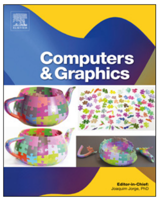
Computers & Graphics 115 (2023) 25–34

Contents lists available at ScienceDirect

Computers & Graphics

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



Technical Section

Domain-specific modeling and semantic alignment for image-based 3D model retrieval☆

Dan Song [a], Xue-Jing Jiang [a], Yue Zhang [b,c], Fang-Lue Zhang [d], Yao Jin [e], Yun Zhang [f,*]

[a] School of Electrical and Information Engineering, Tianjin University, Tianjin, China
[b] School of Biomedical Engineering, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, China
[c] Center for Medical Imaging, Robotics, Analytic Computing & Learning (MIRACLE), Suzhou Institute for Advanced Research, University of Science and Technology of China, Suzhou, China
[d] School of Engineering and Computer Science, Victoria University of Wellington, New Zealand
[e] School of Computer Science and Technology (School of Artificial Intelligence), Zhejiang Sci-Tech University, Hangzhou, China
[f] College of Media Engineering, Communication University of Zhejiang, Hangzhou, China

ARTICLE INFO

Article history:
Received 3 February 2023
Received in revised form 25 June 2023
Accepted 26 June 2023
Available online 28 June 2023

Keywords:
3D model retrieval
Domain adaptation
Semantic alignment

ABSTRACT

Image-based 3D model retrieval aims to search for 3D models according to 2D image queries, which provides a convenient way for the management of large 3D model datasets. Most of the related works put the emphasis on bridging the modality gap between 2D images and 3D models, which faces a lot of challenges due to the huge domain discrepancy. In this paper, we explicitly model and eliminate the domain-specific features of 2D images and 3D models. To alleviate the negative effect of complex background of natural images, we adopt semantic focus loss to constrain networks to learn the most semantically relevant feature representations for both 2D images and 3D models. We conduct extensive experiments on two cross-domain 3D model retrieval datasets, MI3DOR and MI3DOR-2, to show the effectiveness of the proposed method.

1. Introduction

In recent years, with the increasing maturity of 3D modeling technology, 3D models have been widely used in industrial production, e-commerce, biomedicine, and other relevant fields [1–4]. How to effectively organize and utilize the enormous amount of 3D model data becomes an important and challenging problem. 3D model retrieval is an important tool to leverage such massive data. Take the field of biomedicine for example, 3D model retrieval can be used to search for and retrieve 3D models of molecules that can be used in drug discovery. A pharmaceutical researcher could use 3D model retrieval to search for and retrieve 3D models of proteins that are involved in a particular disease to aid in the development of new drugs. According to different query conditions, 3D model retrieval can be categorized into model-based and image-based methods. Compared with 3D models, 2D images are easier to acquire, such as natural images and hand-drawn sketches. In this paper, we focus on the task of retrieving 3D models queried from natural images, which has practical real-world applications.

Traditional 3D model retrieval methods [5,6] mainly rely on manually designed descriptors to represent the model, among which statistical information based methods describe features through statistical information such as angles, normal vectors, and curvature between random points on the model surface. For example, Hamza et al. [5] used a probabilistic shape descriptor to represent an object, which measures the global geodesic distance between two arbitrary points on the object surface. Image-based 3D model retrieval methods aim to query 3D models based on 2D images, and a common way is to convert complex 3D models into a sense of 2D model views. Some traditional methods [7,8] use polar coordinates to represent 2D images and use the bag-of-words method to fuse multi-view features. In recent years, the great success of deep learning in the feature extraction of 2D images has promoted the development of feature extraction for image-based 3D model retrieval. The image-based deep learning method is more mature, and it is the mainstream multi-view feature learning method for 3D model retrieval.

Among deep learning-based methods for 3D model retrieval from 2D images, some previous works [9–13] used supervised learning to solve the problem of matching 3D models according to 2D images. Li et al. [9] fused the image and model into a joint embedding space and calculated their similarity by the distance between the points in the space. Lin et al. [10] performed instance-level and category-level contrastive learning on the images and the models to solve the problem of single image 3D shape retrieval, and achieved a good performance. However, the

☆ This article was recommended for publication by S. Biasotti.
* Corresponding author.
E-mail address: zhangyun@cuz.edu.cn (Y. Zhang).

https://doi.org/10.1016/j.cag.2023.06.033
0097-8493/© 2023 Elsevier Ltd. All rights reserved.

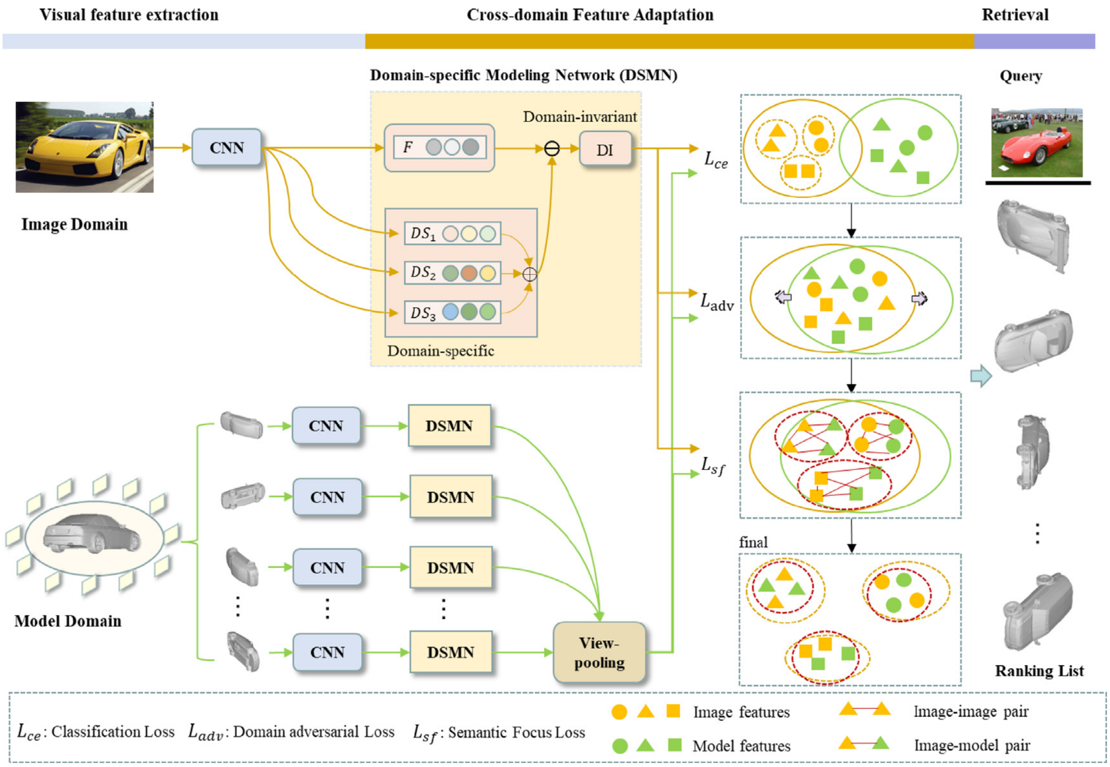


Fig. 1. Overview. The proposed framework mainly consists of three parts. The visual feature extraction module is responsible for extracting image features and multi-view features of the model. In the cross-domain feature adaptation part, we use the DSMN to model domain-specific attributes to obtain domain-invariant representation. And the semantic focus loss is designed, which constructs image-image pairs and image-model pairs to alleviate the interference of negative semantics in natural images in an adversarial manner. The cross-domain adaptation process is constrained by L_{ce}, L_{adv} and L_{sf}, where the feature visualization of loss is shown as four rectangles. The retrieval module uses adaptive features to measure similarity and sort 3D models.

performance highly depends on a large amount of manually labeled data, which needs a time-consuming and expensive process for the explosively emerging 3D models. Therefore, one goal for the task of image-based 3D model retrieval is utilizing existing 2D image datasets with rich labels to boost the feature learning of unlabeled 3D models.

Unsupervised 3D model retrieval methods can be roughly divided into two types: metric learning methods and domain adversarial learning methods. Metric learning methods [14–19] minimize the statistical metric of two domains to align the feature distributions. Adversarial learning-based methods [20–23] simultaneously train a feature extractor and a domain discriminator to force the feature distributions closer, or adds class-level or instance-level constraints on this basis. Despite that great efforts have been made to enable accurate and convenient 3D model retrieval using 2D images, the following issues have not yet been fully resolved:

- Difficulty in eliminating the interference of domain-specific attribute that is harmful to the alignment of two domain features. Due to the obvious discrepancy between 2D images and 3D models, as well as the lack of labels for 3D models, the alignment between the 2D and 3D features is a nontrivial problem. Most of the previous methods [14,23,24] directly operate on the features of 2D images and 3D models to narrow the distance between the two domains. Nevertheless, this may ignore the effect of domain-specific properties that are harmful to cross-domain alignment. Therefore, it becomes necessary to find a suitable representation for both 2D images and 3D models, so that we can obtain domain-invariant features for a better feature alignment and 3D model retrieval.
- Difficulty in alleviating the interference caused by the possible complex background in 2D images. One of the main differences between 2D images and 3D models is that 2D images could contain complex background while a 3D data item only contains the target object. Most of the existing mainstream methods [14,15,25] focus on the features of the entire image, and less consideration is given to the processing of complex backgrounds, causing that the irrelevant semantic information in the backgrounds are also encoded. It will distract the network from learning 2D image features corresponding to 3D models. Therefore, we consider that the network should focus on only the target object in a 2D image for the retrieval task.

To overcome the above problems, we model the domain-specific features of natural images and projections of 3D models and remove them to narrow the gap between the two domains. Furthermore, the semantic focus loss is used to alleviate the influence of the background in natural images by constructing sample pairs, so that the network model can pay more attention to the target object. Specifically, the framework consists of three parts as shown in Fig. 1: visual feature extraction, cross-domain feature adaptation, and retrieval. Firstly, we represent 3D models by rendering from multiple viewpoints, and use convolutional neural networks to extract visual features for both natural images and 3D models. Later in the adaptation module, we explicitly model domain-specific features and eliminate these features to acquire domain-invariant features. The classification loss L_{ce} supervises the network to learn semantic information with 2D image labels. The semantic focus loss L_{sf} constrains the network to focus on the semantic information of the target object via adversarial learning. The domain adversarial loss L_{adv} aims to align the overall feature distribution of the images and the models. In the retrieval phase, given the features of image queries, we obtain a rank list of 3D models in the dataset. In summary, the main contributions of this paper are:

- We model and eliminate the domain-specific features in both 2D natural images and 3D model projections, which can effectively narrow the discrepancy between 2D and 3D domains to improve the performance of image-based 3D model retrieval.
- We adopt a semantic focus loss to alleviate the interference of irrelevant semantic information in 2D images on the retrieval task, which can make the network concentrate on the target object to obtain a better retrieval performance.
- The experimental results on two public cross-domain 3D model datasets, i.e., MI3DOR and MI3DOR-2, show the effectiveness of the proposed method.

2. Related work

2.1. 3D model retrieval

3D model retrieval is a process of searching similar models in a large gallery according to the query and arranging the models in accordance with the similarity measurement. According to different types of queries, we can divide 3D model retrieval methods into two categories, one is image-based methods and the other is model-based methods. For model-based 3D model retrieval methods [26–28], both the query and the search target are 3D models, and various similarity measurements are explored. The representation forms are usually point cloud [26,29], mesh [30], voxel [31] etc. Wu et al. [32] proposed to use Convolutional Deep Belief Network (CDBN) to represent 3D models, which transfer the model into a distribution of binary variables on the 3D grid that can be activated by view planning for object recognition. Li et al. [33] propose the network structure of the Self-Organizing Network for feature extraction of a disordered point cloud, and the self-organization map is constructed to simulate the point cloud spatial distribution. Based on self-organizing map, layered feature extraction is carried out for single point and self-organizing map node, and a feature vector is used to represent the input point cloud.

For image-based 3D model retrieval methods, the query is a 2D image and the search target is the 3D model [34–36], which facilitates practical scenarios. Because of the differences in data distribution between 2D images and 3D models, most image-based 3D model retrieval methods use multiple view images to represent 3D models. Among the traditional methods, Khotanzad et al. [7] used polar coordinates to represent the 2D image, which mapped all the pixels in the image to the unit circle, and converted the Cartesian coordinates to polar coordinates, which have the property of rotation invariance. Ohbuchi et al. [8] first proposed to apply the Bag-of-Words model to 3D model retrieval. It characterizes single-view features through a scale-invariant feature transformation descriptor [37], and then uses a bag-of-words model to fuse features from multiple views. In the deep learning method in recent years, Su et al. [38] designed a Multi-View Convolutional Neural Network, which extracts features of the rendered images from different perspectives of the 3D model via convolution network and fuse view features to a compact descriptor. Massa et al. [39] proposed an end-to-end 2D-3D paradigm detection method. It adopts a cross-domain adaptive method to adjust the characteristics of the 2D image to better align with the rendered view of the CAD model. Its adaptive method is integrated into a convolutional neural network-based pipeline to improve the accuracy of 2D-3D detection. Zhou et al. [23] design an unsupervised two-layer embedded alignment network to reduce the statistical difference between the images and the models for domain alignment, and narrow the distance between the class centers of the 2D image and 3D model domains for class-level alignment. Nie et al. [40] proposed the multi-channel-attention (MCA) convolutional neural network method to represent 3D models. The MCA method can effectively fuse multiple 2D panoramas of 3D models by finding ways of different weights of each panorama view.

2.2. Domain adaptation

In transfer learning, domain adaptation techniques are popular to deal with the situation that the data distributions in the source and target domains are different. Domain adaptation can be roughly divided into metric learning methods [41] and adversarial based methods [22,42]. Sun et al. [43] achieve the adaptation goal by balancing the classification loss and narrowing the second-order statistics of the two domains. Most metric learning methods adopt Maximum Mean Discrepancy [44] to assess the differences between the weighted sum of all order statistical moments of the two domains. Chen et al. [41] designed the Higher-order Moment Matching (HoMM) framework to minimize domain differences, and HoMM is further extended to reproducing kernel Hilbert Spaces for alignment.

The adversarial learning method is to judge whether the feature is from the source domain or the target domain by training the domain discriminator, while the feature extractor learns a domain-invariant representation to confuse the domain discriminator. Ganin et al. [24] first propose adversarial domain adaptation network. On this basis, Long et al. [45] propose Conditional Domain Adversarial Networks (CDANs) that employ the new conditioning methods: multilinear conditioning and entropy conditioning. The former improves the recognition rate of the classifier by capturing the cross-variance between the feature representation and the classifier prediction, and the latter guarantees the portability of the classifier by controlling the uncertainty of the classifier prediction. Liu et al. [46] designed the Transferable Adversarial Training framework to achieve the adaptation of deep classifiers. By tricking the class classifier and domain discriminator, the method generates transferable examples to bridge the gap across domains. Jiang et al. [22] presented simulated recognition as an adversarial reinforcement learning problem, using the learned GAN loss instead of the standard mean squared error to measure the difference between the distributions of transition tuples, which addresses transferring the policy to a new domain with different dynamics.

3. Method

3.1. Overview

In this paper, we aim to match relevant unlabeled 3D models based on a given 2D image as a query. We represent the 2D image domain as $I = \left(x_i^I, y_i^I\right)_{i=1}^{N_I}$, where x_i^I denotes an image, $y_i^I \in \{1, 2, \ldots, C\}$ is the corresponding label, C is the number of categories, and N_I is the number of images. The unlabeled 3D model domain is defined as $M = \left(x_j^M\right)_{j=1}^{N_M}$with N_M samples. 2D images and 3D models share the same label space, but domain discrepancies exist between their feature distributions.

The framework is shown in Fig. 1 which contains three key procedures. The module of visual feature extraction is responsible for taking 2D images and multiple model views into the backbone network to extract features. The cross-domain feature adaptation module is in charge of aligning the cross-domain feature distributions and concentrating on the main semantics. Specifically, the features extracted from the 2D images and 3D model views through CNN are input into the Domain-specific Modeling Network (DSMN) to obtain domain-invariant features. For the 3D model data, after passing through the DSMN, the max-pooling is used to fuse the multiple features into a compact feature descriptor. Then both the image and model features are passed through the classifier. The cross-domain adaptation process is constrained by the classification loss L_{ce}, the domain adversarial loss L_{adv}, and the semantic focus loss L_{sf}. As shown in the visualization of the features in the four rectangles in Fig. 1, L_{ce} uses label information to train network to aggregate the same category features in the image domain, L_{adv} aligns the overall feature distributions of images and models, L_{sf} narrows the distance of the same category in two domains by constructing the image-image pair and the image-model pair. The network is trained under the constraints of the three losses to reduce the distance between features of the same category and expand the decision boundary of different categories. Finally, the retrieval module is responsible for sorting the adapted features. Specifically, the trained network is used to extract features from 2D images and 3D models, and then the euclidean distance between the query image feature representation and each 3D model feature representation in the 3D model database is calculated. Finally, the 3D models are sorted based on the similarity value.

3.2. Visual feature extraction

We use ResNet-50 [47] as the backbone CNN to extract features from 2D images and 3D model multi-views. As MVCNN [38], we represent a 3D model with a set of rendered images captured from different viewpoints around the object by the Phong reflection model [48], one set of rendered images includes 12 views. Specifically, 12 virtual cameras are set around the 3D model, all of which are at an angle of 30 degrees to the horizontal plane, and the interval between two viewing angles is 30 degrees so that 12 views of the 3D model from different viewing angles can be obtained. The model views are fed into CNNs, which share weights with the CNNs that extracts image features.

3.3. Cross-domain feature adaptation

In our cross-domain feature adaptation, we aim to overcome two main problems that exist in the task of image-based 3D model retrieval. On the one hand, domain-specific attributes in 2D images and 3D models negatively impact cross-domain retrieval performance, so we explicitly model the specific features of 2D images and 3D models. On the other hand, 2D images may contain complex background, but 3D model projections only include the target object. Thus, we utilize semantic focus loss to alleviate irrelevant information induced by the background of 2D images. The domain-specific modeling network will be introduced in Section 3.3.1, and the semantic focus loss will be explained in Section 3.3.2.

3.3.1. Domain-specific modeling network

The Domain-specific modeling network consists of a fundamental network (denoted as F) and a domain-specific network (denoted as DS) [49]. The F network extracts all the features of an image, which consists of two parts: domain-specific features and domain-invariant features. Consequently, the domain-invariant representation can be obtained by subtracting the domain-specific network's output from the fundamental network's output.

Formally, we extract the feature of image x_i through CNN and denote it as $G(x_i)$ and feed $G(x_i)$ into the fundamental network and domain-specific network respectively, which are denoted as $F(G(x_i))$ and $DS(G(x_i))$. We use DI to represent domain-invariant network. Thus the relationship can be expressed as:

$$DI(G(x_i)) = F(G(x_i)) - DS(G(x_i)) \tag{1}$$

$DS(G(x_i))$ tends to be dominant by the representation of domain-specific attributes, so the domain-invariant representation can be obtained by subtracting $DS(G(x_i))$.

Specifically, to capture more domain-specific properties to enhance the ability of DS, we adopt the method of integrating multiple sub-networks, and the initialization of these sub-networks is different, which is to enable each sub-network to model local domain-specific properties, resulting in a stronger fit for the entire domain-specific network. We explore the role of different initializations by analyzing the similar relationship between $DI(.)$ and different $DS(.)$, which is also verified by experiments in Section 4.4.4. On the other hand, too many sub-networks can lead to a greater risk of overfitting. The number of sub-networks is shown in Table 4. Let $X = G(x_i)$, then we have $DS(X) = DS_1(X) + DS_2(X) + DS_3(X)$.

3.3.2. Semantic focus loss

The network is trained to capture semantic information with the help of labeled images. As Eq. (2) shows, L_{ce} represents the classification loss on labeled images, where J denotes the category cross-entropy loss function.

$$L_{ce} = \mathbb{E}_{(x,y)\sim D_I}\left[J\left(DI(G(x)), y\right)\right] \tag{2}$$

As natural images have complex background while model views only contain the target object, we use semantic focus loss to constrain the network concentrate on the dominant semantic representation. Suppose the semantic prediction (i.e., the output of classifier) for image x_i is $p_i^I \in \mathbb{R}^d$ where d is the number of category, and the semantic prediction for model x_j is p_j^M. The largest probability within p_i^I indicates the dominant semantic representation, which is probably caused by the main object in the image. For the rest probabilities, different images have different distributions, which are mainly because of various backgrounds. The goal of semantic focus loss is adopting a two-player adversarial strategy [50] to make the feature extraction suppress irrelevant semantic information.

Firstly, two types of pairs are constructed, which are image-image pair and image-model pair with the same dominant prediction. Then the semantic focus loss is defined as:

$$\begin{aligned} L_{sf} = \min_G \max_C &- \frac{T^2}{N_1} \sum_{y_i^I = y_k^I} JS\left(\boldsymbol{p}_i^I, \boldsymbol{p}_k^I\right) \\ &- \frac{T^2}{N_2} \sum_{y_i^I = \hat{y}_j^M} JS\left(\boldsymbol{p}_i^I, \boldsymbol{p}_j^M\right) \end{aligned} \tag{3}$$

The JS divergence aims to measure the discrepancy between the predictions of a pair, G refers to feature extractor and C denotes classifier, and T is the temperature scaling parameter. N_1 and N_2 represent the number of samples that satisfy the conditions of $y_i^I = y_k^I$ and $y_i^I = \hat{y}_j^M$, respectively. For 2D images, we have the ground truth label for each sample. For 3D models, we adopt the pseudo-label which is $\hat{y}_j^M = \mathrm{argmax}_c \left(q_j^{M(c)}\right)$ where $(q_j^{M(c)})$ is the cth element of the softmax output.

As the above two-player formulation shows, the training of the classifier tries to increase the discrepancy between samples with the same dominant prediction. However, for the pair of samples, the prediction probability towards the dominant category performs similarly (i.e., both samples have the largest probability

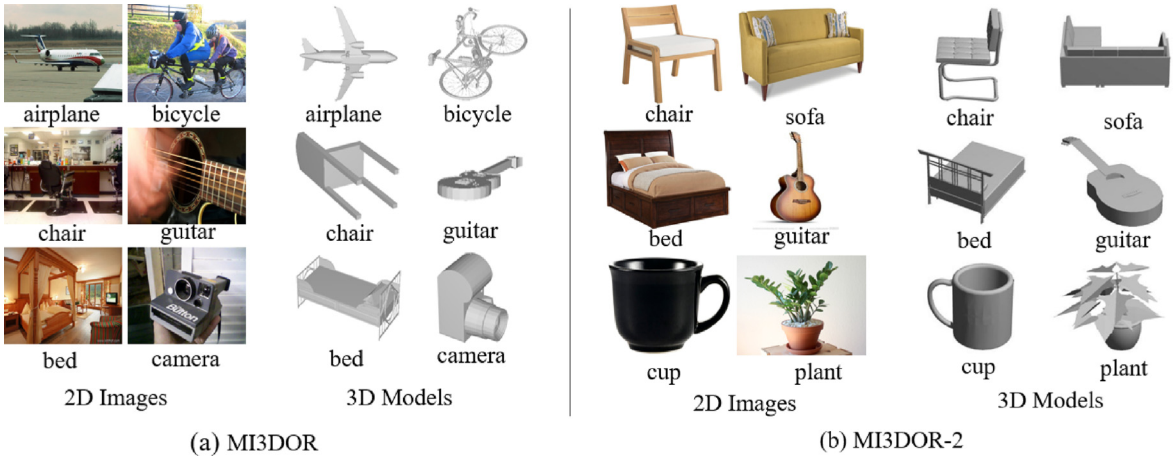


Fig. 2. Samples of MI3DOR and MI3DOR-2.

at the same category). Due to different backgrounds, the probabilities at the rest categories have various distributions for the paired samples. Consequently, maximizing the discrepancy will make the classifier increase the weight of irrelevant category predictions.

In contrast, the training of the feature extractor will suppress the features of these irrelevant semantics in order to reduce the discrepancy of prediction distribution, which makes the feature extractor focus on the dominant semantic information and alleviates the negative effects caused by complex image backgrounds. In the process of realizing adversarial loss, gradient reverse layer (GRL) [24] is adopted to realize parameter optimization through gradient descent.

3.3.3. Overall adaptation loss

The commonly used domain adversarial loss (Eq. (4)) is also adopted to align the feature distributions of images and models, which benefits the knowledge transfer from labeled images to unlabeled models.

$$\begin{aligned} L_{adv} = & -\mathbb{E}_{x\sim D_I} \log(1 - D(DI(G(x)))) \\ & -\mathbb{E}_{x\sim D_M} \log(D(DI(G(x)))) \end{aligned} \tag{4}$$

The above domain adversarial loss aims to align the overall feature distributions of images and models, where D_I and D_M represent the feature distributions of image domain and model domain respectively. $DI(G(x))$ is the output of the domain-specific modeling network. i.e., the domain-invariant feature. D represents the domain discriminator.

The total adaptation loss can be written as follows:

$$L_{\text{total}} = L_{ce} + \beta L_{adv} + \gamma L_{sf} \tag{5}$$

where β and γ are the trade-off parameters.

4. Experiments

4.1. Datasets

We conduct experiments on two cross-domain 3D model retrieval datasets, i.e., MI3DOR and MI3DOR-2. MI3DOR [51] is the first monocular image-based 3D model retrieval benchmark used in the 2019 SHREC competition, in which 2D images are selected from ImageNet [52], and 3D models are selected from ModelNet40 [53], ShapeNetCore55 [54], NTU [55], and PSB [56], containing 21,000 2D images and 7690 3D models, with a total of 21 categories. Among them, 10,500 2D images and 3842 3D models are used as the training set, and the rest of the images and models are used as the test set. Some samples of the MI3DOR dataset are shown in Fig. 2(a).

The MI3DOR-2 [23] dataset consists of 19,694 2D images and 3982 3D models, divided into 40 categories. The 2D images are selected from Google, and the 3D models are selected from the ModelNet40 [53] dataset. 400 2D images and 800 3D models are selected as the test set, and the rest are used as the training set. Some samples of the MI3DOR-2 dataset are shown in Fig. 2(b). The MI3DOR-2 dataset has more categories of 2D images and 3D models than the MI3DOR dataset. Compared with MI3DOR-2, the background of the 2D images in the MI3DOR dataset is more complex.

4.2. Evaluation criteria

In order to evaluate the retrieval performance, we used six evaluation criteria: the Nearest Neighbor (NN), First Tier (FT), Second Tier (ST), F measure (F), Discounted Cumulative Gain (DCG) and Average Normalized Modified Retrieval Rank (ANMRR) as [57]. The values of these evaluation criteria are in the range of 0 to 1. The smaller the value of the evaluation criterion ANMRR is, the better retrieval performance of the model is. The larger the other five evaluation criteria are, the better the model performance is.

4.3. Implementation details

In terms of the implementation details in the paper, we employ Resnet-50 [47] as the backbone for extracting features. We keep the first four layers of backbone, and then add Domain-specific Modeling Network (DSMN) to obtain domain invariant representation. The fundamental network consists of one linear layer, and the domain-specific network consists of three linear layers with different initializations, of which the outputs are added up at the end. Afterward, the obtained domain-invariant features are fed into a linear layer for dimensionality reduction. Finally, we use a two-layer MLP to implement the classifier.

We implemented our network on PyTorch with one NVIDIA GTX 1080Ti (12G) and Intel(R) Xeon(R) CPU (64G). In our experiments, the hyperparameter are set as follows: The trade-off parameter β in L_{total} is set to 0.5. As the training progresses, γ increases from 0 to 1. The parameter T in the L_{sf} loss function is set to T = 10. The network parameters trained on ImageNet [52] are used for initialization. The initial learning rate of the stochastic gradient descent optimizer is set as 0.01, and the momentum is 0.9 and the weight decay is 0.0005. The batch size for training is set to 64, and the training time is about 6 h.

Table 1
Performance on MI3DOR.

	NN	FT	ST	F	DGG	ANMRR
MEDA [16]	0.430	0.344	0.501	0.046	0.361	0.646
JAN [17]	0.446	0.344	0.495	0.085	0.364	0.647
RevGard [24]	0.650	0.505	0.643	0.112	0.542	0.474
DLEA [23]	0.764	0.558	0.716	0.143	0.597	0.421
SC-IFA [15]	0.721	0.584	0.721	**0.163**	0.637	0.363
HIFA [14]	0.778	0.618	0.768	0.151	0.654	0.362
SADA [58]	**0.783**	**0.638**	**0.793**	0.154	**0.672**	**0.343**
Ours	0.7749	0.6303	0.7764	0.1509	0.6657	0.3513

Table 2
Performance on MI3DOR-2.

	NN	FT	ST	F	DGG	ANMRR
MEDA [16]	0.570	0.392	0.523	0.392	0.425	0.590
JAN [17]	0.608	0.501	0.646	0.501	0.527	0.484
RevGard [24]	0.623	0.467	0.614	0.467	0.503	0.514
DLEA [23]	0.700	0.555	0.681	0.555	0.593	0.424
SC-IFA [15]	0.713	0.641	0.738	0.623	0.648	0.415
HIFA [14]	0.725	0.570	0.710	0.570	0.598	0.413
SADA [58]	0.738	0.615	0.746	0.615	0.651	0.366
Ours	**0.7875**	**0.6714**	**0.7797**	**0.6714**	**0.7114**	**0.3077**

4.4. Results and analysis

4.4.1. Comparison with existing methods

The proposed method is compared with previous methods on the MI3DOR and MI3DOR-2 datasets. According to different domain adaptation strategies, we divide these methods into two categories: (1) metric learning-based methods; (2) adversarial learning-based adaptation methods.

- The metric learning-based methods usually reduce the generation error of the 3D model domain by reducing the statistical difference between the two domains. MEDA [16] uses manifold feature learning to dynamically learn the importance of marginal distribution alignment and conditional distribution alignment. JAN [17] utilizes a joint maximum mean error metric to constrain the joint distribution of features in the two domains. They mainly focus on measuring the similarity of features and narrow the gap between the two domains with metric criteria constraints. Compared with adversarial learning methods, these methods are more traditional and have slightly weaker performance.
- Adversarial learning-based adaptation methods mainly align the 2D image domain and the 3D model domain through the confrontation between the feature extractor and the domain discriminator. Related methods are RevGard [24], DLEA [23], SC-IFA [15], HIFA [14], and SADA [58]. The RevGard method was originally proposed to add a domain discriminator after the feature extraction module to align 2D images and 3D models at the domain level. On this basis, some methods use class-level constraints or instance-level constraints to further reduce the data distribution differences between images and 3D models. SADA adopts the method of self-supervised auxiliary domain alignment and divides multiple projections of the 3D model into two sub-target-domains according to the similarity among the projections. The image and model domains are then combined to build an intermediate domain to ease the direct alignment of the image domain and the model domain. Our method is based on adversarial learning like them, but they pay more attention to the entire feature of the image and the feature is processed to align the two domains, but not all features are conducive to domain alignment, our method models domain-specific properties and eliminates them. On the other hand, paying attention to all the features of the image will also make the complex background in the natural image also mixed, so our method also designs the semantic focus loss contrast to constrain.

The retrieval results on two datasets are shown in Tables 1 and 2 respectively. It can be observed that our method can achieve better performance in most evaluation metrics.

We have the following observations on the performance of the proposed method: (1) The adversarial training-based method is better than the traditional metric learning-based method, which benefits from the joint optimization of feature learning and cross-domain adaptation. (2) Compared with adversarial learning-based methods, our method also shows some advantages. DLEA, SC-IFA, and HIFA methods mostly add category constraints or instance constraints to both the image and model domains on the basis of adversarial learning to reduce the domain distance. Our method alleviates the difference in data distribution between the two domains by modeling and eliminating domain-specific attributes, and eliminates the impact of negative semantics in 2D images through semantic focus loss. Compared with SADA, we achieve comparable performance on MI3DOR and superior results on MI3DOR-2, and the slight weakness on MI3DOR mainly attributes to that we do not consider the relationship between multiple views of the 3D model. (3) It can also be observed that our method brings more performance gains on MI3DOR-2 than those on MI3DOR. The reason is that MI3DOR-2 has more object categories than the dataset MI3DOR, our method constructs intra-domain and inter-domain sample pairs when constraining the semantic concentration. Therefore, the more types of samples, the more conducive to the training of sample pairs and the better performance of the model can be achieved.

4.4.2. Ablation study

In this section, we validate the contribution of introducing domain-specific modeling and semantic focus in the task of image-based 3D model retrieval. The ablation results on MI3DOR is shown in Table 3. "$L_{ce} + L_{adv}$" represents the retrieval results obtained under the condition of classification loss and domain adversarial loss, which is also the baseline of this paper. "$L_{ce} + L_{adv}(+DSMN)$" represents the retrieval results obtained by adding domain-specific modeling network to the baseline, and the line of "$L_{ce}+L_{adv}+L_{sf}$" represents the retrieval results obtained by adding semantic focus loss in the presence of a classification loss and a domain adversarial loss. "$L_{ce}+L_{adv}+L_{sf}(+DSMN)$" represents the retrieval result obtained by adding both the domain-specific modeling and semantic focus loss to the baseline, i.e., the proposed method.

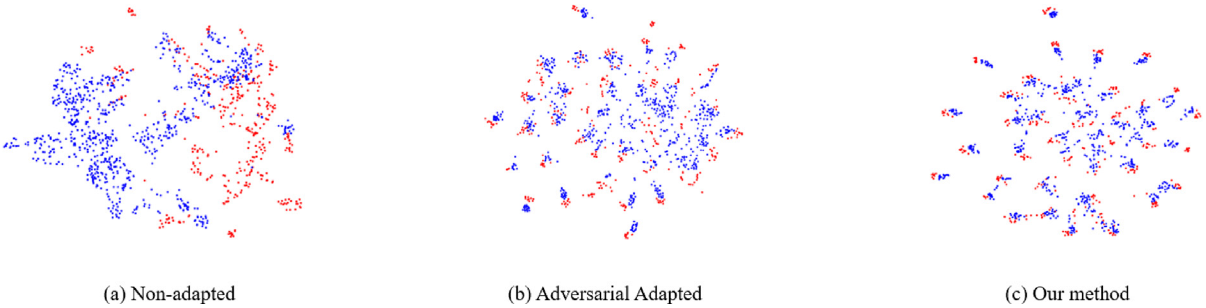
From the ablation results, it can be found that the domain-specific modeling brings the relative improvements of 6.4%, 11.1%, 7.9%, 11.1%, 11.0% and 16.1% towards the baseline in respect of NN, FT, ST, F, DGG and ANMRR. It shows the effectiveness of domain invariant representation by constructing domain-specific networks for the task of image-based 3D model retrieval. Comparatively, semantic focus loss relatively improves the baseline by 0.29%, 12.3%, 8.5%, 8.2%, 11.2% and 17.6% in terms of NN, FT, ST, F, DGG and ANMRR. The proposed method (i.e., introducing both the domain-specific modeling and semantic focus loss to the baseline) performs best, which validates the necessity of each module.

4.4.3. Qualitative evaluation

As shown in Fig. 3, we visualized the feature distributions of images and models via T-SNE [59]. Red dots represent the image features and blue dots are model features. Fig. 3(a) shows the feature distributions with no adaptation, Fig. 3(b) shows the

Table 3
Ablation results.

	NN	FT	ST	F	DGG	ANMRR
$L_{ce} + L_{adv}$	0.7217	0.5373	0.6877	0.1319	0.5717	0.4462
$L_{ce} + L_{adv}(+DSMN)$	0.7678	0.5972	0.7419	0.1466	0.6344	0.3842
$L_{ce} + L_{adv} + L_{sf}$	0.7238	0.6032	0.7459	0.1427	0.6356	0.3795
$L_{ce} + L_{adv} + L_{sf}(+DSMN)$	**0.7749**	**0.6303**	**0.7764**	**0.1509**	**0.6657**	**0.3513**


(a) Non-adapted

(b) Adversarial Adapted

(c) Our method

Fig. 3. Visualization for feature distributions. Red dots represent the image features and blue dots represent model features.

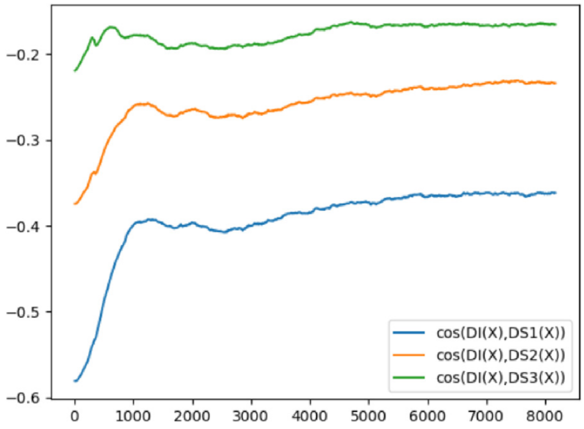


Fig. 4. Cosine similarity of $DI(X)$ and $DS(X)$.

Table 4
Sensitivity analysis of the number of sub-networks on MI3DOR.

The number of *DS*	NN	FT	ST	F	DGG	ANMRR
1	0.7476	0.5574	0.6946	0.1408	0.5976	0.4229
2	0.7580	0.5690	0.7174	0.1421	0.6065	0.4125
3	**0.7749**	**0.6303**	**0.7764**	**0.1509**	**0.6657**	**0.3513**
4	0.7630	0.5896	0.7311	0.1452	0.6274	0.3921
5	0.7472	0.5937	0.7441	0.1429	0.6274	0.3894

Table 5
Sensitivity analysis of the number of sub-networks on MI3DOR-2.

The number of *DS*	NN	FT	ST	F	DGG	ANMRR
1	0.7525	0.6597	0.7677	0.6597	0.6941	0.3217
2	0.7625	0.6634	0.7789	0.6634	0.6954	0.3182
3	**0.7875**	**0.6714**	**0.7797**	**0.6714**	**0.7114**	**0.3077**
4	0.7750	0.6386	0.7554	0.6386	0.6763	0.3405
5	0.7425	0.6104	0.7410	0.6104	0.6583	0.3663

distributions with the participation of adversarial learning (i.e, the baseline of our work), and Fig. 3(c) shows the distributions obtained by the proposed method. On one hand, the results show that there exist a large discrepancy between images and models and it is necessary to align the distributions for the task of image-based 3D model retrieval. On the other hand, compared with the adversarial alignment, the proposed method has a clearer alignment at category level. More discriminative features will boost the retrieval performance.

4.4.4. *Sensitivity study*

In the paper, the domain-specific network (DS) is responsible for modeling domain-specific properties. The domain-invariant representation can be obtained by subtracting the domain-specific network's output from the fundamental network's output. In order to capture more domain-specific attributes to enhance the ability of DS, an approach of integrating multiple sub-networks to the model is adopted. To decide the optimal number of sub-networks, we have supplemented sensitivity experiments as shown in the Tables 4 and 5. When the number of sub-networks increases, the model performance improves, indicating that more sub-networks can enhance the modeling ability of DS. The model achieves the best results when the number is 3. However, when the number of sub-networks exceeds 3, the performance will be degraded due to risks such as too many parameters or overfitting.

Different initializations might have different effects on the performance of the model in terms of capturing different domain-specific properties during training. Setting different initializations for each sub-network makes the similarity between the sub-networks lower during the training process, so that each sub-network can model different local domain-specific properties. We have carried out experiments to verify the above hypothesis and calculated the cosine similarity between the outputs of the three subnetworks (DS_1, DS_2, DS_3) and domain invariant representation (DI) respectively. As shown in Fig. 4, $DI(X)$ and $DS(X)$ are negatively correlated, verifying that the correlation between $DS(X)$ representing domain-specific properties and $DI(X)$ for domain-invariant properties is low. And the similarity distributions of different DS and DI are different, indicating that the domain properties captured by different sub-networks are different.

4.4.5. *Visual retrieval results*

Some retrieval examples are given in Fig. 5, where the left column is the image for query and the right columns are retrieved 3D models ranked at the top 5 positions. Above the dash we show several successful cases while below the dash we show the failure cases. The main reasons for the failure cases mainly lie in the following aspects: (1) shape similarities, e.g., some cameras and radios, and some cups and vases; (2) semantic overlaps, e.g., plant and flower; (3) co-existence of objects, e.g., a desk and a chair locate in the same image.

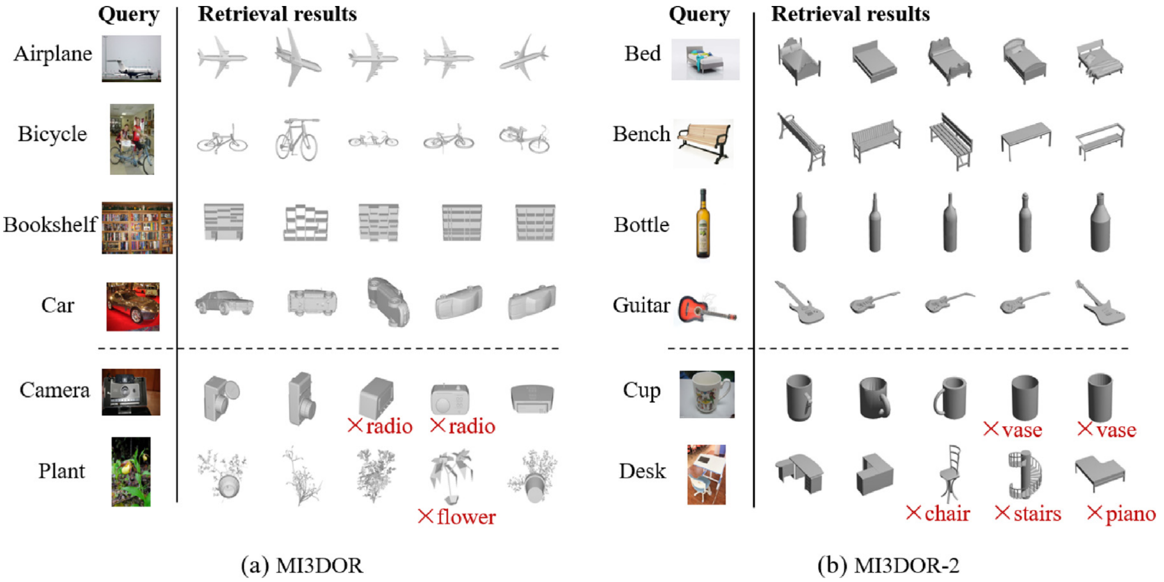


Fig. 5. Visual retrieval results on MI3DOR and MI3DOR-2.

5. Conclusion and future work

In this paper, we try to overcome two problems existing in the image-based 3D model retrieval task. To narrow the modality gap, we model domain-specific attributes for 2D images and multi-view represented 3D models, and align the domain invariant features. To alleviate the negative effects caused by the complex backgrounds of 2D images, we introduce semantic focus loss to make the feature extraction concentrate on the dominant semantics. We have conducted extensive experiments including comparison with existing methods, ablation study, feature distribution visualization and visual retrieval results, and the results show the effectiveness of the proposed method.

Although our method has achieved good results, there is still space for improvements. Our proposed method does not take into account the relationship between multiple views of a 3D model, so future research should consider using graph convolutional neural networks to analyze and calculate the structural information of views from different perspectives and the relationships between them. In addition, our method directly utilizes the output of the classifier as pseudo labels for 3D model samples. However, in the early stages of training, the classifier's classification performance for unlabeled 3D models is not good, resulting in low accuracy of pseudo labels. Therefore, in future work, we can improve the accuracy of image-image pairs and image-model pairs in the semantic focus loss by designing algorithms to improve the accuracy of pseudo-labels, thereby improving the retrieval performance.

CRediT authorship contribution statement

Dan Song: Study conception and design, Material preparation, Data collection and analysis, Writing first draft of the manuscript. **Xue-Jing Jiang:** Study conception and design, Material preparation, Data collection and analysis, Experiments are implemented. **Yue Zhang:** Study conception and design, Material preparation, Data collection and analysis, Experiments are implemented. **Fang-Lue Zhang:** Study conception and design, Revised the draft. **Yao Jin:** Study conception and design, Experiments are implemented. **Yun Zhang:** Study conception and design.

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

Dan Song is supported in part by the National Nature Science Foundation of China (61902277). Yao Jin is supported by the Key R and D Programs of Zhejiang Province (NO. 2023C01224), the Fundamental Research Funds of Zhejiang Sci-Tech University (23232106-Y) and the National Nature Science Foundation of China (61702458). Yue Zhang is supported by Jiangsu Funding Program for Excellent Postdoctoral Talent. Yun Zhang is supported by Zhejiang Province Public Welfare Technology Application Research (No. LGG22F020009) and Key Lab of Film and TV Media Technology of Zhejiang Province (No. 2020E10015). All authors read and approved the final manuscript.

References

[1] Li Z, Xu C, Leng B. Angular triplet-center loss for multi-view 3d shape retrieval. In: Proceedings of the AAAI conference on artificial intelligence, Vol. 33. 2019, p. 8682–9.
[2] Lu X, Zhu L, Cheng Z, Li J, Nie X, Zhang H. Flexible online multi-modal hashing for large-scale multimedia retrieval. In: Proceedings of the 27th ACM international conference on multimedia. 2019, p. 1129–37.
[3] Huang Y-H, He Y, Yuan Y-J, Lai Y-K, Gao L. StylizedNeRF: Consistent 3D Scene Stylization As Stylized NeRF via 2D-3D Mutual Learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. CVPR, 2022, p. 18342–52.
[4] Li X-L, Guo M-H, Mu T-J, Martin RR, Hu S-M. Long range pooling for 3D large-scale scene understanding. 2023, arXiv preprint arXiv:2301.06962.
[5] Hamza AB, Krim H. Geodesic matching of triangulated surfaces. IEEE Trans Image Process 2006;15(8):2249–58.
[6] Cheng H-C, Lo C-H, Chu C-H, Kim YS. Shape similarity measurement for 3D mechanical part using D2 shape distribution and negative feature decomposition. Comput Ind 2011;62(3):269–80.
[7] Khotanzad A, Hong YH. Invariant image recognition by Zernike moments. IEEE Trans Pattern Anal Mach Intell 1990;12(5):489–97.

[8] Ohbuchi R, Osada K, Furuya T, Banno T. Salient local visual features for shape-based 3D model retrieval. In: 2008 IEEE international conference on shape modeling and applications. IEEE; 2008, p. 93–102.

[9] Li Y, Su H, Qi CR, Fish N, Cohen-Or D, Guibas LJ. Joint embeddings of shapes and images via cnn image purification. ACM Trans Graph 2015;34(6):1–12.

[10] Lin M-X, Yang J, Wang H, Lai Y-K, Jia R, Zhao B, Gao L. Single image 3D shape retrieval via cross-modal instance and category contrastive learning. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021, p. 11405–15.

[11] Sun X, Wu J, Zhang X, Zhang Z, Zhang C, Xue T, Tenenbaum JB, Freeman WT. Pix3d: Dataset and methods for single-image 3d shape modeling. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, p. 2974–83.

[12] Aubry M, Russell BC. Understanding deep features with computer-generated imagery. In: Proceedings of the IEEE international conference on computer vision. 2015, p. 2875–83.

[13] Fu H, Li S, Jia R, Gong M, Zhao B, Tao D. Hard example generation by texture synthesis for cross-domain shape similarity learning. Adv Neural Inf Process Syst 2020;33:14675–87.

[14] Zhou H, Nie W, Li W, Song D, Liu A-A. Hierarchical instance feature alignment for 2D image-based 3D shape retrieval. In: Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence. 2021, p. 839–45.

[15] Zhou H, Nie W, Song D, Hu N, Li X, Liu A-A. Semantic consistency guided instance feature alignment for 2D image-based 3D shape retrieval. In: Proceedings of the 28th ACM international conference on multimedia. 2020, p. 925–33.

[16] Wang J, Feng W, Chen Y, Yu H, Huang M, Yu PS. Visual domain adaptation with manifold embedded distribution alignment. In: Proceedings of the 26th ACM international conference on multimedia. 2018, p. 402–10.

[17] Long M, Zhu H, Wang J, Jordan MI. Deep transfer learning with joint adaptation networks. In: International conference on machine learning. PMLR; 2017, p. 2208–17.

[18] Zhang J, Li W, Ogunbona P. Joint geometrical and statistical alignment for visual domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, p. 1859–67.

[19] Wu Z, Zhang Y, Zeng M, Qin F, Wang Y. Joint analysis of shapes and images via deep domain adaptation. Comput Graph 2018;70:140–7.

[20] Xie X, Chen J, Li Y, Shen L, Ma K, Zheng Y. Self-supervised cyclegan for object-preserving image-to-image domain adaptation. In: European conference on computer vision. Springer; 2020, p. 498–513.

[21] Wang J, Jiang J. Conditional coupled generative adversarial networks for zero-shot domain adaptation. In: Proceedings of the IEEE/CVF international conference on computer vision. 2019, p. 3375–84.

[22] Jiang Y, Zhang T, Ho D, Bai Y, Liu CK, Levine S, Tan J. Simgan: Hybrid simulator identification for domain adaptation via adversarial reinforcement learning. In: 2021 IEEE international conference on robotics and automation. ICRA, IEEE; 2021, p. 2884–90.

[23] Zhou H, Liu A-A, Nie W. Dual-level embedding alignment network for 2D image-based 3D object retrieval. In: Proceedings of the 27th ACM international conference on multimedia. 2019, p. 1667–75.

[24] Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation. In: International conference on machine learning. PMLR; 2015, p. 1180–9.

[25] Su Y, Li Y, Nie W, Song D, Liu A-A. Joint heterogeneous feature learning and distribution alignment for 2D image-based 3D object retrieval. IEEE Trans Circuits Syst Video Technol 2019;30(10):3765–76.

[26] Liang Q, Li Q, Nie W, Liu A-A. PAGN: perturbation adaption generation network for point cloud adversarial defense. Multimedia Syst 2022;1–9.

[27] Taha B, Hayat M, Berretti S, Hatzinakos D, Werghi N. Learned 3d shape representations using fused geometrically augmented images: Application to facial expression and action unit detection. IEEE Trans Circuits Syst Video Technol 2020;30(9):2900–16.

[28] Kim J, Hua B-S, Nguyen T, Yeung S-K. Minimal adversarial examples for deep learning on 3d point clouds. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021, p. 7797–806.

[29] Zhou H, Chen D, Liao J, Chen K, Dong X, Liu K, Zhang W, Hua G, Yu N. Lg-gan: Label guided adversarial network for flexible targeted attack of point cloud based deep networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020, p. 10356–65.

[30] Feng Y, Feng Y, You H, Zhao X, Gao Y. Meshnet: Mesh neural network for 3d shape representation. In: Proceedings of the AAAI conference on artificial intelligence, Vol. 33. 2019, p. 8279–86.

[31] Cai W, Liu D, Ning X, Wang C, Xie G. Voxel-based three-view hybrid parallel network for 3D object classification. Displays 2021;69:102076.

[32] Wu Z, Song S, Khosla A, Yu F, Zhang L, Tang X, Xiao J. 3D shapenets: A deep representation for volumetric shapes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015, p. 1912–20.

[33] Li J, Chen BM, Lee GH. So-net: Self-organizing network for point cloud analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, p. 9397–406.

[34] Nie W, Zhao Y, Niea J, Liu A-A, Zhaob S. CLN: Cross-domain learning network for 2D image-based 3D shape retrieval. IEEE Trans Circuits Syst Video Technol 2021.

[35] Qin J, Yuan S, Chen J, Amor BB, Fang Y, Hoang-Xuan N, Chu C-B, Nguyen-Ngoc K-N, Cao T-T, Ngo N-K, et al. SHREC'22 track: Sketch-based 3D shape retrieval in the wild. Comput Graph 2022;107:104–15.

[36] Hu N, Zhou H, Liu A-A, Huang X, Zhang S, Jin G, Guo J, Li X. Collaborative distribution alignment for 2D image-based 3D shape retrieval. J Vis Commun Image Represent 2022;103426.

[37] Lowe DG. Distinctive image features from scale-invariant keypoints. Int J Comput Vis 2004;60:91–110.

[38] Su H, Maji S, Kalogerakis E, Learned-Miller E. Multi-view convolutional neural networks for 3d shape recognition. In: Proceedings of the IEEE international conference on computer vision. 2015, p. 945–53.

[39] Massa F, Russell BC, Aubry M. Deep exemplar 2d-3d detection by adapting from real to rendered views. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, p. 6024–33.

[40] Nie W, Wang K, Liang Q, He R. Panorama based on multi-channel-attention CNN for 3D model recognition. Multimedia Syst 2019;25(6):655–62.

[41] Chen C, Fu Z, Chen Z, Jin S, Cheng Z, Jin X, Hua X-S. Homm: Higher-order moment matching for unsupervised domain adaptation. In: Proceedings of the AAAI conference on artificial intelligence, Vol. 34. 2020, p. 3422–9.

[42] Cui S, Wang S, Zhuo J, Su C, Huang Q, Tian Q. Gradually vanishing bridge for adversarial domain adaptation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020, p. 12455–64.

[43] Sun B, Saenko K. Deep coral: Correlation alignment for deep domain adaptation. In: European conference on computer vision. Springer; 2016, p. 443–50.

[44] Tzeng E, Hoffman J, Zhang N, Saenko K, Darrell T. Deep domain confusion: Maximizing for domain invariance. 2014, arXiv preprint arXiv:1412.3474.

[45] Long M, Cao Z, Wang J, Jordan MI. Conditional adversarial domain adaptation. Adv Neural Inf Process Syst 2018;31.

[46] Liu H, Long M, Wang J, Jordan M. Transferable adversarial training: A general approach to adapting deep classifiers. In: International conference on machine learning. PMLR; 2019, p. 4013–22.

[47] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, p. 770–8.

[48] Phong BT. Illumination for computer generated pictures. Commun ACM 1975;18(6):311–7.

[49] Cui S, Jin X, Wang S, He Y, Huang Q. Heuristic domain adaptation. Adv Neural Inf Process Syst 2020;33:7571–83.

[50] Li S, Xie M, Lv F, Liu CH, Liang J, Qin C, Li W. Semantic concentration for domain adaptation. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021, p. 9102–11.

[51] Li W, Liu A, Bui N-M, Cen Y, Zenian Chen H-HC-N, Diep G-H, Do T-L, Doubrovski EL, Wang CC, Wang S, et al. Shrec 2019-monocular image based 3d model retrieval. In: Eurographics 2019 workshop 3D object retrieval. 2019, p. 1–7.

[52] Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. Ieee; 2009, p. 248–55.

[53] Sedaghat N, Zolfaghari M, Amiri E, Brox T. Orientation-boosted voxel nets for 3D object recognition. 2016, arXiv preprint arXiv:1604.03351.

[54] Savva M, Yu F, Su H, Aono M, Chen B, Cohen-Or D, Deng W, Su H, Bai S, Bai X, et al. Large-scale 3d shape retrieval from shapenet core55. In: Proceedings of the eurographics 2016 workshop on 3d object retrieval. 2016, p. 89–98.

[55] Chen D-Y, Tian X-P, Shen Y-T, Ouhyoung M. On visual similarity based 3D model retrieval. In: Computer graphics forum, Vol. 22. Wiley Online Library; 2003, p. 223–32.

[56] Shilane P, Min P, Kazhdan M, Funkhouser T. The princeton shape benchmark. In: Proceedings shape modeling applications, 2004. IEEE; 2004, p. 167–78.

[57] Liu A-A, Nie W-Z, Gao Y, Su Y-T. View-based 3-D model retrieval: A benchmark. IEEE Trans Cybern 2017;48(3):916–28.

[58] Liu A-A, Zhang C, Li W, Gao X, Sun Z, Li X. Self-supervised auxiliary domain alignment for unsupervised 2D image-based 3D shape retrieval. IEEE Trans Circuits Syst Video Technol 2022.

[59] Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9(11).

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.

E-mail: dan.song@tju.edu.cn

Xue-Jing Jiang received the bachelor's degree from Wuhan University of Technology, China. She is currently pursuing the master's degree with Tianjin University, Tianjin, China. Her research interests include 3D object retrieval and computer vision.

E-mail: jxj970304@163.com

Yue Zhang is with the School of Biomedical Engineering, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui, China and also with the Center for Medical Imaging, Robotics, Analytic Computing & Learning (MIRACLE), Suzhou Institute for Advanced Research, University of Science and Technology of China, Suzhou, Jiangsu, China. She received her B.S. from Southwest Jiaotong University, China in 2016, and a Ph.D. from Zhejiang University, China in 2021. Her research interests include computer vision, medical image analysis and deep learning.

E-mail: yue_zhang@ustc.edu.cn

Fang-Lue Zhang received the Ph.D. degree from Tsinghua University, in 2015. He is currently a Lecturer with the Victoria University of Wellington, Wellington, New Zealand. His research interests include image and video editing, computer vision, and computer graphics. He is a member of ACM. He received the Victoria Early-Career Research Excellence Award, in 2019, and the Fast-Start Marsden Grant from the New Zealand Royal Society, in 2020.

E-mail: fanglue.zhang@vuw.ac.nz

Yao Jin received the B.S. degree in apparel engineering and the M.S. degree in computer science from Zhejiang Sci-Tech University, China, in 2007 and 2010, respectively, and the Ph.D. degree from Zhejiang University, China, in 2015. He is currently a associate professor with the Department of digital media technology, Zhejiang Sci-Tech University. His research interests include computer graphics and digital geometry processing.

E-mail: jinyao@zstu.edu.cn

Yun Zhang is currently a Professor at the Communication University of Zhejiang in China. He received his doctoral degree from Zhejiang University in 2013. Before that, he received Bachelor and Master degrees from Hangzhou Dianzi University in 2006 and 2009, respectively. He visited the Visual Computing Group of Cardiff University in 2018 and 2023, and the Computational Media Innovation Centre of Victoria University of Wellington in 2019. His research interests include Computer Graphics, Image and Video Editing, Computer Vision. He is a Senior Member of CCF.

E-mail: zhangyun@cuz.edu.cn