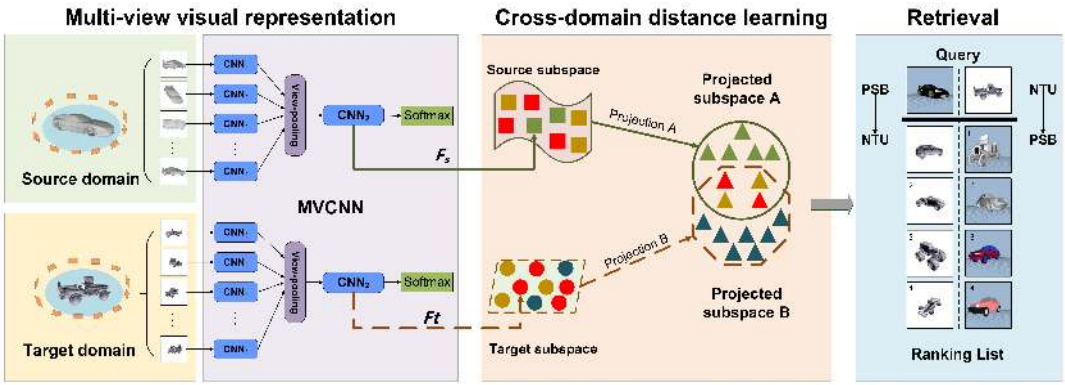



Figure 1: The framework of cross-domain 3D model retrieval via visual domain adaptation. Each 3D model is firstly represented by a set of multi-view 2D images. The image set is passed to the $CNN_1$ layer, each branch of which has the same architecture as AlexNet. The outputs of the $CNN_1$ layer are synthesized by the view-pooling layer. The output of this layer is passed through the $CNN_2$ layer to obtain the compact visual feature for individual 3D model. The output of $CNN_2$ is fed into the module of cross-domain distance learning for feature projection and similarity measure. The shape of each subspace indicates the specific geometrical and statistical distribution. The same color represents the 3D models from the identical category. After visual domain adaptation, the statistical and geometrical discrepancy of two domains can be explicitly reduced and both domain-shared and domain-specific characteristics can be preserved. Then cross-domain retrieval can be conducted with the projected subspaces.

mented to reduce the diverse data distributions of different datasets for cross-domain retrieval. Different from most of traditional domain adaptation methods, our method is free of the strong assumption that there must exist a common space between different datasets. The assumption cannot always hold especially for 3D model datasets, which usually consists of extremely complicated characteristics, caused by multiple views, multiple modalities, complex spatial structures, and so on. This method can learn two coupled projections to map the source and target data into the respective subspaces and reduce domain shift. The transformed visual features by preserving both domain-shared and domain-specific characteristics can augment the discrimination and robustness for the cross-domain task. The proposed method is verified on two popular datasets, the National Taiwan University 3D Model database (NTU) [Chen *et al.*, 2003] and the Princeton Shape Benchmark (PSB) [Shilane *et al.*, 2004], under three cross-domain scenarios. The extensive experimental results show that the proposed method can significantly outperform the state of the arts, especially the most recent method for cross-domain 3D model retrieval and the champion of Shrec'16 Large-Scale 3D Shape Retrieval from ShapeNet Core55.

### 1.2 Contribution

The main contributions are summarized as follows:

- This paper proposes a novel framework for multi-view representation and visual domain adaptation in the unsupervised manner. It can inherit the advantage of deep learning to learn multi-view visual features and can reduce the geometrical and statistical domain divergence to benefit cross-domain retrieval.
- Different from most of current methods, which works on retrieval in the closed set, this paper focuses on open-domain problem. It can make good use of small-scale 3D model datasets to augment the generation ability of algorithms.

## 2 Related Work

This section will review the recent progress in the fields of both 3D model retrieval and domain adaptation.

### 2.1 3D Model Retrieval

Generally, the existing 3D model retrieval methods can be grouped into two types, model-based methods and view-based methods. Model-based methods usually utilize the voxel grid and RGB-D point cloud for 3D model representation. Wu et al. [Wu *et al.*, 2015] represented a 3D model as binary voxel probability distribution. The 3D binary voxel grid can be trained by the CNN framework. Li et al. [Li *et al.*, 2016] represented a 3D model by volumetric fields and replaced the convolutional layer in CNN with Field Probing Filter, which can overcome the sparse problem of 3D voxels. A similar approach is VoxNet [Maturana and Scherer, 2015], which leverages binary voxel grids and supervised Convolutional Neural Network. The advantage of this kind of method is that the three-dimensional voxel completely retains the three-dimensional shape information, which is beneficial to improve the distinguishing ability.

View-based methods usually utilize a set of multi-view images for 3D model representation. Hence, the advanced image processing and machine learning techniques can be utilized for this task. Multi-view Convolutional Neural Networks [Su *et al.*, 2015] was proposed to process all rendered views and utilize view-pooling layer to combine convolutional features. Kalogerakis et al. [Kalogerakis *et al.*, 2016] captured a series of shadow and depth maps of 3D shapes, Fully Convolutional Networks (FCN) was then employed to learn shape descriptors. RotationNet [Kanezaki, 2016] took multiple views as inputs and estimate its pose and category. The discovered pose and category information can benefit augmenting the performance of 3D model retrieval.

### 2.2 Domain Adaptation

Generally, domain adaptation can be divided into two categories: semi-supervised domain adaptation by using both the

labeled and unlabeled data in the target domain, and unsupervised domain adaptation by only using unlabeled data in the target domain. Semi-supervised HFA [Li *et al.*, 2014] can simultaneously learn the target classifier as well as infer the labels of unlabeled target samples. Unsupervised domain adaptation is considered to be more challenging. For instance, Transfer Component Analysis [Pan *et al.*, 2011] minimized the discrepancy between the instance of the source and target data when they were projected into a K-dimensional embedding space. Joint distribution analysis [Long *et al.*, 2013] took the marginal distribution as well as conditional distribution into consideration. Zhang's work [Zhang *et al.*, 2015] utilized causal models to represent the relationship between the feature $X$ and label $Y$, and consider possible situations where different modules of the causal model change with the domain.

Although many sophisticated domain adaptation methods have been developed, seldom of them have been applied for 3D model retrieval. Hong's work MSTM[Hong *et al.*, 2016] can be regarded as the first work towards cross-domain 3D model retrieval. MSTM proposed a multi-scale topic models for this task. MSTM leveraged cross-domain learning directly, while ignoring the discrimination of each domain.

# 3 Methods

## 3.1 Overview

This paper aims to address the cross-domain 3D model retrieval: the source models and the target models are drawn from different datasets, which have unknown and discrepant data distributions. The proposed framework is shown in Fig.1. It consists of two successive steps:

**1) Multi-view visual representation:** For the view-based methods, a 3D model is usually represented by a set of views captured from different directions as shown in Fig. 1. Then, the popular Multi-view Convolutional Neural Network (MVCNN), which can jointly learn the visual and spatial structural characteristics of each category of 3D models, is adopted to generate a discriminative and compact descriptor for individual 3D model. Section 3.2 will illustrate this step.

**2) Cross-domain distance learning:** When pair-wise 3D models come from the same dataset, their similarity can be directly computed based on specific metrics with the extracted visual features. However, for the cross-domain retrieval tasks, it is mandatory to project the visual features from different domains into specific subspaces to reduce domain divergence before similarity measure. To tackle this problem, we present the visual domain adaptation strategy by jointly reducing the statistical and geometrical discrepancy of different domains in an unsupervised manner. Section 3.3 will detail this step. Then we can measure the similarity between two 3D models by computing their Euclidean distance, as most representative methods do, with the transformed visual features after domain adaptation.

## 3.2 Multi-View Visual Representation

To transform each 3D model into a set of images, Phong reflection model [Phong, 1975] is used to capture and render multiple views of 3D models. As most of related works, we created 12 views by placing 12 virtual cameras around the model every 30 degrees. With the image set representation, the popular MVCNN (Fig. 1) is implemented to generate a compact descriptor for individual 3D models. First, the image set are inputted into MVCNN. Individual images pass throughout the convolutional layers ($\text{CNN}_1$), and are synthesized at the view-pooling layer. All branches in $\text{CNN}_1$ share identical architecture as AlexNet and the same parameters. Then they are operated by $\text{CNN}_2$. The view-pooling layer is placed after $\text{conv}_5$ layer, taking element-wise maximum operation over multiple views. In our work, MVCNN is trained on ModelNet40 that is the most popular 3D model dataset and contains the common 20 categories in NTU and PSB. We apply this MVCNN model to extract the descriptors of the source and target 3D models. Especially, the output of $\text{fc}_7$ (4096-D) is used as visual feature.

| Variables | Definition |
|---|---|
| $\mathcal{D}_s = \{(x_i, y_i)\}_{i=1}^{n_s}$ | labeled source domain data, $X_s = \{x_i\}_{i=1}^{n_s} \in \mathbb{R}^{D \times n_s}$ comes from distribution $P_s(X_s)$, $y_i$ is the label |
| $\mathcal{D}_t = \{x_j\}_{j=1}^{n_t}$ | unlabeled target domain data, $X_t = \{x_j\}_{j=1}^{n_t} \in \mathbb{R}^{D \times n_t}$ is drawn from distribution $P_t(X_t)$. |

Table 1: Definition of terminologies. Note $P_s(X_s) \neq P_t(X_t)$ for the cross-domain problem.

## 3.3 Cross-Domain Distance Learning

In this section, we will detail the method of cross-domain distance learning for 3D model retrieval. Motivated by [Zhang *et al.*, 2017], we well adapt the popular Joint Geometrical and Statistical Alignment method for visual domain adaptation. The definition of terminologies are summarized in Table 1. Since the data shift from different datasets (e.g. NTU and PSB) is large, we aim to find two coupled projections (A for source domain, and B for target domain) to reduce shifts between respective domains. An ideal objective function should have the following properties: 1) maximizing the variance of target domain, $VAR_{Target}$; 2) preserving the discriminative information of source domain, which can be represented by inner-class variance, $VAR_{Inner}$, and inter-class variance, $VAR_{Inter}$; 3) minimizing the divergence of source and target distributions, $DIV_{distribution}$; 4)minimizing the divergence between source and target subspaces, $DIV_{Subspace}$. Consequently, the objective function can be formulated, by incorporating all these factors, as follows:

$$max \frac{\mu\{VAR_{Target}\} + \beta\{VAR_{Inter}\}}{\{DIV_{distribution}\} + \lambda\{DIV_{Subspace}\} + \beta\{VAR_{Inner}\}} \quad (1)$$

where $\lambda$, $\mu$, $\beta$ are weighted parameters.

The optimization of the objective function in Eq. 1 has two goals: 1) maximizing the numerator to increase the variance of the target domain and the inter-class variance of the source domain; 2) minimizing the denominator to decrease the domain distribution shifts, and the inner-class variance of the source domain. We will detail the formulation of each term as follows:

**Target Variance**

Maximization of Target Variance can ensure projecting features into relevant dimensions. $VAR_{Target}$ can be formulated as:

$$VAR_{Target} = Tr\left(B^T S_t B\right) \tag{2}$$

where $S_t = X_t H_t X_t^T$ is the target domain scatter matrix. $H_t = I_t - \frac{1}{n_t} 1_t 1_t^T$ is the centering matrix, $1_t \in \mathbb{R}^{n_t}$ is the column vector with all ones.

**Source Discriminative Information**

Preserving the source discriminative information can be realized by maximizing inter-class variance and minimizing inner-class variance. $VAR_{Inter}$ and $VAR_{Inner}$ are defined as follows:

$$VAR_{Inter} = Tr\left(A^T S_b A\right) \tag{3}$$

$$VAR_{Inner} = Tr\left(A^T S_w A\right) \tag{4}$$

where $S_w = \sum_{c=1}^{C} X_s^{(c)} H_s^{(c)} \left(X_s^{(c)}\right)^T$ is the inner-class scatter matrix, $S_b = \sum_{c=1}^{C} n_s^{(c)} \left(m_s^{(c)} - \bar{m}_s\right)\left(m_s^{(c)} - \bar{m}_s\right)^T$ is the inter-class scatter matrix of the source models, where $X_s^{(c)}$ is the set of 3D models belonging to class c, $m_s^{(c)} = \frac{1}{n_s^{(c)}} \sum_{i=1}^{n_s^{(c)}} x_i^{(c)}$, $\bar{m}_s = \frac{1}{n_s} \sum_{i=1}^{n_s} x_i$, $H_s^{(c)} = I_s^{(c)} - \frac{1}{n_s^{(c)}} 1_s (1_s^{(c)})^T$ is the centering matrix of data within class $c$, $I_s^{(c)} \in \mathbb{R}^{n_s^{(c)} \times n_s^{(c)}}$ is the identity matrix, $1_s \in \mathbb{R}^{n_s^{(c)}}$ is the column vector with all ones, $n_s^{(c)}$ is the number of target models in class $c$.

**Distribution Divergence**

Motivated by [Long *et al.*, 2013], the distribution divergence can be formulated as follows, considering both marginal and conditional distribution shift:

$$\min_{A,B} Tr\left(\begin{bmatrix} A^T B^T \end{bmatrix} \begin{bmatrix} M_s & M_{st} \\ M_{ts} & M_t \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix}\right) \tag{5}$$

where

$$M_s = X_s \left(L_s + \sum_{c=1}^{C} L_s^{(c)}\right) X_s^T, \quad L_s = \frac{1}{n_s^2} 1_s 1_s^T,$$
$$\left(L_s^{(c)}\right)_{ij} = \begin{cases} \frac{1}{\left(n_s^{(c)}\right)^2} & x_i, x_j \in X_s^{(c)} \\ 0 & otherwise \end{cases} \tag{6}$$

$$M_{st} = X_s \left(L_{st} + \sum_{c=1}^{C} L_{st}^{(c)}\right) X_t^T, \quad L_{st} = -\frac{1}{n_s n_t} 1_s 1_t^T,$$
$$\left(L_{st}^{(c)}\right)_{ij} = \begin{cases} -\frac{1}{n_s^{(c)} n_t^{(c)}} & x_i \in X_s^{(c)}, x_j \in X_t^{(c)} \\ 0 & otherwise \end{cases} \tag{7}$$

Note $M_t$ can be computed by replace $s$ with t in Eq.6 and $M_{ts}$ can be computed by alternate $s$ and $t$ in Eq.7.

**Subspace Divergence**

To preserve both source class information and target variance, A and B should be optimized simultaneously. Consequently the distance between two subspaces can be reduced. Subspace Divergence can be computed as:

$$\min_{A,B} \|A - B\|_F^2 \tag{8}$$

With the definition of these four critical factors, the objective function can be formulated by:

$$\max_{W} \frac{Tr\left(W^T \begin{bmatrix} \beta S_b & \mathbf{0} \\ \mathbf{0} & \mu S_t \end{bmatrix} W\right)}{Tr\left(W^T \begin{bmatrix} M_s + \lambda I + \beta S_\omega & M_{st} - \lambda I \\ M_{ts} - \lambda I & M_t + (\lambda + \mu) I \end{bmatrix} W\right)} \tag{9}$$

where $W^T = \begin{bmatrix} A^T & B^T \end{bmatrix}$. Because rescaling of $W$ will not affect the optimization of the objective function, we can utilize the denominator as constraint and further rewrite the Eq.9 in the Lagrange function format:

$$\begin{aligned} L = & Tr\left(W^T \begin{bmatrix} \beta S_b & \mathbf{0} \\ \mathbf{0} & \mu S_t \end{bmatrix} W\right) \\ & + Tr\left(\left(W^T \begin{bmatrix} M_s + \lambda I + \beta S_\omega & M_{st} - \lambda I \\ M_{ts} - \lambda I & M_t + (\lambda + \mu) I \end{bmatrix} W - I\right) \Phi\right) \end{aligned} \tag{10}$$

By setting the derivative $\frac{\partial L}{\partial W} = 0$, we get:

$$\begin{bmatrix} \beta S_b & \mathbf{0} \\ \mathbf{0} & \mu S_t \end{bmatrix} W = \begin{bmatrix} M_s + \lambda I + \beta S_\omega & M_{st} - \lambda I \\ M_{ts} - \lambda I & M_t + (\lambda + \mu) I \end{bmatrix} W \Phi \tag{11}$$

where $\Phi = diag\left(\lambda_1, ..., \lambda_k\right)$ are the k leading eigenvalues and $W = [W_1, ..., W_K]$ contains the corresponding eigenvectors, which can be solved analytically by generalized eigenvalue decomposition. Then two subspaces can be computed with the transformation matrix W.

# 4 Experimental Results and Discussion

In this section, we describe the experimental settings, evaluation criteria, and discuss the experimental results.

## 4.1 Experimental Settings

**Implementation Details**

Two popular 3D model datasets with diverse data distribution are utilized for evaluation, the same as the most recent work [Hong *et al.*, 2016] for cross-domain 3D model retrieval. The National Taiwan University (NTU) 3D model dataset contains 549 3D models from 46 categories. Princeton Shape Benchmark (PSB) consists of 1,814 models from 161 categories. There are 20 common categories, such as bike, car, chair, bookshelf, etc., in NTU and PSB, including 226 and 275 models, respectively. Additionally, the models in NTU and PSB are split into two subsets: training and test, with ratio 50% and 50% respectively.

The experiment was conducted under three cross-domain scenarios: 1) NTU→PSB: the target comes from NTU and the source comes from PSB ; 2) PSB→NTU: the target comes from PSB and the source comes from NTU; 3) NTU←→PSB: we synthesize a new dataset by mixing the common categories of NTU and PSB; both target and source are from this mixture.

### Evaluation Criteria

The following popular criteria are employed for evaluation. These criteria range from 0 to 1. The higher value means the better performance except ANMRR. The lower ANMRR indicates the better performance.

Nearest neighbor (NN): the precision of the first retrieved model.

First tier (FT): the recall for the first $K$ relevant samples, where $K$ is the cardinality of the target category.

Second tier (ST): the recall for the first $2K$ relevant match samples.

F-measure: a synthetical measurement of precision and recall of the top retrieved results. The top 20 retrieval results are used in our experiments.

Discounted Cumulative Gain (DCG): a statistical measure that assigns relevant results at the top ranking positions with higher weights under the assumption that a user is less likely to consider lower results.

Average Normalized Modified Retrieval Rank (ANMRR): it considers the ranking information of relevant models among the retrieved models.

Area Under Curve (AUC): AUC can simultaneously evaluate the performance of both precision and recall in the PR curve.

### Competing Methods

We compare the proposed method against MVCNN, which is the champion of Shrec'16 Large-Scale 3D Shape Retrieval from ShapeNet Core55, and MSTM [Hong *et al.*, 2016], the most recent cross-domain method for 3D model retrieval. Moreover, the proposed method is compared against several representative methods, including: 1) distance-based methods: Hausdorff (HAUS) [Liu *et al.*, 2017] and SumMin [Gao and Dai, 2014]; 2) model-based method: Extension Ray-based Descriptor (ERD) [Vranic, 2003], Adaptive Views Clustering (AVC) [Ansary *et al.*, 2007], Elevation Descriptor (ED) [Shih *et al.*, 2007], Bag-of-Visual-Features (BoVF) [Ohbuchi *et al.*, 2008]; 3) domain adaptation method: Heterogeneous Feature Augmentation (HFA) [Li *et al.*, 2014] 4) graph matching-based methods: Weighted Bipartite Graph Matching (WBGM) [Gao *et al.*, 2011], Hypergraph Analysis (HA) [Gao *et al.*, 2012]; 5) deep learning methods: Learning Multi-view Deep Features (LMDF) [Guo *et al.*, 2015].

## 4.2 Comparison against the State of the Arts

The comparison under three cross-domain scenarios are detailed as follows.

1) NTU→PSB: From Fig. 2, it is obvious that the proposed method (VDA) can outperform all competing methods. Specifically, VDA can achieve the gain of 15.10%-178.60%, 62.63%-403.70%, 24.84%-263.90%, 33.43%-168.80%, 49.77%-379.10% in terms of NN, FT, ST, F-measure, and DCG, with the decline of 111.90%-241.7% in terms of ANMRR compared with HAUS, SumMin, ERD, AVC, ED, BoVF, WBGM, HA, LMDF, HFA, MSTM and MVCNN respectively.

2) PSB→NTU: The results of PSB→NTU are shown in Fig. 3. VDA can obtain the best performances with the gain of 9.10%-120.94%, 63.78%-432.13%, 21.70%-260.67%, 24.99%-184.04%, 46.35%-288.55%, in terms of NN, FT, ST, F-measure, and DCG, and 147.13%-432.8% decline in terms of ANMRR.

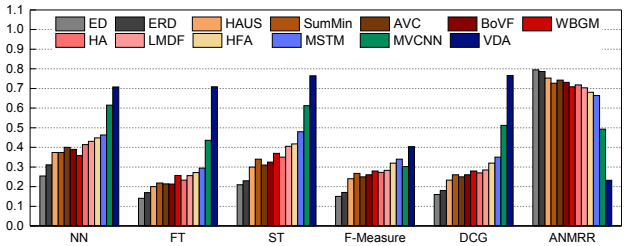



Figure 2: Performance for NTU→PSB.

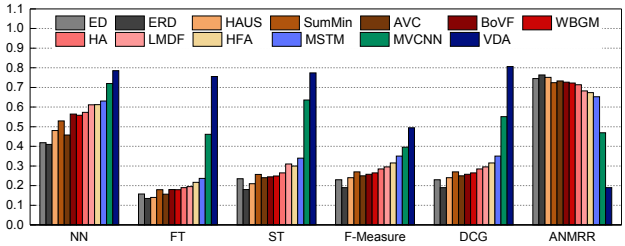



Figure 3: Performance for PSB→NTU.

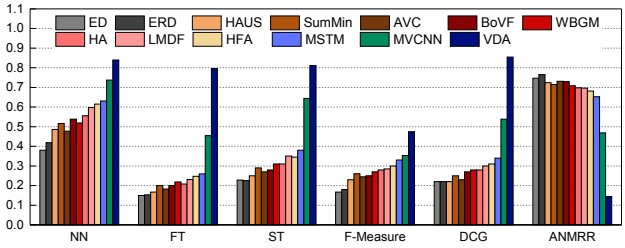



Figure 4: Performance for NTU⟷PSB.

3) NTU⟷PSB: The results of NTU⟷PSB are shown in Figure 4. VDA can obtain the best performances with the gain of 13.96%-120.94%, 75.23%-432.13%, 26.11%-260.67%, 34.53%-184.04%, 58.80%-288.55%, in terms of NN, FT, ST, F-measure, and DCG, and 225.61%-432.8% decline in terms of ANMRR.

According to the experimental results, we have several key observations:

1) VDA vs. MSTM: MSTM is the most recent method, which directly implemented cross-domain learning for this task. By comparison, VDA can significantly outperform MSTM: a) NTU→PSB: VDA can achieve the gains of 52.91%, 140.71%, 59.19%, 18.59%, 119.01% in terms of NN, FT, ST, F-measure, DCG and 185.35% decline in terms of ANMRR, comparing against MSTM; b) PSB→NTU: VDA can achieve the gain of 24.48%, 218.99%, 127.61%, 41.27%, 130.35% in terms of NN, FT, ST, F-measure, and DCG, and 243.94% decline in terms of ANMRR, comparing with MSTM; c) NTU⟷PSB: VDA can achieve the gain of 33.12%, 206.30%, 113.55%, 43.74%, 175.75% in terms of NN, FT, ST, F-measure, and DCG, and 354.32% decline in ANMRR, comparing with MSTM. MSTM represents each 3D model with a set of bag of topics discovered by the topic model. Then it conducts topic clustering for the basic topics from two datasets and then generates the common topic dic-

tionary for new representation. In this way, the two models from different datasets can be aligned to the same common feature space for comparison. The key problem of MSTM is that assuming that there exists a unified transformation to map two domains into one common domain, which cannot always hold especially when the dataset shift is large. Moreover, MSTM loses the discriminative information of individual domains. Comparatively, the proposed method is independent of this assumption and aims to learn two projections to map each dataset into individual subspaces. The projected subspaces can preserve both domain-shared and domain-specific characteristics, which can benefit cross-domain 3D model retrieval.

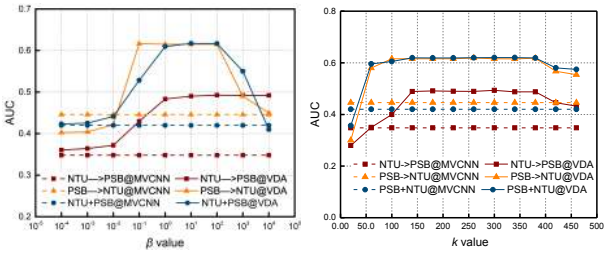
2) VDA vs. MVCNN: MVCNN is commonly regarded as one of the best methods for this task since it can leverage deep learning for data-driven feature learning, which can have better generalization ability. When the visual domain adaptation module is implemented after feature extraction by MVCNN, VDA can significantly outperform MVCNN: a) NTU→PSB: VDA can achieve the gain of 15.10%, 62.60%, 24.84%, 33.43%, 49.77% in terms of NN, FT, ST, F-measure, DCG and 111.90% decline in terms of ANMRR, comparing against MVCNN; b) PSB→NTU: VDA can achieve the gain of 9.10%, 63.78%, 21.70%, 24.99%, 46.35% in terms of NN, FT, ST, F-measure, and DCG, and 147.13% decline in terms of ANMRR, comparing with MVCNN; c) NTU←→PSB: VDA can achieve the gain of 13.96%, 75.23%, 26.11%, 34.53%, 58.80% in terms of NN, FT, ST, F-measure, and DCG, and 225.61% decline in ANMRR, comparing with MVCNN. This demonstrates that the VDA module can benefit reducing the divergence between different domains by heterogeneous information mapping. Consequently, the transformed features after visual domain adaptation can be more suited for similarity measure for cross-domain data.

3) VDA vs. others: By comparison, it is obvious that VDA can consistently outperform other competing methods under all three scenarios. The proposed method has two main advantages: a) VDA inherits the advantage of MVCNN to generate the discriminative visual descriptors for 3D models by leveraging multi-view information. b) VDA can reduces the domain divergence by exploiting both domain-shared and domain-specific features. It can further augment the discrimination and robustness of visual representation for cross-domain 3D model retrieval, comparing against these traditional methods. Therefore, VDA can outperform all the distance-based, model-based, and graph matching-based methods, which only utilized the hand-crafted visual features without visual domain adaptation.

### 4.3 Sensitivity Study

For sensitivity study, we vary the weights of Source Discriminative Information ($\beta$) and the subspace dimension ($k$) within different ranges to study how they affect the performance. The evaluation shows that $\lambda$ and $\mu$ cannot have significant influence on performances. We experimentally fix $\lambda$=1, and $\mu$=1. We directly compare the proposed method against the second best method, MVCNN, in our evaluation. Fig. 5 shows the results on three cross-domain scenarios.

We vary $\beta$ to study how source discriminative information affects the performance by fixing $k$ with the optimal value. From Fig. 5(a), $\beta$ can achieve stable results under three scenarios within a wide range of [1,$10^2$] and significantly outperform MVCNN within a wide range of [$10^{-4}$,$10^4$]. It demonstrates that VDA is robust with respect to $\beta$.




(a) $\beta$ value (b) $k$ value

Figure 5: Sensitivity study on different types of datasets.

We vary $k$ to study how the subspace dimension affects the performance by fixing $\beta$ with the optimal value. From Fig. 5(b), $k$ can achieve stable results under three scenarios within a wide range of [140, 380], and significantly outperform MVCNN within a wide range of [60,460]. If $k$ is too small or too big, the coupled projections may lose much discriminative information and consequently degrade the performance. It demonstrates that VDA is robust with respect to $k$.

## 5 Conclusion

This paper proposes a novel framework for multi-view representation and visual domain adaptation in the unsupervised manner. The proposed method can inherit the advantages of deep learning to generate the visual features. Moreover, considering the challenge by diverse divergence of different datasets, it can reduce both geometrical and statistical shifts and preserve domain-shared and domain-specific features. Consequently, it can augment the discrimination and robustness of visual representation for cross-domain retrieval. This method were validated on two popular datasets, under three retrieval scenarios, comparing with several state-of-the-art methods. In our future work, we will implement feature learning and domain adaptation in an end-to-end deep learning framework, which can inherit more informative data and more practical for real applications in Artificial intelligence.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (61772359, 61472275, 61502337).

## References

[Ansary *et al.*, 2007] Tarik Filali Ansary, Mohamed Daoudi, and Jean-Philippe Vandeborre. A bayesian 3-d search engine using adaptive views clustering. *IEEE Trans. Multimedia*, 9(1):78–88, 2007.

[Chang *et al.*, 2015] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical

Report arXiv:1512.03012 [cs.GR], Stanford University–Princeton University–Toyota Technological Institute at Chicago, 2015.

[Chen *et al.*, 2003] Ding-Yun Chen, Xiao-Pei Tian, Yu-Te Shen, and Ming Ouhyoung. On visual similarity based 3d model retrieval. *Comput. Graph. Forum*, 22(3):223–232, 2003.

[Cheng *et al.*, 2017] Zhiyong Cheng, Jialie Shen, Liqiang Nie, Tat-Seng Chua, and Mohan S. Kankanhalli. Exploring user-specific information in music retrieval. In *ACM SIGIR*, pages 655–664, 2017.

[Gao and Dai, 2014] Yue Gao and Qionghai Dai. View-based 3d object retrieval: Challenges and approaches. *IEEE MultiMedia*, 21(3):52–57, 2014.

[Gao *et al.*, 2011] Yue Gao, Qionghai Dai, Meng Wang, and Naiyao Zhang. 3d model retrieval using weighted bipartite graph matching. *Sig. Proc.: Image Comm.*, 26(1):39–47, 2011.

[Gao *et al.*, 2012] Yue Gao, Meng Wang, Dacheng Tao, Rongrong Ji, and Qionghai Dai. 3-d object retrieval and recognition with hypergraph analysis. *IEEE Trans. Image Processing*, 21(9):4290–4303, 2012.

[Guo *et al.*, 2015] Haiyun Guo, Jinqiao Wang, Min Xu, Zheng-Jun Zha, and Hanqing Lu. Learning multi-view deep features for small object retrieval in surveillance scenarios. In *ACM MM*, pages 859–862, 2015.

[He *et al.*, 2017] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *WWW*, pages 173–182, 2017.

[Hong *et al.*, 2016] Richang Hong, Zhenzhen Hu, Ruxin Wang, Meng Wang, and Dacheng Tao. Multi-view object retrieval via multi-scale topic models. *IEEE Trans. Image Processing*, 25(12):5814–5827, 2016.

[Kalogerakis *et al.*, 2016] Evangelos Kalogerakis, Melinos Averkiou, Subhransu Maji, and Siddhartha Chaudhuri. 3d shape segmentation with projective convolutional networks. *CoRR*, abs/1612.02808, 2016.

[Kanezaki, 2016] Asako Kanezaki. Rotationnet: Learning object classification using unsupervised viewpoint estimation. *CoRR*, abs/1603.06208, 2016.

[Li *et al.*, 2014] Wen Li, Lixin Duan, Dong Xu, and Ivor W. Tsang. Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(6):1134–1148, 2014.

[Li *et al.*, 2016] Yangyan Li, Sören Pirk, Hao Su, Charles Ruizhongtai Qi, and Leonidas J. Guibas. FPNN: field probing neural networks for 3d data. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, pages 307–315, 2016.

[Liu *et al.*, 2017] Anan Liu, Weizhi Nie, Yue Gao, and Yuting Su. View-based 3-d model retrieval: A benchmark. *IEEE Transactions on Cybernetics*, PP(99):1–13, 2017.

[Long *et al.*, 2013] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S. Yu. Transfer feature learning with joint distribution adaptation. In *ICCV*, pages 2200–2207, 2013.

[Maturana and Scherer, 2015] Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *IROS*, pages 922–928, 2015.

[Nie *et al.*, 2016] Liqiang Nie, Xuemeng Song, and Tat-Seng Chua. *Learning from Multiple Social Networks*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2016.

[Ohbuchi *et al.*, 2008] Ryutarou Ohbuchi, Kunio Osada, Takahiko Furuya, and Tomohisa Banno. Salient local visual features for shape-based 3d model retrieval. In *2008 International Conference on Shape Modeling and Applications*, pages 93–102, 2008.

[Pan *et al.*, 2011] Sinno Jialin Pan, Ivor W. Tsang, James T. Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Trans. Neural Networks*, 22(2):199–210, 2011.

[Phong, 1975] Bui Tuong Phong. Illumination for computer generated pictures. *Commun. ACM*, 18(6):311–317, 1975.

[Shih *et al.*, 2007] Jau-Ling Shih, Chang-Hsing Lee, and Jian Tang Wang. A new 3d model retrieval approach based on the elevation descriptor. *Pattern Recognition*, 40(1):283–295, 2007.

[Shilane *et al.*, 2004] Philip Shilane, Patrick Min, Michael M. Kazhdan, and Thomas A. Funkhouser. The princeton shape benchmark. In *SMI*, pages 167–178, 2004.

[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.

[Tang *et al.*, 2017] Sheng Tang, Yu Li, Lixi Deng, and Yongdong Zhang. Object localization based on proposal fusion. *IEEE Trans. Multimedia*, 19(9):2105–2116, 2017.

[Vranic, 2003] Dejan V. Vranic. An improvement of rotation invariant 3d-shape based on functions on concentric spheres. In *ICIP*, pages 757–760, 2003.

[Wu *et al.*, 2015] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *CVPR*, pages 1912–1920, 2015.

[Zhang *et al.*, 2015] Kun Zhang, Mingming Gong, and Bernhard Schölkopf. Multi-source domain adaptation: A causal view. In *AAAI*, pages 3150–3157, 2015.

[Zhang *et al.*, 2017] Jing Zhang, Wanqing Li, and Philip Ogunbona. Joint geometrical and statistical alignment for visual domain adaptation. In *CVPR*, pages 5150–5158, 2017.