Improved Semantic Representation Learning by Multiple Clustering for Image-Based 3D Model Retrieval

Jinghui Chu, Tianjin University, China
Xiaoqian Zhao, Tianjin University, China
Dan Song, Tianjin University, China*
Wenhui Li, Tianjin University, China
Shenyuan Zhang, People’s Daily, China
Xuanya Li, Baidu, China
An-An Liu, Tianjin University, China

ABSTRACT

Under the heavy management on the increasing 3D models, the topic of image-based 3D model retrieval which organizes unlabeled 3D models based on abundant knowledge learned from labeled 2D images has drawn attention. However, prior methods are limited in aligning semantically at corresponding categories of two domains due to the lack of label information in the 3D domain. To this end, this paper proposes an improved semantic representation learning by multiple clustering approach, which improves the reliability of pseudo labels for 3D models, so as to achieve class-level semantic alignment. Specifically, this paper first extracts features for 2D images and 3D models. Then it clusters combining the 3D features with the semantic information from multiple clustering on 3D model features to obtain more reliable target pseudo labels. Extensive experiments have shown that the proposed method has achieved the gain of 3.0%-205.0% averagely for popular retrieval metrics on the benchmark of monocular image-based 3D object retrieval (MI3DOR), and 1.3%-69.7% on another advanced benchmark, MI3DOR-2.

KEYWORDS

Domain Adaptation, Image-Based 3D Model Retrieval, Label Reliability, Multiple Clustering, Pseudo Labels, Semantic Representation Learning

INTRODUCTION

With the wide applications of 3D techniques in academic, medical and other fields (Abburu, 2019), it becomes urgent to manage the exploding unlabeled 3D data in an efficient manner. An intuitive idea is to transfer the knowledge from available annotated data into these unlabeled 3D data. In reality, existing organized 2D datasets have advantages over 3D datasets in terms of category number and sample diversity. For example, the public 2D datasets used for classification or retrieval task such as ImageNet (Deng et al., 2009) and COCO (Lin et al., 2014) have over 20,000 categories and 15,000,000 samples, which is tens of times than those of the popular 3D datasets used for classification or retrieval task such as ModelNet (Wu et al., 2015) and ShapeNet (Chang et al., 2015). And it is more
convenient to use widely available 2D images to manage 3D models. Thus, this paper leverages the model trained with labeled 2D dataset to manage these increasing unlabeled 3D data. However, this attractive task suffers from the domain shift induced by the distinctions of illumination, texture and background between 2D images and 3D models.

Domain adaptation is a mainstream direction to address the huge domain shift in cross-domain problem such as image-based 3D model retrieval in this paper. Existing domain adaptation methods can be divided into two categories (Xia & Ding, 2020), i.e., discrepancy measurement (Long et al., 2015; Long et al., 2017) and domain adversarial learning (Liu et al., 2019; Zhang et al., 2019). The former measures and reduces the statistical cross-domain distance. The latter inherits the concept of generative adversarial networks (GAN) (Goodfellow et al., 2014) to bridge the gap between two distinct domains. Specifically, a domain discriminator is added upon the feature extractor to distinguish features of source domain from those of target domain. Meanwhile, the feature extractor is trained to generate domain-invariant feature for both domains to confuse the discriminator. However, these series of domain adaptation methods are unlikely to be further improved since they only focus on domain-level alignment (global statistics alignment) but ignore the class-level alignment (semantic alignment) across domains.

To alleviated the above issue, numerous works have continuously sprung up, where one of the most effective and convincing methods is to assign relatively reliable pseudo labels for the target samples, on which the class-level alignment is promoted. For example, the authors in (Xie et al., 2018; Zhou et al., 2019a; Zhou et al., 2020) all adopted the source classifier that was well-trained on source domain to label the target domain and then performed their own designed semantic alignment. However, the source classifier is biased towards the source domain characteristics, which is not well applicable to the target domain. Thus, Zuo et al. (2021) proposed to train a target-specific classifier to assign pseudo labels for the target samples. Specifically, they utilized cross-entropy on easy target samples with the prediction from source classifier as ground-truth and a tailored GAN with two discriminators on tough target samples to train the target-specific classifier. Although such method considers target samples, the prediction accuracy of pseudo labels still relies on source classifier and domain transferability. To get rid of limitation of the source classifier and improve the reliability of pseudo labels for target samples, this paper explores the semantic similarity between target samples from different classification level by multiple clustering and then combines the similarity information with the target features to more accurately classify the target samples.

As illustrated in Figure 1, the method first initializes the source and target caches storing the features of all samples in source and target domain. Next, in each training iteration, the method performs multiple K-means (Bahmani et al., 2012) clustering with different number of cluster heads on target cache and then conducts the final clustering on the joint space uniting target cache with the multiple clustering results in order to obtain more reliable pseudo labels for target samples. Finally, the class-level centroid alignment is carefully carried out based on the computed labeled source centroids and pseudo-labeled target centroids. Moreover, the hard triplet loss (Hermans et al., 2017) is employed to make the target samples scattered among classes but compacted within classes.

This paper summarizes the main contributions as follows:

● This article proposes an improved semantic representation learning with multiple clustering approach for image-based 3D model retrieval problem, which provides more reliable pseudo-labels for 3D models domain, thereby enhancing the class-level cross-domain adaptation.

● To promote the reliability of pseudo-labels for 3D models, the paper develops multiple K-means clustering on target feature space to discover the semantic similarity between samples from different classification levels, based on which it designs centroid aligning loss and hard triplet loss to achieve alignment between the corresponding categories of two domains.

● Extensive experiments on two public cross-domain datasets, i.e., Monocular Image based 3D Object Retrieval (MI3DOR) and another advanced version MI3DOR-2, have demonstrated that the
The proposed method is superior to the state-of-the-art methods. On MI3DOR/MI3DOR-2, it improves the performance in terms of Nearest Neighbor, First Tier, Second Tier, F-Measure, Discounted Cumulative Gain and Average Normalized Modified Retrieval Rate metrics for 88.0%/40.4%, 109.8%/53.8%, 92.1%/43.9%, 120.3%/42.0%, 105.5%/51.3% and 81.9%/68.8% gains averagely over the approaches based on reducing discrepancy between domains (Gong et al., 2012; Long et al., 2017; Sun et al., 2016; Wang et al., 2018; Zhang et al., 2017), 10.0%/17.6%, 12.5%/19.0%, 10.7%/10.8%, 9.2%/20.0%, 10.1%/18.2% and 14.3%/36.3% gains averagely over the methods based on domain adversarial learning (Ganin & Lempitsky, 2015; Zhou et al., 2019a; Zhou et al., 2020) and 1.9%/4.4%, 12.5%/0.2%, 10.7%/0.4%, 9.2%/0.2%, 10.1%/0.03% and 14.3%/0.3% gains over the single clustering based method (Fu et al., 2019).

BACKGROUND

3D Model Retrieval

3D model retrieval aims to search the relevant 3D data in the forms of mesh, voxel, point cloud and multi-view from the gallery when given a query. According to the type of the query, this paper simply divides the existing 3D model retrieval methods into two categories: model-based and image-based.

The model-based 3D model retrieval methods (Qi et al., 2017a; Qi et al., 2017b; Wu et al., 2015) attempt to extract features containing the spatial and structure information for 3D models in both domains. For example, Wu et al. (2015) proposed 3D ShapeNets which learned useful representation for volumetric object by convolutional deep belief network. Qi et al. (2017a) designed a novel deep learning algorithm termed PointNet for extracting the effective feature of point cloud data. PointNet++ (Qi et al., 2017b) improved over PointNet (Qi et al., 2017a) by introducing a hierarchical neural network to generalize to more complex scenes. Due to the expensive computation in 3D convolution, Su et al. (2015) first used multi-view to represent each 3D data and then max-pooled the multi-view features which are from the 2D CNN into a compact 3D descriptor. Also, Zhou et al. (2019b) mined the potential information between views of each 3D model to enhance the discriminative ability of the ultimate representation.

Image-based 3D model retrieval methods (Liao et al., 2018; Su et al., 2020; Zhou et al., 2019a; Zhou et al., 2020) first extract features for 2D images domain and 3D models domain respectively, and then return relevant 3D models based on the distance with the 2D image query. However, the huge gap between images and models domain severely obstructs the retrieval performance. To address this problem, many works have been proposed recently. For example, Zhou et al. (2019a) proposed dual-level embedding alignment network (DLEA), which bridged the domain gap by simultaneously aligning the global statistics and the centroid of each category from both domains. Zhou et al. (2020) also presented semantic consistency guided instance feature alignment network (SC-IFA) to further improve the retrieval performance. Concretely, they designed a feature translator module to transform the features from one domain to another domain and then minimized the discrepancy between original feature space and transformed feature space. Su et al. (2020) mapped the 2D and 3D features into a Grassmann manifold space to reduce the difference between them. Considering the easy accessibility of the large scale labeled 2D images datasets, the image-based 3D model retrieval topic has gradually received increasing interests.

Unsupervised Domain Adaptation

Unsupervised domain adaptation is a crucial technique for solving cross-domain problem such as image-based 3D model retrieval. The common domain adaptation methods can be classified into two types (Xia & Ding, 2020): discrepancy measurement based and domain adversarial confusion based.

The former discrepancy measurement-based methods reduce the global statistics distribution between two domains to learn a domain-invariant feature. For instance, Tzeng et al. (2014) constrained
the Maximum Mean Discrepancy (MMD) of two distinct feature spaces. Based on MMD, Long et al. (2015) proposed multi-kernel Maximum Mean Discrepancy (MK-MMD) measurement. Zellinger et al. (2017) tried to narrow the gap of both domains from the aspect of Central Moment Discrepancy (CMD). Shen et al. (2018) computed and minimized the Wasserstein distance between source and target domains. Moreover, a series of new discrepancy measurement criteria are emerged in (Long et al., 2017; Pan & Yang, 2020; Si et al., 2021; Yan et al., 2017). The latter domain adversarial confusion-based methods additionally add a domain discriminator to distinguish the source samples from target samples while the feature extractor is fooled to generate a domain-invariant feature. This idea was first carried out in (Ganin & Lempitsky, 2015). In spite of the remarkable improvement by the above domain adaptation approaches, the mismatch of feature and its category is not taken into account due to the lack of target domain labels. To this end, Xie et al. (2018) aligned the labeled source centroid and corresponding pseudo-labeled target centroid to address semantic mismatch, where the pseudo-labels of target domain came from the source classifier. Instead of directly trusting the pseudo labels, Zuo et al. (2021) designed a GAN with two target classifiers to learn discriminative features for tough samples by minmax game while leveraging supervised information of easy samples to facilitate tough ones. However, there still existed false pseudo labels harming the classification or recognition for target domain. Thus, reliable pseudo labels for target samples are critical to domain adaptation.

Unsupervised Clustering

The objective of unsupervised clustering is to divide the data into several distinct classes of similar characteristics. The most classical clustering methods are K-Means (Bahmani et al., 2012), DBSCAN (Ester et al., 1996) and HAC (Lukasová, 1979), and many advanced algorithms (Fu et al., 2019; Hämäläinen et al., 2021; Lin et al., 2019) have developed upon them. For example, an improving scalable K-Means++ with new initialization method was proposed in (Hämäläinen et al., 2021). Fu et al. (2019) applied DBSCAN algorithm on the whole/upper/lower body of each image to produce three pseudo-label sets for every image, and then trained the model with these pseudo-labeled data using triplet loss. Lin et al. (2019) designed a bottom-up clustering (BUC) approach, which began with regarding each sample as one group, and then iteratively merged clusters according to the minimum distance criterion and trained the network using the new clustering label results until the model reached stability. Inspired by BUC, Zhao et al. (2020) adopted inter-cluster and intra-cluster distance as new enhanced merging strategy. Furthermore, to make the clustering results more convincing, Gupta et al. (2020) extracted highly confident data samples from the clustering pseudo labeled dataset to perform semi-supervised training. Park et al. (2021) introduced co-training strategy and label-smoothing to refine noisy pseudo labels from the initial clustering methods. Chen et al. (2020) proposed to apply multiple clustering on the whole/upper/lower parts of each image to promote the performance of (Fu et al., 2019). Motivated by this, this paper considers employing multiple clustering to improve the reliability of pseudo labels on target domain and then performs semantic alignment by reducing the distance between samples within the same class but across different domains to tackle imaged-based 3D model retrieval problem.

PROPOSED METHOD

Overview

For image-based 3D model retrieval problem, the researchers have access to $n_s$ labeled 2D images $\{ (x^i_S, y^i_S) \}_{i=1}^{n_s}$ from source domain $D_S$ and $n_t$ unlabeled 3D models $\{ x^i_T \}_{i=1}^{n_t}$ from target domain $D_T$, where $x^i_S \in X_S$, $x^i_T \in X_T$ and $y^i_S \in Y_S$. $Y_T'$ refers to the pseudo label set of target domain.

Figure 1 illustrates the overall pipeline of the proposed method. It is composed of three modules: feature extraction, cross-domain adaptation and retrieval. In the feature extraction, the source and
target features are extracted. In the cross-domain adaptation, the source classification information, domain-level global alignment with adversarial learning and class-level centroid alignment of which the pseudo labels of target domain are from multiple clustering procedure are jointly used to learn discriminative features for target samples. In retrieval, the adapted features are used for retrieval task. Specifically, first of all, the method pre-trains the entire model with the source supervised information and domain adversarial learning between source and target domains. Then the labeled 2D images unlabeled 3D models are input into the pretrained model for feature extraction. Based on the extracted 2D/3D features, the authors utilize the source classification information, domain-level global alignment and class-level centroid alignment to achieve cross-domain adaptation. In domain-level global alignment, a domain discriminator is used to assist the feature extractor to align the global statistics between two domains. In class-level centroid alignment which is an innovative work in this paper, the authors strive to align the centroids computed from same category but different domains. Specifically, the authors firstly get a pseudo label matrix by multiple K-means clustering on 3D features space, where the number of cluster heads in each time is different. Secondly, they combine the target feature cache with the pseudo label matrix horizontally to perform the final K-means clustering on it for the purpose of obtaining more reliable pseudo labels, so as to successfully accomplish class-level centroid alignment. When the training process of this model reaches an optimal state, the model can be adopted to extract features for query image and 3D gallery and then compute a ranking list for gallery based on distance metric. With the ranking list, 6 popular criteria, namely NN, FT, ST, F, DCG and ANMRR are used to assess the retrieval performance.

Model Pre-training

In order to ensure the reliability of clustering results at the early stage of training, which is beneficial to the class-level centroid alignment, this paper firstly pre-trains the model on the specified dataset, i.e., MI3DOR or MI3DOR-2. Specifically, it jointly pre-trains the feature extractor $F$, domain discriminator $D$ and source classifier $G$ by minimizing the source classification error and adversarially narrowing discrepancy across two domains on MI3DOR or MI3DOR-2 dataset. Meanwhile, to avoid the time consumption of pre-training from scratch, we download the existing parameters that are from the ResNet50 architecture well-trained on ImageNet (Deng et al., 2009) dataset to initialize the feature extractor $F$ before the pre-training process. The domain discriminator $D$ and source classifier $G$ are pre-trained from scratch. The objective for the pre-training of the feature extractor $F$, domain discriminator $D$ and source classifier $G$ is defined as:
\[ L = L_{\text{CE}}(X_S, Y_S) + L_{\text{ADV}}(X_S, X_T) \]  

(1)

\[ L_{\text{CE}}(X_S, Y_S) = E_{(x, y) \sim D_S} J(y, G(F(x))) \]  

(2)

\[ L_{\text{ADV}}(X_S, X_T) = -E_{x \sim D_S} \left[ \log \left( D(F(x)) \right) \right] - E_{x \sim D_T} \left[ \log \left( 1 - D(F(x)) \right) \right] \]  

(3)

Where \( J(\cdot) \) represents the cross-entropy loss function and \( L_{\text{ADV}} \) is the domain adversarial loss used for narrowing the difference across domains. For domain adversarial learning, the discriminator \( D \) learns to distinguish whether the feature comes from source or target domain while the feature extractor \( F \) aims to learn domain-invariant feature to confuse the discriminator \( D \). When this two-player game reaches an equilibrium, the gap between two domains is reduced to the satisfactory case.

**Feature Extraction**

As shown in Figure 1, inspired by multi-view convolutional neural networks (MVCNN) (Su et al., 2015), the proposed method first represents each 3D model with \( N \) view images rendered from virtual cameras uniformly distributed around the model. Then, it employs the feature extractor \( F \) to extract the image/view features for source/target domain. For multi-view features of each 3D model, it fuses them into a compact 3D descriptor via max-pooling operation. Formally, \( f_{F,X}^S = (\cdot) \) represents the visual feature of 2D image \( x_S^i \). \( f_{F,X}^T = (\cdot) \) represents the visual feature of 3D model \( x_T^i \).

**Cross-domain Adaptation**

This module takes the source/target features generated from feature extractor \( F \) as input and performs domain adaptation with source classification information, domain-level global alignment and class-level centroid alignment. Minimizing source classification error and achieving domain-level global alignment by Equation (1) to (3) were detailed in Section Model Pre-training. Thus, this paper carefully elaborates the class-level centroid alignment below.

**Class-level Centroid Alignment Based on Multiple Clustering**

Only applying the source classification information and domain-level global alignment can not ensure source and target features belonging to the same class mapped nearby. Motivated by this, this paper designs class-level centroid alignment to reduce the distance of centroids of same class across domains. However, the label information is required in computing the center of each class in both domains. To obtain a pseudo labeled target domain, previous methods either resorted to source classifier or trained a target-specific classifier to assign pseudo labels for target domain. But the source classifier is biased towards source domain characteristics and the prediction accuracy of pseudo labels from target-specific classifier still relies on source classifier. Thus, to get rid of limitation of the source classifier and improve the reliability of pseudo labels for target samples, this paper explores the semantic similarity between 3D models from different classification level by multiple clustering and then combines the similarity information with 3D model features to assign reliable pseudo labels for 3D models.
As shown in Figure 1 and the Algorithm 1 in Table 1, the authors first initialize a source cache of size \( n_s \times d \) and a target cache of size \( n_t \times d \) to store feature vectors of all samples in the source and target domains, respectively. Here, \( d \) is the dimension of source/target feature vector. Then during each iteration, they update the corresponding rows in the source/target cache with source/target features generated from feature extractor \( F \) in the current batch. Based on the updated target cache, they perform K-means clustering for \((C - 1)\) times with \( K \) ranging from 2 to \( C \) to get corresponding \((C - 1)\) clustering results, where \( C \) is the number of common categories in both domains. Note that the number of cluster heads in each time, i.e., the setting of value \( K \) in K-means function is different and varies from 2 to \( C \), which helps to discover the semantic similarity between samples from different classification levels. Therefore, the authors can obtain the pseudo-label matrix \( L_{pseudo} \), which is defined as:

\[
L_{pseudo} = \left[ L^1_T, L^2_T, \ldots, L^n_T \right]
\]

(4)

\[
L^i_T = \left[ L^i_T(2), L^i_T(3), \ldots, L^i_T(C) \right], \quad \forall i \in \{1, 2, \ldots, n_t\}
\]

(5)

where \( C \) is the number of common categories in both domains, \( L^i_T \) is the pseudo label vector of model \( x^i_T \), \( L^i_T(j) \) is a pseudo label of model \( x^i_T \) produced by the K-means clustering with \( j \) cluster heads.

For making effective use of semantic similarity information between samples from different classification level in the pseudo-label matrix \( L_{pseudo} \), the method concatenates the target cache matrix and pseudo-label matrix horizontally to get the final matrix \( M \). Then it conducts the final K-means clustering with \( C \) cluster heads on the matrix \( M \) to obtain the final pseudo label set of the target domain \( Y^T_T \), which are more reliable for subsequent cross-domain centroid alignment work. However, \( Y^T_T \) is a clustering label set whose semantic category of each class is not consistent with source ground-truth set \( Y^S_S \). Thus, the method needs to translate \( Y^T_T \) to \( Y^T_T' \) that strictly corresponds with the source label set \( Y^S_S \). Specifically, the method collects the source sample label closest to each target sample in each cluster and then regards the source label that appears most in current cluster as the objective pseudo label.

With the ground-truth set \( Y^S_S \) and the objective pseudo label set \( Y^T_T' \), the method computes the centroid of each class for source and target domains respectively and then restricts the distance between pair of centroids of the same class in these two domains. Formally,

\[
L_{CA} \left( X^S_S, Y^S_S, X^T_T, Y^T_T' \right) = \sum_{c=1}^{C} \phi \left( O^c_S, O^c_T \right)
\]

(6)

Here, \( C \) is the number of common categories in both domains, \( O^c_S \) and \( O^c_T \) are centroids of the \( c^{th} \) class in source and target domains, and \( \phi \left( \cdot, \cdot \right) \) is Euclidean distance function. Consequently, the class-level centroid alignment is perfectly implemented by minimizing the loss \( L_{CA} \). Moreover, this paper applies hard triplet loss on target feature space to make the target samples scattered among classes but compact within classes.
$$L_{\text{triplet}}(X_T, Y_T) = \sum_{i=1}^{n_i} \left[ m + \max_{j=1...n_i} f_a^i - f_p^j - \min_{j=1...n_i} f_a^i - f_n^j \right]$$ (7)

where $m$ is the margin hyperparameter between positive and negative samples, and $f_a^i$, $f_p^i$ and $f_n^j$ are features extracted from the anchor, positive and negative samples of target domain, respectively.

### Overall Loss Function

The paper uses a combination of source classification loss $L_{\text{CE}}$, domain-level adversarial loss $L_{\text{ADV}}$, class-level centroid loss $L_{\text{CA}}$ and hard triplet loss $L_{\text{triplet}}$ to jointly re-train the entire network:

$$L_{\text{total}} = L_{\text{CE}}(X_S, Y_S) + \lambda_1 \cdot L_{\text{ADV}}(X_S, X_T) + \lambda_2 \cdot L_{\text{CA}}(X_S, Y_S, X_T, Y_T) + \lambda_3 \cdot L_{\text{triplet}}(X_T, Y_T)$$ (8)

where $\lambda_1$, $\lambda_2$ and $\lambda_3$ are hyperparameters to balance these four loss values to reach the best optimization situation.

### EXPERIMENTS

#### Setup

**Dataset**

This paper chooses two public datasets, MI3DOR (Li et al., 2019) and MI3DOR-2 (Zhou et al., 2019a) for evaluation of the proposed method.

MI3DOR includes 21,000 2D images as source domain and 7,690 3D models as target domain, among which 10,500 images and 3,842 models are used for training while the rest are used for testing. There are 21 common categories existing in both domains.

MI3DOR-2 contains 19,694 2D images as source domain and 3,982 3D models as target domain that both domains share the common 40 categories. The training set consists of 19,294 images and 3,182 models and the testing set is composed of the remaining 400 images and 800 models. Compared with MI3DOR, images in MI3DOR-2 have clearer background.

**Evaluation Protocol**

In experiments, the researchers use the popular criteria as (Liu et al., 2017), including Nearest Neighbor (NN), First Tier (FT), Second Tier (ST), F Measure (F), Discounted Cumulative Gain (DCG) and Average Normalized Modified Retrieval Rank (ANMRR) to evaluate the performance on image-based 3D model retrieval task. Suppose that $N_q$ is the number of images in query set, $N_g$ is the number of models in gallery set, $G_q$ of size $N_q \times 1$ is the label set of the query, $G_g$ of size $N_g \times 1$ is the label set of the gallery, $NC_q$ of size $C \times 1$ represents the sample number contained within each category in the query, $NC_g$ of size $C \times 1$ represents the sample number contained within each category in the gallery. $T(i) = NC_g(G_q(i))$ represents the number of matched samples in gallery for the $i^{th}$ query. With extracted features for query samples and gallery samples, the returned results...
are ranked by similarity computation. Then let $R$ of size $N_q \times N_g$ represents the retrieval indicator, where 1/0 denotes that the label of returned result matches/mismatches the query. Then the six metrics are defined as follows:

1) NN (Nearest Neighbor) represents the average accuracy of the first returned retrieval result of all query images: 

\[ \text{NN} = \frac{\sum_{i} R(i,1)}{N_g}; \]

2) FT (First Tier) represents the average accuracy of top $T$ retrieval results of all query images:

\[ \text{FT} = \frac{\sum_{i} \left( \frac{R(i,1; T(i))}{T(i)} \right)}{N_g}; \]

3) ST (Second Tier) represents the average accuracy of top $2T$ retrieval results of all query images:

\[ \text{ST} = \frac{\sum_{i} \left( \frac{R(i,1; T(i) \times 2)}{T(i)} \right)}{N_g}; \]

4) F-Measure is used to jointly evaluate the average recall (AR) and average precision (AP) of the retrieval results:

\[ \text{AR} = \frac{\sum_{i} R(i,1; :20)}{N_q \times N_g}, \quad \text{AP} = \frac{\sum_{i} R(i,1; :20)}{N_q \times 20}, \quad \text{and} \]

\[ F = \frac{2 \times \text{AR} \times \text{AP}}{\text{AR} \times \text{AP} + \text{AR} + \text{AP}}; \]

Table 1. The training procedure of the proposed method

<table>
<thead>
<tr>
<th>Algorithm 1: The training procedure of the approach of improved semantic representation learning with multiple clustering.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input:</strong> labeled 2D images set $\left( X_S, Y_S \right)$, unlabeled 3D models set $\left( X_T \right)$, initial network parameter $\theta$ from the well pre-trained feature extractor $F$, domain discriminator $D$ and classifier $G$ on MI3DOR/MI3DOR-2, maximum iteration round $R$, the number of common categories in both domains $C$.</td>
</tr>
<tr>
<td><strong>Output:</strong> Trained network parameter $\theta$.</td>
</tr>
<tr>
<td>1: Initialize: source cache, target cache.</td>
</tr>
<tr>
<td>2: for $r = 1 \text{ to } R$ do</td>
</tr>
<tr>
<td>3: Update source/target cache with the most recent extracted features.</td>
</tr>
<tr>
<td>4: Cluster ($C - 1$) times with varying cluster heads from 2 to $C$ on target cache to obtain pseudo-label matrix $L_{pseudo}$ defined in Eq. (4) and Eq. (5).</td>
</tr>
<tr>
<td>5: Combine target cache with $L_{pseudo}$ and cluster one time with cluster head $C$ to obtain pseudo label set $Y_T$.</td>
</tr>
<tr>
<td>6: Translate $Y_T$ to $Y_T'$ semantically consistent to the source ground truth set $Y_S$.</td>
</tr>
<tr>
<td>7: Compute the distance between the same class across domains by Eq. (6) and the hard triplet loss by Eq. (7).</td>
</tr>
<tr>
<td>8: Update network parameter $\theta$ by optimizing the overall loss Eq. (8).</td>
</tr>
<tr>
<td>9: end for</td>
</tr>
</tbody>
</table>
5) DCG (Discounted Cumulative Gain) represents a practical retrieval evaluation which gives the returned results falling behind a discount: for the \( i^{th} \) query, its discounted cumulative gain is

\[
DCG(i) = R(i,1) + \sum_{k=2}^{\infty} \frac{r(i,k) R(i,k)}{\log_2 k}.
\]

Then normalize \( DCG(i) \) with the ideal maximum value:

\[
DCG_{\text{norm}}(i) = \frac{DCG(i)}{1 + \sum_{k=2}^{\infty} \frac{1}{\log_2 k}},
\]

and finally compute the average DCG of all query samples:

\[
DCG = \frac{\sum_i DCG_{\text{norm}}(i)}{N_q}.
\]

6) ANMRR (Average Normalized Modified Retrieval Rate) also takes the ranking of retrieval results into consideration to evaluate the performance. Suppose \( r(k) \) is the ranking of the returned results for the \( i^{th} \) query:

\[
r(k) = \begin{cases} 
  k, & \text{if } R(i,k) = 1 \\
  K(i), & \text{if } R(i,k) = 0
\end{cases},
\]

where \( K(i) = \min(4 \times T(i), 2 \times \max(NCG)) \). Then

\[
ARR(i) = \frac{\sum_{k=1}^{T(i)} r(k)}{T(i)}, \quad \text{NMRR}(i) = \frac{ARR(i) - \frac{T(i)}{2} - 0.5}{\frac{T(i)}{2} + 0.5},
\]

and the average NMRR is computed as: \( \text{ANMRR} = \frac{\sum_i \text{NMRR}(i)}{N_q} \).

For the first five criteria, the higher value means the better performance. On the contrary, for the last criterion, the lower value means the better performance. Generally, these criteria have consistent performance, where for better performance the values of NN, FT, ST, F and DCG tend to be higher and the value of ANMRR tend to be lower.

Comparisons

The researchers compare their method with the following representative cross-domain adaptation algorithms: 1. Discrepancy measurement based: a. CORrelation Alignment (CORAL) (Sun et al., 2016) aligns the second-order statistics information between source and target domains to bridge the domain gap. b. Geodesic Flow Kernel (GFK) (Gong et al., 2012) first maps the source and target samples into Grassmann manifolds, then computes and minimizes the discrepancy from geometric and statistical properties in multi-subspaces. c. Manifold Embedded Distribution Alignment (MEDA) (Wang et al., 2018) first maps the feature into Grassmann manifold space and then conducts distribution alignment by dynamically assigning weights to marginal and conditional distribution. d. Joint Geometrical and Statistical Alignment (JGSA) (Zhang et al., 2017) reduces the domain shift by projecting the source and target data into a low-dimensional subspace and then minimizing the geometrical and distribution discrepancy based on this subspace. e. Joint Adaptation Network (JAN) (Long et al., 2017) aligns the distributions of outputs from several task-specific layers between source and target domain via joint maximum mean discrepancy. 2. Domain adversarial confusion based: a. Reversal Gradient (RevGard) (Ganin & Lempitsky, 2015) adds a domain classifier to learn a domain-invariant feature space. b. Dual-level Embedding Alignment (DLEA) (Zhou et al., 2019a) not only learns domain-invariant feature by domain-level embedding alignment but also semantically aligns the centroids of same
class in both domains. c. Semantic Consistency guided Instance Feature Alignment (SC-IFA) (Zhou et al., 2020) first utilizes the adversarial training to achieve domain alignment and then transforms the feature from one domain to another domain based on which the class-level semantic alignment is implemented by minimizing the translation loss and correlation loss. 3. Clustering based: a. Self-similarity Grouping (SSG) (Fu et al., 2019) starts with pre-training the model on source dataset and then applies DBSACN clustering on whole/upper/lower part feature of each image to obtain clustering pseudo labels set as supervised information to re-train the source model.

Implementation Details

In the experiments, the researchers employ the ResNet50 as the backbone of the feature extractor $F$, which consists of five convolutional layers, an average pool layer and a 256-d fully connected layer. The architecture in the domain discriminator $D$ includes three-layer fully connected layers with activation function as $ReLU$. The source classifier $G$ is composed by two-layer fully connected layers and a $Softmax$ function. The weights and biases in all layers of the feature extractor $F$, domain discriminator $D$ and source classifier $G$ are defined as the network parameters $\theta$ (in Algorithm 1). In the pre-training phase, this paper jointly trains the feature extractor $F$, domain discriminator $D$ and source classifier $G$ for 10,000 steps. Stochastic gradient descent (SGD) is exploited with momentum 0.9 and weight decay 0.0005 during back-propagation. The initial learning rate is set to 0.01 and annealed by $\mu_p = \frac{H_0}{(1 + \alpha \cdot p)^\beta}$, where $\alpha = 10$, $\beta = 0.75$ and $p$ is the progress of training steps linearly varying from 0 to 1 by following (Zhou et al., 2019a). In re-training phase, they reduce the initial learning rate to 3e-4. The total steps are still set to 10,000. Multi-clustering procedure on target cache is performed every 1,200 steps. For the weight balance parameters, they empirically set $\lambda_1 = 1$. $\lambda_2$ and $\lambda_3$ are selected from $\{10^{-3}, 5 \times 10^{-3}, 10^{-2}, 5 \times 10^{-2}, 10^{-1}, 5 \times 10^{-1}, 1\}$. The best setting of $\lambda_2$ and $\lambda_3$ is obtained in the Section Parameter Analysis.

Comparison with State-of-the-Arts

The comparisons with the state-of-the-art algorithms on MI3DOR and MI3DOR-2 datasets are shown in Table 2 and Table 3, respectively. Note that the number of all comparison methods are directly quoted from (Zhou et al., 2020), except for the SSG which is reproduced from (Fu et al., 2019) by the authors. They adopt the same single split setting as other comparison methods to evaluate the retrieval performance. Obviously, it can be seen that the proposed method in this paper outperforms all comparison counterparts in almost all criteria by a large margin. On MI3DOR, the proposed method achieves 1.9%-142.6%, 4.5%-253.6%, 3.8%-198.9%, 3.8%-222.8% (except SC-IFA), 1.9%-226.0% and 7.2%-122.3% (except SC-IFA) improvements in terms of NN, FT, ST, F, DCG and ANMRR metrics over the other comparisons. On MI3DOR-2, the proposed method exceeds others by a margin of 6.4%-55.0%, 0.2%-75.9%, 0.4%-55.2%, 0.2%-75.9%, 0.03%-70.3%, 0.3%-85.8% in terms of NN, FT, ST, F, DCG and ANMRR metrics. Moreover, for fair justification on the effectiveness of the proposed method, the authors additionally employ 10-fold cross-validation setting to assess the performance. The results reported under 10-fold cross-validation setting are comparable to those under single split setting, which further demonstrates the effectiveness of the proposed method. This significant improvement is mainly due to the procedure of combining target features with multi-level semantic information which promotes the clustering accuracy and facilitates the subsequent semantic alignment work. Furthermore, the authors also conclude as followings:

● In general, the proposed method is superior to all other comparison counterparts, which verifies the reliability of target pseudo label set generated from the process of clustering ($C - 1$) times on the whole target features space and then clustering the last one time on the joint space uniting
target features with the \((C-1)\) clustering results. The resulting target pseudo labels facilitate the subsequent cross-domain centroid alignment work.

- The domain adversarial confusion-based methods (RevGard, DLEA, SC-IFA) outperform the discrepancy measurement-based methods (CORAL, GFK, MEDA, JGSA, JAN), which demonstrates that domain adversarial training is more beneficial to cross-domain adaptation than the way of reducing statistical discrepancy between two domains.
- SSG trains the network with simple triplet loss based on single DBSCAN clustering pseudo labels. Comparing with discrepancy measurement-based methods, SSG has a significant improvement over them. This suggests unsupervised clustering is useful for cross-domain retrieval task. Thus, this paper considers improving the existing single clustering to multiple clustering for the objective of discovering the semantic similarity between samples from different classification level. The results of SSG and the proposed method in this paper also confirm the superiority of multiple clustering over single clustering.

Table 2. Retrieval performance on MI3DOR

<table>
<thead>
<tr>
<th>Method</th>
<th>NN</th>
<th>FT</th>
<th>ST</th>
<th>F</th>
<th>DCG</th>
<th>ANMRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>CORAL (Sun et al., 2016)</td>
<td>0.362</td>
<td>0.174</td>
<td>0.256</td>
<td>0.060</td>
<td>0.199</td>
<td>0.816</td>
</tr>
<tr>
<td>GFK (Gong et al., 2012)</td>
<td>0.323</td>
<td>0.309</td>
<td>0.338</td>
<td>0.065</td>
<td>0.314</td>
<td>0.688</td>
</tr>
<tr>
<td>MEDA (Wang et al., 2018)</td>
<td>0.430</td>
<td>0.344</td>
<td>0.501</td>
<td>0.046</td>
<td>0.361</td>
<td>0.646</td>
</tr>
<tr>
<td>JGSA (Zhang et al., 2017)</td>
<td>0.612</td>
<td>0.443</td>
<td>0.599</td>
<td>0.116</td>
<td>0.473</td>
<td>0.541</td>
</tr>
<tr>
<td>JAN (Long et al., 2017)</td>
<td>0.446</td>
<td>0.344</td>
<td>0.495</td>
<td>0.085</td>
<td>0.364</td>
<td>0.647</td>
</tr>
<tr>
<td>RevGard (Ganin &amp; Lempitsky, 2015)</td>
<td>0.650</td>
<td>0.505</td>
<td>0.643</td>
<td>0.112</td>
<td>0.542</td>
<td>0.474</td>
</tr>
<tr>
<td>DLEA (Zhou et al., 2019a)</td>
<td>0.764</td>
<td>0.558</td>
<td>0.716</td>
<td>0.143</td>
<td>0.597</td>
<td>0.421</td>
</tr>
<tr>
<td>SC-IFA (Zhou et al., 2020)</td>
<td>0.721</td>
<td>0.584</td>
<td>0.721</td>
<td>0.163</td>
<td>0.637</td>
<td>0.363</td>
</tr>
<tr>
<td>SSG (Fu et al., 2019)</td>
<td>0.7650</td>
<td>0.5887</td>
<td>0.7374</td>
<td>0.1425</td>
<td>0.6216</td>
<td>0.3935</td>
</tr>
<tr>
<td>ours (single-split)</td>
<td>0.7794</td>
<td>0.6152</td>
<td>0.7653</td>
<td>0.1485</td>
<td>0.6488</td>
<td>0.3670</td>
</tr>
<tr>
<td>ours (10-fold)</td>
<td>0.7776</td>
<td>0.6060</td>
<td>0.7536</td>
<td>0.1462</td>
<td>0.6397</td>
<td>0.3762</td>
</tr>
</tbody>
</table>

Table 3. Retrieval performance on MI3DOR-2

<table>
<thead>
<tr>
<th>Method</th>
<th>NN</th>
<th>FT</th>
<th>ST</th>
<th>F</th>
<th>DCG</th>
<th>ANMRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>CORAL (Sun et al., 2016)</td>
<td>0.538</td>
<td>0.369</td>
<td>0.497</td>
<td>0.369</td>
<td>0.399</td>
<td>0.614</td>
</tr>
<tr>
<td>GFK (Gong et al., 2012)</td>
<td>0.513</td>
<td>0.471</td>
<td>0.495</td>
<td>0.471</td>
<td>0.484</td>
<td>0.527</td>
</tr>
<tr>
<td>MEDA (Wang et al., 2018)</td>
<td>0.570</td>
<td>0.392</td>
<td>0.523</td>
<td>0.392</td>
<td>0.425</td>
<td>0.590</td>
</tr>
<tr>
<td>JGSA (Zhang et al., 2017)</td>
<td>0.585</td>
<td>0.405</td>
<td>0.533</td>
<td>0.405</td>
<td>0.433</td>
<td>0.577</td>
</tr>
<tr>
<td>JAN (Long et al., 2017)</td>
<td>0.608</td>
<td>0.501</td>
<td>0.646</td>
<td>0.501</td>
<td>0.527</td>
<td>0.484</td>
</tr>
<tr>
<td>RevGard (Ganin &amp; Lempitsky, 2015)</td>
<td>0.623</td>
<td>0.467</td>
<td>0.614</td>
<td>0.467</td>
<td>0.503</td>
<td>0.514</td>
</tr>
<tr>
<td>DLEA (Zhou et al., 2019a)</td>
<td>0.700</td>
<td>0.555</td>
<td>0.681</td>
<td>0.555</td>
<td>0.593</td>
<td>0.424</td>
</tr>
<tr>
<td>SC-IFA (Zhou et al., 2020)</td>
<td>0.713</td>
<td>0.641</td>
<td>0.738</td>
<td>0.623</td>
<td>0.648</td>
<td>0.415</td>
</tr>
<tr>
<td>SSG (Fu et al., 2019)</td>
<td>0.7475</td>
<td>0.6480</td>
<td>0.7650</td>
<td>0.6490</td>
<td>0.6793</td>
<td>0.3320</td>
</tr>
<tr>
<td>ours (single-split)</td>
<td>0.7950</td>
<td>0.6490</td>
<td>0.7683</td>
<td>0.6490</td>
<td>0.6795</td>
<td>0.3309</td>
</tr>
<tr>
<td>ours (10-fold)</td>
<td>0.7825</td>
<td>0.6481</td>
<td>0.7634</td>
<td>0.6481</td>
<td>0.6799</td>
<td>0.3328</td>
</tr>
</tbody>
</table>
Ablation Study

In this section, the researchers conduct an ablation study to evaluate the effectiveness of each component in the proposed method on MI3DOR and MI3DOR-2 datasets. As shown in Table 4 and Table 5, they first compare 4 variants: model trained with classification loss only on source data (denoted as $L_{CE}$); model trained with source classification loss and domain adversarial loss for achieving fundamental domain-level alignment (denoted as $L_{CE} + L_{ADV}$); the proposed method in this paper without triplet loss (denoted as $L_{CE} + L_{ADV} + L_{CA}$); the proposed method with triplet loss (denoted as $L_{CE} + L_{ADV} + L_{CA} + L_{triplet}$). It can be clearly found that $L_{CE} + L_{ADV}$ surpasses $L_{CE}$ by a large margin especially in FT, ST, DCG, ANMRR metrics on MI3DOR and FT, ST, F, DCG, ANMRR metrics on MI3DOR-2, which proves that domain alignment is greatly essential for cross-domain retrieval. Then, compared to $L_{CE} + L_{ADV}$, the proposed method with only class-level centroid alignment $L_{CE} + L_{ADV} + L_{CA}$ improves the performance in NN, FT, ST, F, DCG and ANMRR metrics by 4.7%, 6.4%, 6.3%, 2.3%, 5.2% and 9.2% on MI3DOR, and 3.0%, 1.1%, 0.3%, 1.1%, 1.3% and 2.3% on MI3DOR-2. This is because class-level centroid alignment enforces the features in the same class but different domains mapped nearby based on the target pseudo label set produced by the carefully designed clustering process in this paper and the given source labels set. Moreover, adding hard triplet loss $L_{triplet}$ to $L_{CE} + L_{ADV} + L_{CA}$ can further improve the retrieval performance, especially in NN value on both datasets, since $L_{triplet}$ guides the target samples scattered among classes but compacted within classes, making the features more discriminative.

<table>
<thead>
<tr>
<th>Table 4. Ablation results on MI3DOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>nn</td>
</tr>
<tr>
<td>----</td>
</tr>
<tr>
<td>$L_{CE}$</td>
</tr>
<tr>
<td>$L_{CE} + L_{ADV}$</td>
</tr>
<tr>
<td>$L_{CE} + L_{ADV} + L_{CA}$</td>
</tr>
<tr>
<td>$L_{CE} + L_{ADV} + L_{CA} + L_{triplet}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 5. Ablation results on MI3DOR-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>nn</td>
</tr>
<tr>
<td>----</td>
</tr>
<tr>
<td>$L_{CE}$</td>
</tr>
<tr>
<td>$L_{CE} + L_{ADV}$</td>
</tr>
<tr>
<td>$L_{CE} + L_{ADV} + L_{CA}$</td>
</tr>
<tr>
<td>$L_{CE} + L_{ADV} + L_{CA} + L_{triplet}$</td>
</tr>
</tbody>
</table>
Visualization and Parameter Analysis

Figure 2 and Figure 3 show several retrieval examples from MI3DOR and MI3DOR-2, where each row contains an image query and its returned top-5 3D models. From the top-3 rows in both figures, it can be seen that all the returned results exactly match with the query image, verifying the effectiveness of the proposed method in this paper in solving cross-domain retrieval task. For the row 4, row 5, row 6 in both figures, there happens mismatched 3D models marked by red forks. The authors analyze and summarize the following three main reasons for these failed cases: 1. The incomplete object in query, such as the stairs query in row 4 in Figure 2. 2. The similarity existing in the local or whole parts between different classes. For example, the cup and flower pot have the same columnar shape; the mantel is like a part structure out of the table; the bench and bookshelf are all constituted by many wooden battens. 3. There are two or more objects of different classes appearing in an image/model. For instance, the plant always occurs together with the flower pot or vase. It can be inferred from the above failed cases that the proposed method has limitation in distinguishing two objects of very similar appearance but from different categories. Thus, the authors will consider incorporating the effective label refinement strategies into this work for learning more discriminative features in the future.

The hyperparameters $\lambda_2$ and $\lambda_3$ are used to trade off the effect of class-level centroid loss and hard triplet loss which are of great significant for successfully achieving cross-domain adaptation. The authors select value for $\lambda_2$ and $\lambda_3$ from $\{10^{-3}, 5 \times 10^{-3}, 10^{-2}, 5 \times 10^{-2}, 10^{-1}, 5 \times 10^{-1}, 1\}$ and evaluate the impact of $\lambda_2$ and $\lambda_3$ on the testing set of MI3DOR. The results are reported in Figure 4 and Figure 5. In case of $\lambda_2$, the accuracy in NN, FT, ST, F, DCG and ANMRR metrics reach the highest at $\lambda_2 = 0.01$ and start to slightly decrease after it. Meanwhile, the accuracy in these six metrics keep stable when changing the value of $\lambda_3$ from which $\lambda_3 = 0.1$ is better. Thus, the setting of $\lambda_2 =0.01$ and $\lambda_3 =0.1$ is optimal in this paper.

The margin between positive and negative samples $m$ in hard triplet loss is crucial for making a scattered inter-class and compact intra-class relationship. The researchers evaluate the influence of $m$ uniformly sampled from interval (0,1]. As shown in Figure 6 and Figure 7, it can be observed that the best performance on MI3DOR and MI3DOR-2 is obtained by setting $m=0.3$ and $m=0.6$, respectively. Setting $m$ less or greater than 0.3/0.6 on MI3DOR/MI3DOR-2 all brings negative effects to some extent, due to too large or small $m$ all unable to accurately distinguish any two samples. Therefore, $m=0.3/m=0.6$ is the optimal selection on MI3DOR/MI3DOR-2 dataset for cross-domain retrieval task.

CONCLUSION

This paper proposes improved semantic representation learning with multiple clustering approach, which can produce more reliable pseudo labels for target 3D models to tackle the scarcity of 3D model label information existing in the image-based 3D model retrieval task. It jointly trains the network with source classification supervision, domain adversarial global alignment and class-level centroid alignment based on the given labeled 2D images domain and pseudo-labeled 3D models domain. Specifically, the pseudo label set of target domain is produced by first clustering multiple times on the whole target features and then clustering the last one time on the joint space uniting the target features with the multiple clustering results before. The multiple clustering is capable of capturing semantic similarity between 3D samples from different classification levels beneficial for recognizing these 3D models in a higher accuracy. Experimental results demonstrate the superiority of the proposed method over the state-of-the-art cross-domain adaptation methods on image-based 3D model retrieval datasets MI3DOR and MI3DOR-2. Besides, we sincerely thank to the Baidu Program for the Paddlepaddle platform.
Figure 2. Visual retrieval results on MI3DOR

<table>
<thead>
<tr>
<th>Query</th>
<th>Retrieval results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motorcycle</td>
<td><img src="image" alt="Motorcycle Retrieval Results" /></td>
</tr>
<tr>
<td>Rifle</td>
<td><img src="image" alt="Rifle Retrieval Results" /></td>
</tr>
<tr>
<td>Tent</td>
<td><img src="image" alt="Tent Retrieval Results" /></td>
</tr>
<tr>
<td>Stairs</td>
<td><img src="image" alt="Stairs Retrieval Results" /></td>
</tr>
<tr>
<td>Vase</td>
<td><img src="image" alt="Vase Retrieval Results" /></td>
</tr>
<tr>
<td>Wardrobe</td>
<td><img src="image" alt="Wardrobe Retrieval Results" /></td>
</tr>
</tbody>
</table>

Figure 3. Visual retrieval results on MI3DOR-2

<table>
<thead>
<tr>
<th>Query</th>
<th>Retrieval results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Curtain</td>
<td><img src="image" alt="Curtain Retrieval Results" /></td>
</tr>
<tr>
<td>Keyboard</td>
<td><img src="image" alt="Keyboard Retrieval Results" /></td>
</tr>
<tr>
<td>Sofa</td>
<td><img src="image" alt="Sofa Retrieval Results" /></td>
</tr>
<tr>
<td>Cup</td>
<td><img src="image" alt="Cup Retrieval Results" /></td>
</tr>
<tr>
<td>Bench</td>
<td><img src="image" alt="Bench Retrieval Results" /></td>
</tr>
<tr>
<td>Table</td>
<td><img src="image" alt="Table Retrieval Results" /></td>
</tr>
</tbody>
</table>
Figure 4. The impact of $\lambda_2$ on the testing set of MI3DOR

![Figure 4]

Figure 5. The impact of $\lambda_3$ on the testing set of MI3DOR

![Figure 5]
ACKNOWLEDGMENT

This work was supported in part by the National Nature Science Foundation of China (61902277, 61772359), the China Postdoctoral Science Foundation (2021T140511, 2020M680884), the grant of Tianjin New Generation Artificial Intelligence Major Program (19ZXZNGX00110) and the Baidu Program.
REFERENCES


Jinghui Chu received the B.Eng. degree in radio technology, and M.Eng. and PhD. degrees in signal and information processing all from Tianjin University, Tianjin, China, in 1991, 1997, and 2006 respectively. She is currently an associate professor in the School of Electrical and Information Engineering, Tianjin University. Her teaching and research interests include digital video technology and pattern recognition.

Zhao Xiaoqian is currently pursuing the master’s degree in Tianjin University, Tianjin, China. Her research interests include computer vision and machine learning.

Dan Song received the Ph.D. degree in computer science and technology from Zhejiang University, China. She is currently an associate professor with the School of Electrical and Information Engineering, Tianjin University. Her research interests include computer graphics and computer vision.

Wen-Hui Li received the Ph.D. degree in the School of Electrical and Information Engineering, Tianjin University. He was an Intern Student with the SeSaMe Center, National University of Singapore. His research interests include computer vision, machine learning, and 3D model retrieval.

Shenyuan Zhang received his Master’s degree from the Communications University of China. Currently, she is a product director of the New Media Center of People’s Daily, and has rich experience in searching and recommending user products. Also she has researched in mobile terminal.

Xuanya Li received the PhD. degree from Beijing Institute of Technology, Beijing, China, in 2012. He is currently the director of Baidu Campus and the executive member of China Computer Federation. His main research interests include Internet of Things and artificial intelligence.

An-An Liu received the Ph.D. degree in electronic engineering from Tianjin University, China. He is currently a Professor with the School of Electrical and Information Engineering, Tianjin University. He was a Visiting Professor with the SeSaMe Center of National University of Singapore, and a Visiting Scholar with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA. His current research interests include computer vision and machine learning.