Computer Vision and Image Understanding 238 (2024) 103858

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Adaptive semantic transfer network for unsupervised 2D image-based 3D model retrieval

Dan Song [a,b], Yuanxiang Yang [a], Wenhui Li [a,*], Zhuang Shao [c], Weizhi Nie [a], Xuanya Li [d], An-An Liu [a,b]

[a] The School of Electrical and Information Engineering, Tianjin University, Tianjin 300110, China
[b] The Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230088, China
[c] Warwick Manufacturing Group, University of Warwick, CV4 7AL, United Kingdom
[d] Baidu Inc., Beijing, China

ARTICLE INFO

Communicated by Jingyi Yu

Keywords:
Semantic alignment
Unsupervised domain adaptation
Unsupervised 2D image-based 3D model retrieval

ABSTRACT

Unsupervised 2D image-based 3D model retrieval has been a highlighted research topic to enable flexible retrieval from 2D photos to 3D shapes. Although what methods we have so far have been great progress in this aspect, it still exists some issues in learning discriminative features and to well align the distribution diversity from various domains because of the huge cross-domain interval. According to our paper, we propose an adaptive semantic transfer network (ASTN) to improve the discrimination of feature representations and conveniently narrow the discrepancy of different domains by utilizing the intermediate domain to conduct semantic alignment. Our ASTN composes of the adaptive feature encoding module (AFE) and the dynamic semantic alignment module (DSA). To improve the quality of feature representation, the AFE module deploys a new strategy that trains the learnable parameters on multiple convolutional layers, which can adaptively pay different attentions to these layers. The DSA module dynamically constructs an intermediate domain that aims to convert the familiar direct alignment into the sum of two alignments which are source-intermediate and target-intermediate alignments, effectively narrowing the domain gap and further realizing the semantic alignment. On two arduous datasets, MI3DOR-1 and MI3DOR-2, we design abundant experiments that have demonstrated the effectiveness of our suggested method.

1. Introduction

With the 3D technology's high-speed improvement and the progress of computer modeling, 3D models have made wide use in diverse fields, including digital entertainment (Wong et al., 2007), medical diagnosis (Guetat et al., 2006), computer-aided design, augmented reality and e-business applications (Gao et al., 2017). Various applications generate massive 3D models and pose a great challenge for effective and convenient 3D model retrieval. Hence, the above retrieval's assignment is to give a query and design an algorithm to discover the similar target 3D model. The formats of query include point cloud, multiple views, 2D images and so on. As the 2D photos are handier to be obtained, the 2D image-based 3D model retrieval has drawn relevant workers' attentions to multimedia and computer vision fields.

In addition to the modality of the query model, how to train the feature extractor also becomes a hot topic. The effective 3D model retrieval methods (Qi et al., 2017a; Wu et al., 2015; Furuya and Ohbuchi, 2016; Feng et al., 2019) always train the deep neural network with extensive human-labeled information, which requires expensive labor costs and limits the broad application of 3D models to a large extent. Inspired by the unsupervised domain adaptation strategy (Ganin and Lempitsky, 2015; Wang et al., 2018; Zhang et al., 2017; Fu and Liu, 2022; Xu et al., 2022), unsupervised 2D image-based 3D model retrieval makes use of 2D images regarded as queries and trains the deep learning network through the unsupervised way by transferring the ability from the labeled 2D source domain to the 3D target domain, which has no label. Specifically, MEDA (Wang et al., 2018) dynamically learned the classifier to gain the domain-invariant features in the Gaussian manifold and at the same time aligned the conditional distributions from various domains. JGSA (Zhang et al., 2017) mapped the data features of each domain into a uniform region to lessen geometric and distributed displacement. DANN (Ganin and Lempitsky, 2015) used an adversarial domain adaptation algorithm to confound the information of different domains so as to lessen the discrepancy of domains. DRPC (Yue et al., 2019) adopted the image translation method to further improve domain adaptation by performing the pixel-level alignment. Aiming to align the gap of features' distribution,

* Corresponding author.
E-mail address: liwenhui@tju.edu.cn (W. Li).

https://doi.org/10.1016/j.cviu.2023.103858
Received 23 February 2023; Received in revised form 8 September 2023; Accepted 8 October 2023
Available online 11 October 2023

SSDDA (Chen et al., 2021) adopted a strategy that can mix the image-level feature to accomplish domain adaptation. SAN-SAW (Peng et al., 2022) paid attention to the alignment in the category-level center between the features belonging to the diverse styles and enforced it to achieve the distributed alignment for the center features that had been already aligned.

Although the existing algorithms have made significant breakthroughs in the unsupervised 2D image-based 3D model retrieval task, there still exist two shortcomings that have been ignored in many works: (1) the existing methods only attempt to align the different domains in high-level semantic space and ignore the low-level feature alignment, which increases the difficulty of learning domain-invariant features. For example, existing methods usually constrain the outputs of the fully connected layer before the classifier conducts domain adaptation while not imposing any operation on the convolutional layer. (2) most methods directly align the distributions derived from several domains without fully exploring the semantic correlations of these domains. However, the huge discrepancy between two domains causes the inability to effectively reduce the domain gap in direct alignment.

To settle the aforesaid problems, we recommend a fresh adaptive semantic transfer network (ASTN) approach for the responsibility of unsupervised 2D image-based 3D model retrieval. This proposed approach can effectively overcome the limitations of previous methods to lessen the difficulty of learning domain-invariant features and improve the performance of unsupervised 2D image-based 3D model retrieval. Specifically, this proposed ASTN approach mainly contains two modules: an adaptive feature encoding module and a dynamic semantic alignment module. The adaptive feature encoding module aims to pay different attentions on the features encoded by different convolutional layers, which can be implemented by multiplying different trainable parameters on corresponding convolutional layers. The dynamic semantic alignment module is performed by building an intermediate domain alignment to work out the difficulty that the discrepancies between different representations of data are too large to be well aligned. In short, the prime contributions of our approach are as follows:

- We recommend a fresh adaptive feature encoding module to mine diverse information by adaptively learning the weights of different convolutional layers. This module can effectively enhance the cross-domain feature representation in shallow spaces and reduce the low-level differences of cross-domain data.
- A novel dynamic semantic alignment module is designed specifically for domain adaptation, which converts the alignment between different domains into the new alignments that both source and target are close to the intermediate domain. This module can address the problem that the data discrepancy between different styles of domains is too large to become directly aligned and fundamentally reduce the difficulty of conventional cross-domain alignment.
- Extensive experiments establish that our proposed network outperforms others in two challenging datasets of unsupervised 2D image-based 3D model retrieval.

2. Related works

2.1. 3D model retrieval

The ambition of the 3D model retrieval task is to measure the affinity between the query and the other samples in the 3D dataset, and return the relevant 3D model in accordance with the affinity order. The typical 3D model retrieval methods are mainly segregated into two approaches. One is model-based 3D model retrieval methods and the other one is image-based 3D model retrieval methods. With regard to the model-based methods, they usually extract features of 3D formats, for example, point cloud and voxel. To be specific, VoxNet (Maturana and Scherer, 2015) obtained 3D shapes represented by binary voxels through the convolutional neural network. 3D ShapeNets (Wu et al., 2015) represented a model on a 3D voxel grid by using probability distributions of binary variables. PointNet++ (Qi et al., 2017b) utilized the points that are close together at multiple scales to apprehend the local structural elements. With regard to image-based methods, their core lies in the use of 3D models expressed in the form of multi-view, that is, 2D images. MVCNN (Su et al., 2015) used the traditional convolutional neural network to glean features from different views individually and next adopted the max-pool operation to collect the comprehensive feature of the 3D model. PCNN (Gao et al., 2021) utilized patch-level features to make the extracted multi-view features contain spatial features. Aggregating VF (Watanabe et al., 2021) employed a minimal number of multiple views in each 3D shape to get a compact representation. However, the existing 3D model retrieval methods mainly include supervised and unsupervised methods. Most of these methods take the object models as the inputs, which is highly inconvenient for users. Compared with the object models, 2D images are easily accessible to users. Therefore, there has been a strong focus on how to use domain adaptation techniques to learn relevant knowledge on a large number of marked 2D images and assist in processing the 3D model retrieval task.

2.2. Domain adaptation

The ambition of domain adaptation (DA) is to provide the ability that can convert the awareness from the one domain with many samples to the other domain with few samples. The domain adaptation methods are divided into unsupervised, semi-supervised, and supervised domain adaptation. Among them, the supervised domain adaptation methods consider all samples should be marked. For example, DDCH_{ms} (Yu et al., 2022) supervised the feature learning process with the joint supervision of multifarious similarities to learn discrete hashing code for cross-modal retrieval with multiple supervision. Semi-supervised domain adaptation (SSDA) methods exploit not only source data but also a few target data with labels during training. For example, SDSS (Kim et al., 2022) eliminated irrelevant source domain samples and benefited from a subset of source domain data only containing samples highly relevant to the target domain. LIRR (Li et al., 2021) jointly learned invariant depictions and shared optimal predictions to alleviate the certainty disparity across domains for exceptional generalizations. ASFS (Yu et al., 2019) devised the label graph propagation to explore the potential relationship between two domain's data for label propagation, and it kept the geometric structure consistency between raw feature and label spaces. Unsupervised domain adaptation (UDA) methods aim to lessen the domain divergence with the different domains' images, in which the target domain is without any labels on the target images, like (Soviany et al., 2021; Dubourvieux et al., 2022). A popular strategy for the UDA methods is to grasp domain-invariant features. For instance, DANN (Ganin and Lempitsky, 2015) constructed a method to extract domain-invariant information based on the convolutional neural network through the adversarial network. DLEA (Zhou et al., 2019) considered alignment at the class level by aligning the category centroids of the diverse domains based on eliminating the offsets at the domain level. JAN (Long et al., 2017) used the joint maximum mean difference criterion and utilized multiple domain-specific layers for adaptation. For the sake of solving the domain displacement problem, MultiLayerDA (Li et al., 2019) made the maximum mean divergence between multi-layer and multi-kernel on different domains. AUDA (Wang et al., 2020) designed a robust deep neural network to transform all domains to a mutual feature area and made the classifier obtained on the source domain apply successfully to the target dataset. Continual UDA (Saporta et al., 2022) sequentially trained a model on new target domains, while maintaining its performance on the previous target domains without forgetting. HRDA (Hoyer et al., 2022) was trained with pseudo-labels aimed to coalesce from multiple resolutions, and used an overlapping sliding window mechanism to advance the

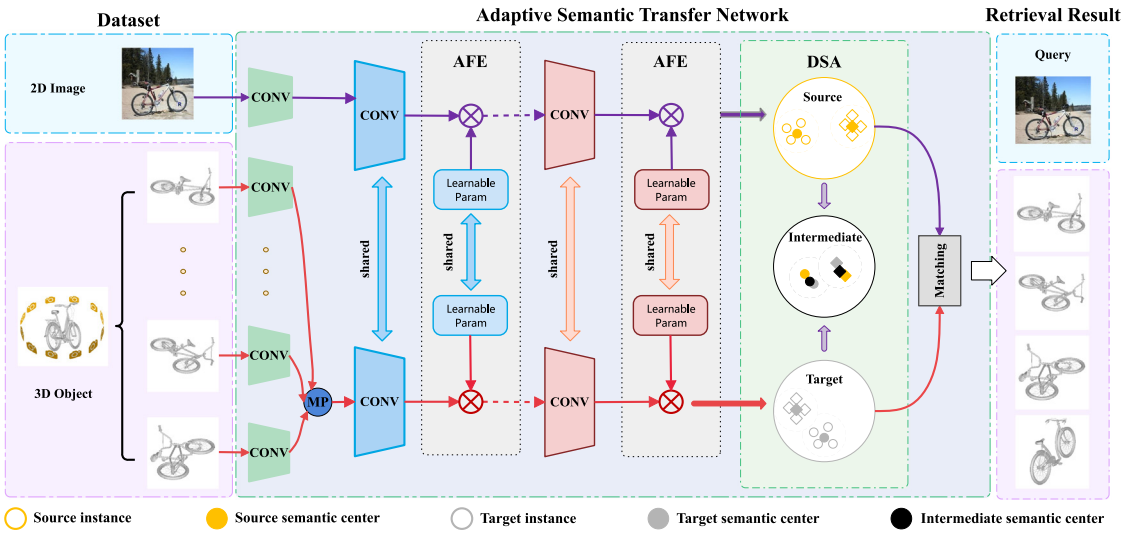


Fig. 1. The general structure of the proposed ASTN method. It first learns the basic feature encoding from the bottom convolutional layers. Then the AFE modules adaptively learn different features, thus assigning learnable parameters Ψ to different low-level representations. Next, the DSA module implements an indirect alignment strategy that makes the source or target domain semantic centers close to the intermediate domain. Finally, instance features from the source with 2D images and the target with 3D models have been matched to output the retrieval results. MP is the Max-pooling layer and Conv is the convolutional layer.

robustness of fine-grained pseudo-labels. Wang et al. (2022) considered the relative relationships of sample pair distance and assigned different weights to these pairs as the target pseudo labels instead of clustering results, which effectively increases the reliability of these pseudo labels. However, most unsupervised domain adaption methods ignore the mining of low-level feature information, which increases the difficulty of domain-invariant feature learning. Moreover, those methods usually directly reduced the discrepancy between two domains, and ignored the problem that the discrepancy cannot be well aligned due to the large difference between domains. Therefore, we propose an ASTN approach to explore the low-level feature information and explore the intermediate semantic information to facilitate the domain alignment.

3. Method

3.1. Overview

The objective of unsupervised 2D image-based 3D model retrieval is to design the cross-domain algorithm so that we can exactly find similar 3D models when we offer a 2D image as the input from the source dataset. In this task, the 2D image data of the source domain contains the images and their labels, denoted as $S = \{X^s, Y^s\}$, while the 3D model data of the target domain only contains the 3D models without labels, denoted as $T = \{X^t\}$. To address the huge domain discrepancy without target label guidance, we propose the adaptive semantic transfer network method to gain the domain-invariant features. Our method first assigns the adaptive weights to the low-level convolutional layers through the adaptive feature encoding module to mine different low-level information, which can form better sample representations. Moreover, according to the sample representations of different domains on the same categories, our dynamic semantic alignment module can adaptively construct an intermediate domain to achieve indirect alignment at the semantic-level. At the same time, the dynamic semantic alignment module can solve the problem that the data samples are too different to be aligned directly. The overall idea of our access is decorated in Fig. 1 and our work mainly contains two modules including adaptive feature encoding and dynamic semantic alignment:

Adaptive Feature Encoding (AFE): The existing methods only attempt to align the different domains in high-level semantic space while ignoring the importance of low-level features for alignment, which increases the difficulty of domain-invariant feature learning. To deal with this problem, we introduce an adaptive feature encoding module to explore discriminative features in low-level layers. This module plays an important role in paying different attentions on different convolutional layers to capture more informative representations. The detail information is introduced in Section 3.2.

Dynamic Semantic Alignment (DSA): Most methods directly align the distributions of different domains without fully exploring the semantic correlations of these domains because of the domain discrepancy. However, the huge discrepancy between two domains causes the inability to effectively reduce the domain gap in direct alignment. Dynamic Semantic Alignment generates the intermediate domain by automatically setting the weights of source features and target features. Thus, it is able to convert the current source-target domain alignment into a new alignment method that all domains are aligned to the same new domain we build, which can adequately lessen the problem of domain alignment.

3.2. Adaptive feature encoding

Given a 2D data included in the source dataset, set as $x^s \in X^s$ with label $y^s \in Y^s$. At the same time, the 3D model belonging to the target dataset is denoted as $x^t \in X^t$ without any label. The task of the unsupervised 2D image-based 3D model retrieval aims to learn the discrimination and domain-shared features so that the instances attached to the identical class from cross domains can be well matched. Therefore, we first introduce how to encode the features to represent the cross-domain characteristics. The conventional methods usually use the convolutional neural network for feature extraction, but these works do not take into account the importance of different conventional layers. To this end, we propose the adaptive feature encoding module to train the parameters to dynamically consider the importance of the features encoded from different convolutional layers. The output of each convolutional layer can be formulated by:

$$\left(x^d\right)^{(i)} = \Psi^{(i)} \cdot G^{(i)}\left(\left(x^d\right)^{(i-1)}\right) \tag{1}$$

where i means the ith convolutional layer, G is the feature extractor and $d \in \{s, t\}$. Where $\Psi^{(i)}$ approximates a dynamic weight calculation, which can adaptively focus on the feature encoding with the training procedure. The calculation of $\Psi^{(i)}$ is expressed as follows:

$$\Psi^{(i)} = \left(1 + e^{-w^{(i)} \cdot h^{(i)}}\right)^{-1} \tag{2}$$

where $h^{(i)}$ is the trainable unbounded model parameter and $w^{(i)}$ controls the plasticity of activation. We set $w^{(i)} = 1 - q$, where q is the proportion of the total number of training times to the current training time. The value of the plasticity varies between extreme 0 and 1.

It is well-known that the bottom layers of the convolutional neural network extract the low-level characteristics of the sample. To reconsider this information ignored by the conventional methods, the dynamic weights of these layers are applied in our adaptive feature encoding module. As shown in Eq. (1). the weights of different layers are adaptively changed with the training processing. By calculating trainable weights for different convolutional layers, we obtain the sample descriptors to better represent the domain-invariant characteristic. Because of the optimization of domain-invariant feature extraction, that is, the optimization of low-level feature learning, we can construct more appropriate semantic-level category centers through instance-level features, so as to improve the knowledge transfer at the semantic-level. Besides learning discriminative representations, another key step is how to correctly measure cross-domain similarity, which will be further discussed in the next section.

3.3. Dynamic semantic alignment

The AFE module considers the different contributions of different convolutional layers, which improves the discrimination of two-domain features. In this section, we propose the dynamic semantic alignment module to align the encoded features. DSA module translates the direct domain alignment into the source domain to the intermediate domain and the target domain to the intermediate domain alignments. It also presents that the DSA module aims to lessen the huge trouble of direct cross-domain alignment. DSA module mainly consists of the intermediate domain construction and cross-domain semantic alignment.

Intermediate Domain Construction: The traditional direct alignment (Zhou et al., 2019; Xie et al., 2018) aimed to narrow the gap of cross domains by using source labels and pseudo target labels predicted by the category classifier f. The pseudo labels of target samples can be formulated as:

$$\hat{y}_i^t = Labeling\left(G, f, x_i^t\right) \tag{3}$$

where x_i^t represents samples of target domain, G is feature extractor, and f can be implemented as the category classifier. $Labeling(\cdot,\cdot,\cdot)$ is a pseudo-label generation process for the target samples. The traditional direct alignment considers the category-level semantics of two domains, which are given by:

$$C_k^S \leftarrow \frac{1}{\|S_k\|} \sum_{(x_i^s, y_i^s) \in S_k} G(x_i^s) \tag{4}$$

$$C_k^T \leftarrow \frac{1}{\|T_k\|} \sum_{(x_i^t, \hat{y}_i^t) \in T_k} G(x_i^t) \tag{5}$$

where C_k^S performs the kth center of the source domain in the category-level and C_k^T performs the kth center of the target domain in the category-level. And i represents the ith instance among the kth category. $\|\cdot\|$ is prepared as the digit of instances in the same category.

The traditional methods constrained the distance between source and target category-level centers. However, these methods ignored that the distribution of different domains is too huge that the direct alignment hardly narrows the domain gap. To figure out this problem, we come up with generating the intermediate domain by weighting the source and target centers. Thus, the direct alignment of those domains is converted to the indirect alignment of source-intermediate and target-intermediate domains. After obtaining the respective category-level centers of two domains, we utilize the numbers of different domains' samples belonging to the same class within the batch as the weights to generate the category-level centers of the intermediate domain C_k^I, where I means the intermediate domain. As shown in Fig. 2,

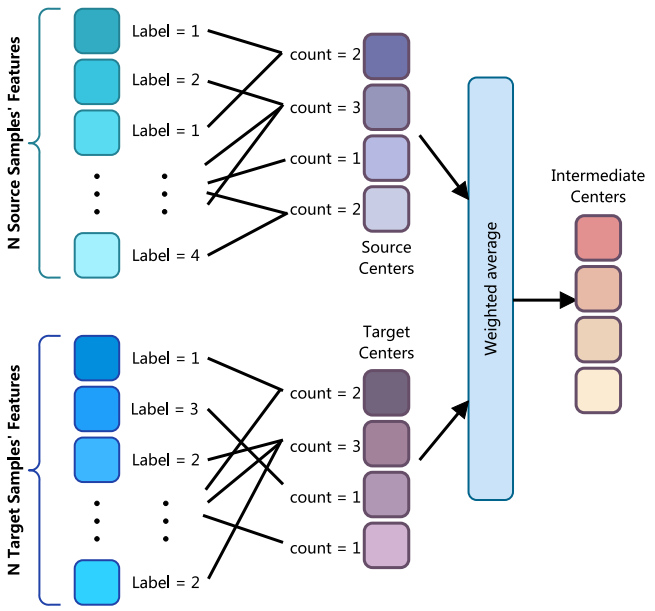


Fig. 2. Intermediate domain construction flow.

we divide the source domain (target domain) sample features of each batch according to the source domain real labels (target domain pseudo labels) and obtain the source domain (target domain) category-level center features by meaning. Then, the category-level center features representation of the intermediate domain is obtained by the weighted average of two domains' the category-level center features according to the corresponding number of samples in the same category. The intermediate centers are adaptively weighted by the category-level centers of source and target domains, which are mathematically formulated by:

$$C_k^I \leftarrow \frac{\|S_k\|}{\|S_k\| + \|T_k\|} C_k^S + \frac{\|T_k\|}{\|S_k\| + \|T_k\|} C_k^T \tag{6}$$

Cross-domain semantic alignment. By generating the intermediate centers, we translate the difference between C_k^S and C_k^T into C_k^S, C_k^I, and C_k^I, C_k^T to optimize the procedure of cross-domain semantic alignment. As shown in Fig. 1, the orange and gray solid graphs represent the data of the source and target domains respectively, and the centers of the intermediate domain are generated between the two domains, which palliates the difficulty of domain alignment. Moreover, as the training process evolves, the semantic centers are dynamically changed because of the batch training strategy. To preserve global semantics as well as the semantics of different batches, we dynamically merge the global category-level center representation with the category-level center representation of the current batch in a certain proportion, which can be expressed as:

$$C_k^S \leftarrow \theta C_{k_b}^S + (1-\theta) C_k^S \tag{7}$$

$$C_k^T \leftarrow \theta C_{k_b}^T + (1-\theta) C_k^T \tag{8}$$

$$C_k^I \leftarrow \theta C_{k_b}^I + (1-\theta) C_k^I \tag{9}$$

where C_k^S, C_k^T, C_k^I represent the global kth global semantic centers of the source, target and intermediate domains. $C_{k_b}^S$, $C_{k_b}^T$, $C_{k_b}^I$ represent the kth global semantic centers of these domains on current batch b. The θ is a hyper-parameter and it aims to adjust the contribution of current semantic centers. After obtaining the updated semantic centers, we further reduce the distinction between different domains by minimizing the distance metric of the category-level center features between the intermediate domain and the other domains, respectively.

3.4. The objective loss function

Following the conventional domain adaptation methods, we adopt classification loss $\mathcal{L}_{ce}$ and domain classification loss $\mathcal{L}_{DC}$ to optimize the whole network. Classification loss $\mathcal{L}_{ce}$ is applied on the source domain to capture class-level representations. Domain classification loss $\mathcal{L}_{DC}$ is trained in a max–min adversarial manner to obtain domain-invariant representations. We also design the semantic loss $\mathcal{L}_s$ to facilitate the domain alignment, which is explained as follows:

$$\mathcal{L}_s = \sum_{k=1}^{K} \phi(C_k^S, C_k^T) + \alpha \left(\sum_{k=1}^{K} \phi(C_k^S, C_k^I) + \sum_{k=1}^{K} \phi(C_k^T, C_k^I) \right) \tag{10}$$

where α is the hype-parameter to balance the importance of direct and intermediate alignments, which is validated in the experiment. C_k^S, C_k^T, and C_k^I indicate global category-level centers in the source dataset, target dataset, and intermediate dataset, individually. $\phi(\cdot,\cdot)$ is the similarity measure, prescribed as the Euclidean distance to reflect the difference between two domains. For the overall loss of this method, we define it as follows:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda \left(\mathcal{L}_{DC} + \mathcal{L}_s \right) \tag{11}$$

where λ is used to control the contributions among the loss of the source domain's classification and others. We set $\lambda = \frac{2}{1+e^{-10 \cdot p}} - 1$, where p stands for the training progress and its value varies between 0 and 1. Our method's pseudo code is displayed in Algorithm 1. The intermediate domain construction method is intrinsically a weighted process. The center of the intermediate domain in each batch contains the information of two domains by adding source centers and target centers proportionally. The center update aims to merge the global semantic representations with the current semantic representation in a certain proportion. Then, the domain alignment operation is put into effect on these semantic depictions. Finally, we extract the features of each instance by the feature extraction G, and utilize the Euclidean distance to measure their similarities.

Algorithm 1 One step of our proposed method optimization.

Require: Feature extractor G, Classifier f, Source domain S, Target domain T, K is the number of classes, Training classifier f, Global centers for three domains: $\{C_k^S\}_{k=1}^K$, $\{C_k^T\}_{k=1}^K$ and $\{C_k^I\}_{k=1}^K$

Ensure: G and f

1: Initialize the parameters of G and f.
2: Random select the samples of source and target domains to construct the current batch data $\{x^s, y^s\}$ and x^t.
3: Obtain pseudo labels $\hat{y}_i^t$ of x^t by Eq. (3).
4: **for** $k = 0 \rightarrow K$ **do**
5: Compute k-th source semantic representation by Eq. (4)
6: Compute k-th target semantic representation by Eq. (5)
7: Compute k-th intermediate semantic representation by Eq. (6)
8: Update these semantic representations by Eq. (7)-Eq. (9)
9: **end for**
10: Compute the overall loss $\mathcal{L}_s$ by Eq. (10) and Eq. (11)
11: Update the G and f

4. Experiment and analysis

To prove our ASTN approach's reliability, we carry out some studies on two popular datasets. In this part, we first recommend the datasets and experiment settings, and next, we analyze the retrieval performance with other representative methods. At the same time, we structure ablation tests to evaluate the convincingness of our approach. Finally, we visualize the retrieval results and conduct qualitative analysis.

4.1. Datasets and evaluation criteria

4.1.1. Dataset

We put into practice variance experiments on two frequently-applied datasets for domain adaptation. The source dataset contains 2D images of different categories, and the target dataset contains multiple 3D models represented by 12 views.

- **MI3DOR-1** (Zhou et al., 2019): The MI3DOR-1 dataset makes up of 21,000 2D real images and 7690 3D virtual shapes. The dataset includes 21 categories. There are 10,500 images and 3845 models in the training set, while 10,500 images and 3845 models are in the testing set.
- **MI3DOR-2** (Zhou et al., 2019): The MI3DOR-2 dataset makes up of 2D real images and 3D virtual models belonging to 40 categories, where the number of 2D images is 19,694 and the number of 3D models is 3982. There are 19,294 images and 3182 models in the training set, while 400 images and 800 models are in the testing set.

4.1.2. Evaluation criteria

To appraise the feasibility of our network, we select seven popular evaluation criteria as (Zhou et al., 2019) in our experiment. NN can calculate the retrieval results' certainty returned by the nearest neighbor, which refers to the model's accuracy most similar to the retrieval target. FT and ST are described as the return values of the first K and 2 K items, respectively, where K means the number of related objects retrieved. These two indicators can evaluate the retrieval results' recall rate. The F-measure is used to cooperatively assess the accuracy and return rate of the results of the former K item. DCG is equivalent to a statistical method that accredits interrelated results to the top position with a higher weight without considering the lower results. AUC value can represent the overall quality of the system. ANMRR is on behalf of a rank-based measure which evaluates the ordering message of related objects in the retrieved object. The higher value indicates better performance in terms of all evaluation criteria except ANMRR.

4.2. Implementation details

We employ the AlexNet that is trained in advance on the ImageNet dataset for gleaning the image features of each 2D image or multi-view of each 3D model. The 3D model multi-view feature representations are merged as a tight 3D model characterization by the Max-pooling function. During training, we set the momentum to 0.9 and optimized the network using Stochastic Gradient Descent (SGD). $\mu_p = \frac{\mu_0}{(1+\tau \cdot p)^\beta}$ represents as the learning rate, where μ_0 is the initial learning rate intended to be 0.01, $\tau = 10$ and $\beta = 0.75$. For the balance hype-parameter, we adopt $\lambda = \frac{2}{1+e^{-\zeta \cdot p}} - 1$, where the value of ζ is 10 and p means the training progress. We prescribe the batch size to 76 for every domain. Furthermore, the maximum epoch is set to 20,000. Our model is implemented on the PyTorch framework with one NVIDIA 3080Ti GPU.

4.3. Retrieval performance

4.3.1. Compared methods

To validate the presentation of domain alignment, we mainly analyze our recommended approach with a few representative methods:

Basic deep learning methods, such as AlexNet (Krizhevsky et al., 2012). This method directly extracts features of 3D multiple views and 2D images by convolution neural network without any transfer learning.

Traditional transfer learning methods, such as MEDA (Wang et al., 2018) and JGSA (Zhang et al., 2017). MEDA jointly trained the domain-invariant classifier in the Grassmann manifold and implemented dynamic alignment of cross-domain distributions in the manifold. JGSA

Table 1
Comparison results with other methods on MI3DOR-1. The best results are in bold.

Methods	Retrieval						
	NN↑	FT↑	ST↑	F↑	DCG↑	ANMRR↓	AUC↑
AlexNet (Krizhevsky et al., 2017)	0.424	0.323	0.469	0.099	0.345	0.667	–
DANN (Ganin and Lempitsky, 2015)	0.650	0.505	0.643	0.112	0.542	0.474	–
JAN (Long et al., 2017)	0.446	0.343	0.495	0.085	0.363	0.647	–
JGSA (Zhang et al., 2017)	0.612	0.443	0.599	0.116	0.473	0.541	–
MEDA (Wang et al., 2018)	0.430	0.344	0.501	0.046	0.361	0.646	–
MSTN (Xie et al., 2018)	0.789	0.622	0.779	0.154	0.657	0.358	0.557
DLEA (Zhou et al., 2019)	0.764	0.558	0.716	0.143	0.597	0.421	–
HIFA (Zhou et al., 2020)	0.778	0.618	0.768	0.151	0.654	0.362	–
UCM-GCN (Liang et al., 2023)	0.781	0.574	0.723	**0.172**	0.617	0.398	–
HDA (Wen et al., 2022)	0.785	0.623	0.785	0.152	0.673	0.344	0.569
Ours	**0.790**	**0.647**	**0.797**	0.156	**0.679**	**0.334**	**0.576**

Table 2
Comparison results with other methods on MI3DOR-2. The best results are in bold.

Methods	Retrieval						
	NN↑	FT↑	ST↑	F↑	DCG↑	ANMRR↓	AUC↑
AlexNet (Krizhevsky et al., 2017)	0.518	0.355	0.488	0.355	0.383	0.629	–
DANN (Ganin and Lempitsky, 2015)	0.625	0.501	0.635	0.510	0.547	0.477	–
JAN (Long et al., 2017)	0.608	0.501	0.646	0.501	0.527	0.484	–
JGSA (Zhang et al., 2017)	0.585	0.405	0.533	0.405	0.433	0.577	–
MEDA (Wang et al., 2018)	0.570	0.392	0.523	0.392	0.425	0.590	–
MSTN (Xie et al., 2018)	0.750	0.573	0.700	0.583	0.614	0.405	0.485
DLEA (Zhou et al., 2019)	0.700	0.555	0.681	0.555	0.593	0.434	–
HIFA (Zhou et al., 2020)	0.725	0.570	0.710	0.570	0.598	0.413	–
HDA (Wen et al., 2022)	0.758	0.588	**0.742**	0.592	0.625	0.387	0.504
Ours	**0.763**	**0.597**	0.722	**0.609**	**0.636**	**0.380**	**0.510**

made a regulation on the coupled projections of source and target domains, and projected them into a universal low-dimensional subspace to reduce the geometric and distribution shift.

Deep transfer learning methods, e.g. JAN (Long et al., 2017), DLEA (Zhou et al., 2019), DANN (Ganin and Lempitsky, 2015), HIFA (Zhou et al., 2020), MSTN (Xie et al., 2018), UCM-GCN (Liang et al., 2023) and HDA (Wen et al., 2022). JAN utilized the maximum mean difference criterion and multiple domain-specific layers to reduce the domain discrepancy in an adversarial training strategy. DLEA implemented an adversarial domain loss and a center alignment loss to jointly realize the feature learning and distribution alignment. DANN introduced the adversarial learning into transfer learning for the first time, and focused on how to select transferable features between different domains. HIFA minimized the maximum mean feature discrepancy of two domains' distributions in order to conduct domain alignment. MSTN minimized the Euclidean distance between cross domain's category-level centers to achieve the semantic transfer learning. UCM-GCN proposed a novel unsupervised GCN model to align the feature distribution from different modalities, which applied the GCN to update the descriptors according to the cross-media graph. HDA used local features for migration and hierarchical alignment to connect different domains, which was re-implemented and transferred to our 3D model retrieval task.

4.3.2. Performance comparison

Following the existing method (Zhou et al., 2020), we employ seven criteria described above to evaluate the presentation of the unsupervised 2D image-based 3D model retrieval. The experimental results on two datasets are demonstrated in Tables 1 and 2, severally. From the results, we have three key observations:

Although the basic deep learning method has been trained in advance on a wholesale dataset (ImageNet), it is still too troublesome to eliminate the gap between different domains. All the transfer learning methods outperform this no-adaptation method.

The performances of these two tables show that the conventional methods (e.g. MEDA and JGSA) are superior to the no-adaptation method in terms of all evaluation criteria, indicating that coupling the two domains into the low latitude subspace and learning domain invariant classifier by Grassmann manifold can enhance domain alignment. However, these transfer methods cannot consider the semantic information during the alignment, which results in inadequate alignment. To solve this problem, our method not only realizes different concerns in the feature coding of different convolution layers, but also considers the semantic information to implement the indirect alignment, which shows the availability of our recommended method.

We discover that the representative deep transfer learning methods (Long et al., 2017; Zhou et al., 2019; Ganin and Lempitsky, 2015; Zhou et al., 2020; Xie et al., 2018; Liang et al., 2023) enable higher evaluation criterion than the traditional methods, showing that the domain-level alignment of the source domain and target domain, even the category-level alignment, by an adversarial training strategy, can further reduce the gap of domain differences. Yet, these methods ignore the problem that the distance of different domains is too large to reduce the domain gap effectively. Comparatively, ASTN translates the direct alignment of those domains into the indirect of source-intermediate and target-intermediate domains to further implement semantic migration. Therefore, ASTN outperforms the representative deep transfer learning methods.

Overall, the results of these retrieval performances consistently demonstrate that our recommended method achieves exceptional performance than the extant algorithms.

4.3.3. Ablation study

To verify the reliability of our mentioned modules, we devise the experiment of ablation study on MI3DOR-1. As shown in Table 3, when we add adaptive feature encoding module (AFE) on basic AlexNet, the retrieval performance is significantly improved, which demonstrates that adaptive weights of different convolutional layers indeed contribute to better domain alignment. When we adopt a dynamic semantic alignment module (DSA) on the basic network, the performance is further improved by a large margin. This observation verifies that the strategy in which all domains align to the intermediate domain can truly reduce the difficulty of cross-domain alignment. When these two modules are jointly adopted, it outperforms the architectures of

Table 3
Effectiveness with respect to different structures in our method. The best result is bold.

Methods	MI3DOR-1						
	NN↑	FT↑	ST↑	F↑	DCG↑	ANMRR↓	AUC↑
AlexNet	0.424	0.323	0.469	0.099	0.345	0.667	–
ASTN(Only AFE)	0.607	0.516	0.682	0.122	0.538	0.470	0.402
ASTN(Only DSA)	0.774	0.621	0.773	0.153	0.655	0.361	0.543
ASTN(DSA+AFE)	**0.790**	**0.647**	**0.797**	**0.156**	**0.679**	**0.334**	**0.576**

Table 4
Effectiveness with respect to different datasets in our method. The best result is bold.

Methods	testing on MI3DOR-1						
	NN↑	FT↑	ST↑	F↑	DCG↑	ANMRR↓	AUC↑
MI3DOR-2(P) + MI3DOR-1(F)	**0.793**	**0.651**	**0.800**	**0.158**	**0.681**	**0.330**	**0.578**
MI3DOR-1 + MI3DOR-2	0.762	0.614	0.760	0.143	0.648	0.353	0.542
Methods	testing on MI3DOR-2						
	NN↑	FT↑	ST↑	F↑	DCG↑	ANMRR↓	AUC↑
MI3DOR-1(P) + MI3DOR-2(F)	**0.769**	**0.601**	**0.767**	**0.612**	**0.640**	**0.373**	**0.518**
MI3DOR-1 + MI3DOR-2	0.723	0.567	0.718	0.589	0.603	0.410	0.491

Table 5
Complexity analysis including the FLOPs and number of parameters.

Method	Params	FLOPs	NN↑
AlexNet (Krizhevsky et al., 2017)	57.92M	723.19M	0.424
MSTN (Xie et al., 2018)	59.24M	724.51M	0.789
ASTN	59.24M	724.51M	0.790

Table 6
Adaptive feature encoding analysis on MI3DOR-1.

Method	Retrieval						
	NN↑	FT↑	ST↑	F↑	DCG↑	ANMRR↓	AUC↑
layer1–5	0.691	0.586	0.741	0.139	0.612	0.397	0.508
layer2–5	0.773	0.630	0.785	0.153	0.662	0.350	0.567
layer3–5	0.761	0.632	0.781	0.151	0.662	0.349	0.561
layer4–5	**0.790**	**0.647**	**0.797**	**0.156**	**0.679**	**0.334**	**0.576**
only layer5	0.765	0.621	0.770	0.154	0.655	0.360	0.555

using them alone, which shows our AFE and DSA modules can learn complementary information to facilitate domain alignment.

In addition, we try to perform training and testing together on the two datasets MI3DOR-1 and MI3DOR-2. We first place the same training samples of these two datasets together and merge different categories into the same training dataset, and the method is named MI3DOR-1+MI3DOR-2. MI3DOR-1+MI3DOR-2 was tested on MI3DOR-1 and MI3DOR-2, separately. In Table 4, we discover that due to the inconsistency of the number of samples in each category and the different sample distributions in these two datasets, the retrieval performance of 3D models is not improved. Moreover, we explore a new method that utilizes samples from these two datasets for training. In other words, we use the parameters obtained by pre-training (named P) on one dataset and fine-turning (name F) on the other dataset. For example, the model was pre-trained on MI3DOR-1 and fine-turned on MI3DOR-2, which was named MI3DOR-1(P)+MI3DOR-2(F). From Table 4, we can find that the retrieval performances are further improved when testing on both MI3DOR-1 and MI3DOR-2 dataset.

4.3.4. Computational complexity analysis

To further explore the complexity of our method, we count the number of parameters (Params) to represent space complexity and floating-point operations per second (FLOPs) to represent time complexity. The smaller number of parameters indicates the smaller model size. With regard to the FLOPs, the smaller value of the FLOPs represents the faster computing power. As shown in Table 5, we compare our approach with AlexNet and the representative MSTN. Compared with AlexNet, although the complexity of our method is slightly increased, our retrieval performance is greatly improved, which is due to our mining of low-level features and the implementation of the indirect alignment at the semantic-level. In addition, compared to the MSTN, we also dramatically improve the effectiveness of the semantic transfer without increasing the computational complexity.

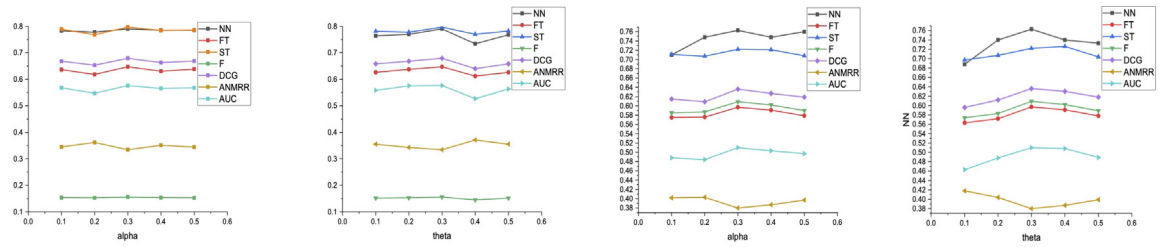
4.4. Discussion

4.4.1. Adaptive feature encoding analysis

To more appraise the reliability of our AFE module, we implement a detailed study on different trainable parameters of different convolutional layers. We set five combinations to add the trainable parameters on different layers, including layer 1 to layer 5, layer 2 to layer 5, layer 3 to layer 5, layer 4 to layer 5 and only layer 5. The experiment results are demonstrated in Table 6. By observation, we discover without a hitch that while the trainable parameters are added to lower convolutional layers, the performances are degraded. The reason for this observation might be that the lower layers contain more detailed clues and the comprehensive information without weighting can help the neural network encode more characteristics. When we add the AFE module on layer 4 and layer 5, our approach reaches the greatest achievement and obtains the gains of 9.9%, 6.1%, 5.6%, 1.7%, 6.7%, 6.3% and 6.8% in terms of seven criteria, which demonstrate the effectiveness of our AFE module.

4.4.2. Dynamic semantic alignment analysis

To explore the impact of our DSA alignment strategy, we examine the retrieval results between the alignment of source or target instances to intermediate centers, named *Instance*, and the alignment of source or target centers to intermediate centers, named *Center*. The results are displayed in Table 7. Compared to *Instance* strategy, we detect that the *Center* strategy enables higher retrieval performances than it, showing the superiority of introducing the novel semantic alignment method between source to intermediate domain and target to the intermediate domain. The reason why instance alignment retrieval is not effective is that semantic loss mutation is caused by excessive deviation of individual instances. However, the *Center* strategy eliminates the problem of large differences between individual instances by averaging the instance features. Besides, the t-SNE visualization results of the *Instance* strategy and the *Center* strategy are illustrated in Fig. 5(a) and (b), separately. It indicates more intuitively that the *Center* strategy makes the feature points of different classes disperse rather than mix, which further demonstrates the effectiveness of our DSA module.



(a) Parameter α on MI3DOR-1 (b) Parameter θ on MI3DOR-1 (c) Parameter α on MI3DOR-2 (d) Parameter θ on MI3DOR-2

Fig. 3. Evaluation criterion results by varying the parameter settings of α and θ on MI3DOR-1 and MI3DOR-2.

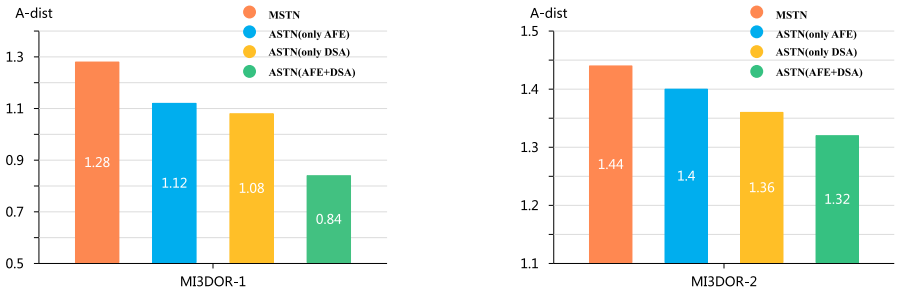


Fig. 4. A-distance results of the representative deep transfer learning MSTN and our method on MI3DOR-1 and MI3DOR-2.

Table 7
Comparison of retrieval performance on MI3DOR-1 and MI3DOR-2 between the instance and center strategies during semantic alignment.

Retrieval	MI3DOR-1		MI3DOR-2	
	Instance	Center	Instance	Center
NN↑	78.72%	**79.00%**	67.00%	**76.25%**
FT↑	63.32%	**64.71%**	51.19%	**59.73%**
ST↑	78.16%	**79.67%**	65.31%	**72.23%**
F↑	15.57%	**15.57%**	52.08%	**60.89%**
DCG↑	66.71%	**67.85%**	54.60%	**63.55%**
ANMRR↓	34.74%	**33.42%**	46.71%	**38.03%**
AUC↑	55.98%	**57.61%**	41.47%	**50.97%**

4.4.3. Parameter analysis

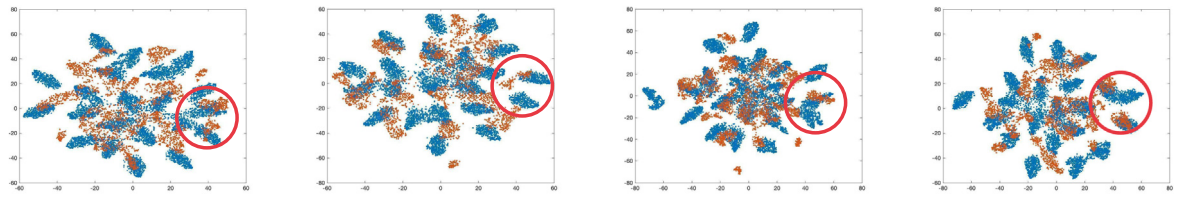
To appraise the leverage of hyper-parameters on the retrieval performance, we examine all evaluation criteria of our method by varying the parameter setting of θ and α. The parameter θ is the updated weight of the global category-level centers and the current category-level centers from each domain. Fig. 3(a) and (c) show the impact of θ on the MI3DOR-1 dataset and MI3DOR-2 dataset, from which we can see that when $\theta = 0.3$, our approach reaches the outstanding presentation on the two datasets we used. In particular, the transformation of θ has little effect on the values of retrieval criterion on the MI3DOR-1 dataset. This may be because the MI3DOR-1 dataset includes fewer categories and more instances in each class than the MI3DOR-2 dataset, so the category-level centers generated by our method are more stable. The parameter α intends to balance the loss of the direct alignment and the indirect alignment. Fig. 3(b) and (d) show the impact of α on both MI3DOR-1 and MI3DOR-2 datasets. For the MI3DOR-2 dataset, we observe that the retrieval performances first increase slowly with the growth of α, and then begin to decline. The best parameter settings of α are 0.3 in the above two datasets, which demonstrates the generalization in our method.

4.4.4. A-distance

To further assess the domain gap after training, we introduce A-distance to evaluate our method. A-distance (Ben-David et al., 2010) is an assessment method of domain discrepancy, illustrated as $d_A = 2(1 - 2\varepsilon)$, where ε is the error ratio of the trained domain classifier to identify which domain the instances come from. For comparison, we compare the deep representative transfer learning MSTN and our method with different modules. The results are presented in Fig. 4. We notice that MSTN obtains a farther distance than our method on those datasets, demonstrating the convincingness of each component in our proposed approach. Specifically, compared to MSTN, the A-distance of using our DSA greatly is greatly reduced, which indicates that a strategy in which both domains are aligned to the intermediate domain indeed addresses the adversity of directly source-target domain alignment. When using our AFE and DSA modules together, we obtain the minimized A-distance, which demonstrates that our method can well narrow the gap between two domains.

4.4.5. Feature visualization

To further prove the superiority of the recommended approach, the learned features of our proposed method are visualized by t-SNE on the MI3DOR-1 dataset. We set the 2D photo and the 3D shape as the blue point and the red point, respectively. We show the visualization results in Fig. 5. Specifically, Fig. 5(a) and (b) are further visualizations of the DSA module's strategies. We can discover from the red circle in the two figures that there is confusion among samples of different categories in Fig. 5(a), while samples in Fig. 5(b) can be clearly divided between categories. Moreover, Fig. 5(b), (c) and (d) are the visualization results of our proposed network with diverse modules. From the red circle in Fig. 5(b) and (c), we can easily notice that in Fig. 5(b), the classification of blue samples is clearer, while in Fig. 5(c), the boundary between categories is not so prominent but the distributions of different domains are more similar. In general, it can be seen that the transferable feature mapping when we only use the AFE module mixes up two colored points, which indicates that this module can indeed reduce the domain discrepancy. We also can see that when we only use the DSA module, we can easier classify the points, but the degree of mixing between the two colored dots is lower. Therefore, when we use both modules, we can observe from Fig. 5(d) that the two categories of source and target domains can be well aligned, and the categories can be clearly separated, that is to say, the samples from two domains are well classifier and the cross-domain distribution is well aligned, which further shows the effectiveness and complementary of the proposed modules.



(a) ASTN(Instance) (b) ASTN(only DSA) (c) ASTN(only AFE) (b) ASTN(AFE+DSA)

Fig. 5. Visualization results of 3D model and 2D image features with diverse modules by t-SNE on the MI3DOR-1 dataset.

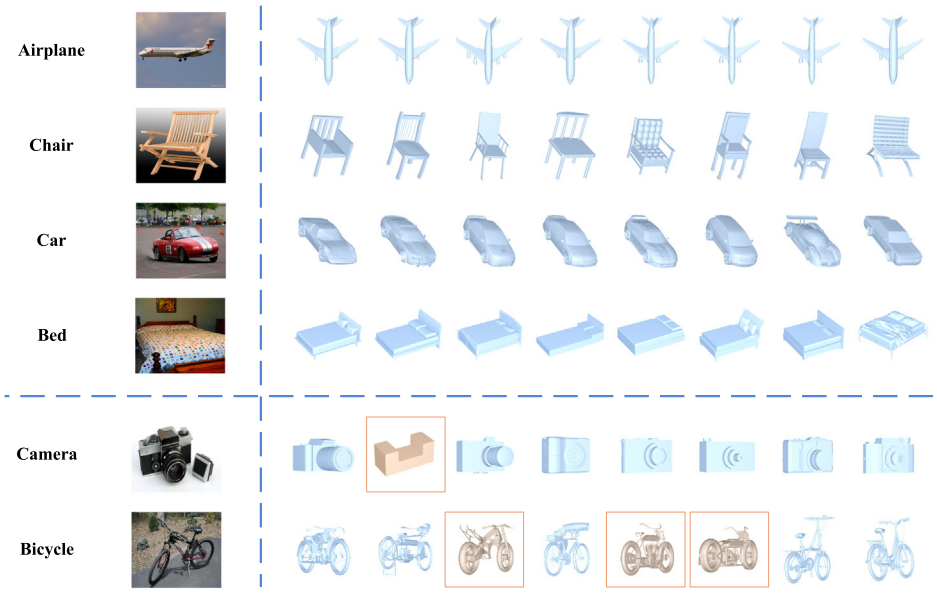


Fig. 6. Demonstration of a few retrieval instances on MI3DOR-1. The inquiry 2D image is arranged in the first column. Moreover, the top 8 retrieved objects are arranged in the right side towards the similarity sequence. The accurate, wrong retrieval samples are flagged as blue and orange colors, orderly. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.4.6. Qualitative retrieval results

The qualitative results are shown in Fig. 6. We randomly visualize six image-model retrieval results on the MI3DOR-1 dataset. The first column represents 2D image inquiries in various categories. Moreover, each row presents the top 8 results after the retrieval task. The erroneous retrieval is tagged as orange, while the accurate one is tagged as blue. From the retrieval examples, we can find that our method can find the matched results, and even though there are some mismatched samples in the retrieval list, they still have similar visual appearances to the query sample. For example, the motorcycles are mismatched when we use the bicycle as the query sample due to the motorcycles and bicycles sharing many same characteristics. This observation verifies the effectiveness of our method in settling the task of unsupervised 2D image-based 3D model retrieval.

5. Conclusion

In conclusion, we propose an adaptive semantic transfer learning method to enhance feature discrimination and further address the cross-domain gap. Our method has two modules, including adaptive feature encoding (AFE) and dynamic semantic alignment (DSA). The AFE devises a strategy to pay different attentions on different convolutional layers, which aims to adaptively utilize the information from different convolutional layers. The DSA translates the conventional domain alignment into the novel alignment in which all domains are aligned to the intermediate domain to address the difficulty of direct alignment. These two modules form complementary information for each other in terms of feature encoding and semantic alignment. The extensive experiment on the challenging datasets demonstrates the dependability of our recommended approach. In contrast, 3D models in real-world applications also contain complex backgrounds and rich visual features just like 2D images, which further increases the difficulty of the 3D model retrieval task. Therefore, it needs to be further explored how to enhance the correlation similarity between 3D model features corresponding to the 2D image features to obtain more efficient 3D model retrieval for real scenes.

CRediT authorship contribution statement

Dan Song: Conceptualization, Methodology. **Yuanxiang Yang:** Conceptualization, Writing – original draft. **Wenhui Li:** Methodology, Investigation, Writing – review & editing. **Zhuang Shao:** Formal analysis, Visualization. **Weizhi Nie:** Software, Validation. **Xuanya Li:** Supervision, Validation. **An-An Liu:** Supervision, Funding acquisition.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (U21B2024, U22A2068, 62202327), National Key Research and Development Program of China (2021YFF0704000), the China Postdoctoral Science Foundation, China (2022M712369) and the Baidu Program, China.

References

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W., 2010. A theory of learning from different domains. Mach. Learn. 79 (1–2), 151–175.

Chen, Y., Ouyang, X., Zhu, K., Agam, G., 2021. Semi-supervised domain adaptation for semantic segmentation. CoRR abs/2110.10639.

Dubourvieux, F., Audigier, R., Loesch, A., Ainouz, S., Canu, S., 2022. A formal approach to good practices in pseudo-labeling for unsupervised domain adaptive re-identification. Comput. Vis. Image Underst. 223, 103527.

Feng, Y., Feng, Y., You, H., Zhao, X., Gao, Y., 2019. MeshNet: Mesh neural network for 3D shape representation. In: The Thirty-Third AAAI Conference on Artificial Intelligence, the Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI, the Ninth AAAI Symposium on Educational Advances in Artificial Intelligence. EAAI, pp. 8279–8286.

Fu, Y., Liu, Y., 2022. Contrastive transformer based domain adaptation for multi-source cross-domain sentiment classification. Knowl.-Based Syst. 245, 108649.

Furuya, T., Ohbuchi, R., 2016. Deep aggregation of local 3D geometric features for 3D model retrieval. In: Wilson, R.C., Hancock, E.R., Smith, W.A.P. (Eds.), Proceedings of the British Machine Vision Conference. BMVC.

Ganin, Y., Lempitsky, V.S., 2015. Unsupervised domain adaptation by backpropagation. In: Bach, F.R., Blei, D.M. (Eds.), Proceedings of the 32nd International Conference on Machine Learning. ICML 2015, Lille, France, 6-11 July 2015, In: JMLR Workshop and Conference Proceedings, vol. 37, pp. 1180–1189.

Gao, Z., Shao, Y., Guan, W., Liu, M., Cheng, Z., Chen, S., 2021. A novel patch convolutional neural network for view-based 3D model retrieval. arXiv: Computer Vision and Pattern Recognition.

Gao, Y., Zhang, H., Zhao, X., Yan, S., 2017. Event classification in microblogs via social tracking. ACM Trans. Intell. Syst. Technol. 8 (3), 35:1–35:14.

Guetat, G., Maitre, M., Joly, L., Lai, S., Lee, T., Shinagawa, Y., 2006. Automatic 3-D grayscale volume matching and shape analysis. IEEE Trans. Inf. Technol. Biomed. 10 (2), 362–376.

Hoyer, L., Dai, D., Gool, L.V., 2022. HRDA: Context-aware high-resolution domain-adaptive semantic segmentation. CoRR abs/2204.13132.

Kim, D., Seo, M., Park, J., Choi, D., 2022. Source domain subset sampling for semi-supervised domain adaptation in semantic segmentation. CoRR abs/2205.00312.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25, 1097–1105.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2017. ImageNet classification with deep convolutional neural networks. Commun. ACM 60 (6), 84–90.

Li, B., Wang, Y., Zhang, S., Li, D., Keutzer, K., Darrell, T., Zhao, H., 2021. Learning invariant representations and risks for semi-supervised domain adaptation. In: IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2021, Virtual, June 19-25, 2021, Computer Vision Foundation / IEEE, pp. 1104–1113.

Li, X., Zhang, W., Ding, Q., Sun, J.-Q., 2019. Multi-layer domain adaptation method for rolling bearing fault diagnosis. Signal Process..

Liang, Q., Li, Q., Nie, W., Liu, A., 2023. Unsupervised cross-media graph convolutional network for 2D image-based 3D model retrieval. IEEE Trans. Multim. 25, 3443–3455.

Long, M., Zhu, H., Wang, J., Jordan, M.I., 2017. Deep transfer learning with joint adaptation networks. In: Precup, D., Teh, Y.W. (Eds.), Proceedings of the 34th International Conference on Machine Learning. ICML, In: Proceedings of Machine Learning Research, vol. 70, pp. 2208–2217.

Maturana, D., Scherer, S., 2015. VoxNet: A 3D convolutional neural network for real-time object recognition. Intell. Robots Syst..

Peng, D., Lei, Y., Hayat, M., Guo, Y., Li, W., 2022. Semantic-aware domain generalized segmentation. CoRR abs/2204.00822.

Qi, C.R., Su, H., Mo, K., Guibas, L.J., 2017a. PointNet: Deep learning on point sets for 3D classification and segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pp. 77–85.

Qi, C.R., Yi, L., Su, H., Guibas, L.J., 2017b. PointNet++: Deep hierarchical feature learning on point sets in a metric space. arXiv: Computer Vision and Pattern Recognition.

Saporta, A., Douillard, A., Vu, T., Pérez, P., Cord, M., 2022. Multi-head distillation for continual unsupervised domain adaptation in semantic segmentation. CoRR abs/2204.11667.

Soviany, P., Ionescu, R.T., Rota, P., Sebe, N., 2021. Curriculum self-paced learning for cross-domain object detection. Comput. Vis. Image Underst. 204, 103166.

Su, H., Maji, S., Kalogerakis, E., Learned-Miller, E.G., 2015. Multi-view convolutional neural networks for 3D shape recognition. In: 2015 IEEE International Conference on Computer Vision. ICCV 2015, Santiago, Chile, December 7-13, 2015, pp. 945–953.

Wang, Q., Du, P., Liu, X., Yang, J., Wang, G., 2020. Adversarial unsupervised domain adaptation for cross scenario waveform recognition. Signal Process..

Wang, J., Feng, W., Chen, Y., Yu, H., Huang, M., Yu, P.S., 2018. Visual domain adaptation with manifold embedded distribution alignment. In: Boll, S., Lee, K.M., Luo, J., Zhu, W., Byun, H., Chen, C.W., Lienhart, R., Mei, T. (Eds.), 2018 ACM Multimedia Conference on Multimedia Conference. MM, pp. 402–410.

Wang, S., Zhang, L., Chen, W., Wang, F., Li, H., 2022. Refining pseudo labels for unsupervised domain adaptive re-identification. Knowl.-Based Syst. 242, 108336.

Watanabe, S., Takahashi, S., Wang, L., 2021. Aggregating viewpoints for effective view-based 3D model retrieval. In: 2021 25th International Conference Information Visualisation. IV.

Wen, J., Yuan, J., Zheng, Q., Liu, R., Gong, Z., Zheng, N., 2022. Hierarchical domain adaptation with local feature patterns. Pattern Recognit. 124, 108445.

Wong, H., Ma, B., Yu, Z., Yeung, P.F., Ip, H.H., 2007. 3-D head model retrieval using a single face view query. IEEE Trans. Multim. 9 (5), 1026–1036.

Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J., 2015. 3D ShapeNets: A deep representation for volumetric shapes. In: IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2015, Boston, MA, USA, June 7-12, 2015, pp. 1912–1920.

Xie, S., Zheng, Z., Chen, L., Chen, C., 2018. Learning semantic representations for unsupervised domain adaptation. In: Dy, J.G., Krause, A. (Eds.), Proceedings of the 35th International Conference on Machine Learning. ICML, In: Proceedings of Machine Learning Research, vol. 80, pp. 5419–5428.

Xu, G., Zhang, Z., Zhang, T., Yu, S., Meng, Y., Chen, S., 2022. Aspect-level sentiment classification based on attention-BiLSTM model and transfer learning. Knowl.-Based Syst. 245, 108586.

Yu, E., Ma, J., Sun, J., Chang, X., Zhang, H., Hauptmann, A.G., 2022. Deep discrete cross-modal hashing with multiple supervision. Neurocomputing 486, 215–224.

Yu, E., Sun, J., Li, J., Chang, X., Han, X., Hauptmann, A.G., 2019. Adaptive semi-supervised feature selection for cross-modal retrieval. IEEE Trans. Multim. 21 (5), 1276–1288.

Yue, X., Zhang, Y., Zhao, S., Sangiovanni-Vincentelli, A.L., Keutzer, K., Gong, B., 2019. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. CoRR abs/1909.00889.

Zhang, J., Li, W., Ogunbona, P., 2017. Joint geometrical and statistical alignment for visual domain adaptation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pp. 5150–5158.

Zhou, H., Liu, A., Nie, W., 2019. Dual-level embedding alignment network for 2D image-based 3D object retrieval. In: Amsaleg, L., Huet, B., Larson, M.A., Gravier, G., Hung, H., Ngo, C., Ooi, W.T. (Eds.), Proceedings of the 27th ACM International Conference on Multimedia. MM, pp. 1667–1675.

Zhou, H., Nie, W., Li, W., Song, D., Liu, A., 2020. Hierarchical instance feature alignment for 2D image-based 3D shape retrieval. In: Bessiere, C. (Ed.), Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. IJCAI, pp. 839–845.