Hierarchical Instance Feature Alignment for 2D Image-Based 3D Shape Retrieval

Heyu Zhou , **Weizhi Nie*** , **Wenhui Li** , **Dan Song** and **An-An Liu***
School of Electrical and Information Engineering, Tianjin University, China
zhy_std@163.com, {weizhinie, liwenhui, dan.song}@tju.edu.cn, anan0422@gmail.com

Abstract

2D image-based 3D shape retrieval has become a hot research topic since its wide industrial applications and academic significance. However, existing view-based 3D shape retrieval methods are restricted by two settings, 1) learn the common-class features while neglecting the instance visual characteristics, 2) narrow the global domain variations while ignoring the local semantic variations in each category. To overcome these problems, we propose a novel hierarchical instance feature alignment (HIFA) method for this task. HIFA consists of two modules, cross-modal instance feature learning and hierarchical instance feature alignment. Specifically, we first use CNN to extract both 2D image and multi-view features. Then, we maximize the mutual information between the input data and the high-level feature to preserve as much as visual characteristics of an individual instance. To mix up the features in two domains, we enforce feature alignment considering both global domain and local semantic levels. By narrowing the global domain variations we impose the identical large norm restriction on both 2D and 3D feature-norm expectations to facilitate more transferable possibility. By narrowing the local variations we propose to minimize the distance between two centroids of the same class from different domains to obtain semantic consistency. Extensive experiments on two popular and novel datasets, MI3DOR and MI3DOR-2, validate the superiority of HIFA for 2D image-based 3D shape retrieval task.

1 Introduction

1.1 Motivation

3D shape retrieval aims to match the similar shapes in the gallery given a query object, which can be the sketch [Li and et al., 2012], 2D image [Zhou *et al.*, 2019a] or 3D shape [Zhou *et al.*, 2019b]. Because of its significant applications in 3D printing, digital entertainment, medical engineering and computer aided design [Lu *et al.*, 2019a; 2019b; Cheng *et al.*, 2018; Hong *et al.*, 2016], 3D shape retrieval has attracted much attention from both industry and academic fields. Although deep learning has been adopted for 3D shape retrieval community and achieved significant performances, few literature focuses on 2D image-based 3D shape retrieval, which has not been well studied to date. They still yield weak performances caused by the following critical problems:

Difficulty in narrowing the gap between the 2D and 3D domains without strong dependence on annotated 3D shapes. Most existing 3D shape retrieval methods employ supervised learning and are usually based on learning multi-view context [Zhou *et al.*, 2019b], learning discriminative features from point cloud [Qi *et al.*, 2017a] or discovering the structure information from voxel [Maturana and Scherer, 2015]. These methods rely on substantial annotated 3D shapes while manually labeling may be unreliable and is time-consuming, which limits the practicability and usability of the supervised learning methods for the real applications. [Liu *et al.*, 2018] focuses on the unsupervised feature learning to solve this problem by transferring the knowledge from the label-rich 3D domain to the unlabled 3D domain. However, this method is based on the assumption that there must exist a common subspace between the two domains. This assumption can not hold for 2D image-based 3D shape retrieval task, because the gap between 2D and 3D is significant due to the background complexity, light intensity, viewpoint variance and modality discrepancy. Therefore, it's mandatory to develop sophisticated algorithms to narrow the gap between the 2D and 3D domains without annotated 3D data for retrieval.

Difficulty in enforcing the alignment of global domain statistics and local semantic information for cross-modal data. 2D image-based 3D shape retrieval is challenging mainly due to the great global domain and local semantic variations. By global domain variation we mean that 3D shape and 2D image are essentially heterogeneous since the data formats are completely different as the basic unit of the 2D image and 3D shape are pixel and voxel, respectively. By local semantic variation we mean the class-level difference across two domains including background, resolution, size, lightness, occlusion, viewpoint, etc. can be significant even when we describe the shape by multi-view images. However, the current domain adaptation methods only focus on

*Corresponding authors: Weizhi Nie & An-An Liu

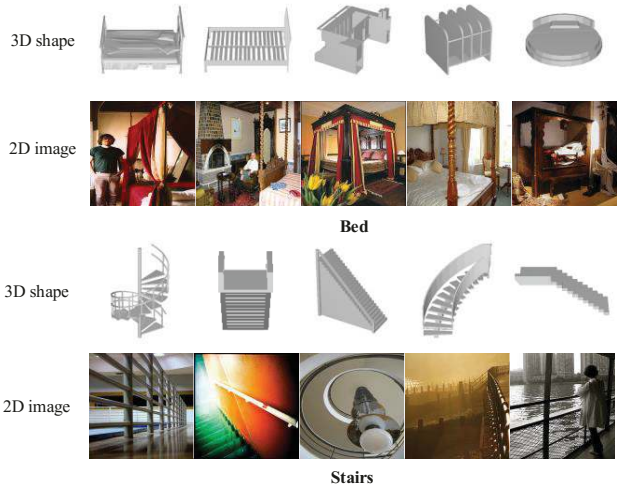


Figure 1: Some 2D image and 3D shape samples on MI3DOR.

the global domain statistics alignment by confusing the domain discriminator or minimizing the domain-level distances, such as CORAL [Long *et al.*, 2013], MMD [Borgwardt *et al.*, 2006], RevGard [Ganin and Lempitsky, 2015] and so on. As shown in Figure. 1, the gap between two domains may be difficult to align in the local semantic level even with the well domain confusion. Thus, it is urgently desired to design a 2D image-based 3D shape retrieval method with both global domain statistics and local semantic information alignment.

Difficulty in learning the discriminative instance feature for both 2D image and 3D shape. Although deep learning has achieved great progress on feature learning for both 2D image and 3D shape, it only focuses on the correlations between the encoder output and the label while neglecting the connections between the output feature and the input. Specifically, for different instances of the same class, the existing methods tend to extract the common-class features while ignoring the unique visual characteristics of individual instances, which may have negative influence on the performance. For examples, in Figure.1, there are two classes, bed and stairs, which have five different instances in each domain, respectively. The common features of both classes are not discriminative enough to keep the visual details clearly, which may weaken the ability of feature representation. Therefore, it's necessary to learn the discriminative instance feature for preserving as much visual characteristics of the cross-data as possible.

To overcome these problems, we propose a novel hierarchical instance feature alignment (HIFA) network for 2D image-based 3D shape retrieval. As shown in Figure. 2, we first employ the identical CNN to extract the visual features for both 2D image and 3D shape (a set of multi-view images) with the max-pooling operation [Su *et al.*, 2015]. To preserve the instance visual characteristics, we maximize the mutual information between the 2D image / 3D shape and the high-level features. To narrow the gap between the two domains, we propose the hierarchical instance feature alignment module to enforce both global and local alignment. Since domain shift is decided by misaligned feature-norm expectations [Xu *et al.*, 2019], we impose the identical large norm restriction on both the 2D and 3D feature-norm expectations to facilitate more transferable possibility. However, well domain alignment is not enough for retrieval since the semantic information for each class may be neglected. We further propose to eliminate the class-level variations by minimizing the distance between the two centroids of the same class from different domains.

1.2 Contributions

In summary, the main contributions of this paper as follows:

- We propose an unsupervised 2D-image based 3D shape retrieval method, which can jointly eliminate the global domain and local semantic variations by the hierarchical instance feature alignment module.
- Different from the existing 3D shape retrieval methods concentrating on learning the common-class features, we first propose to learn the discriminative instance feature, which can preserve as much as visual characteristics for retrieval task.
- Experimental analysis on two widely used datasets, MI3DOR and MI3DOR-2, demonstrates that our method can outperform the state-of-the-art 2D image-based 3D shape retrieval approaches.

2 Related Work

2.1 3D Shape Retrieval

The typical 3D shape retrieval methods mainly includes view-based methods and model-based methods. View-based methods usually describe 3D shape by the multi-view image set and encode individual views by the 2D CNN. [Su *et al.*, 2015] proposed Multi-view Convolutional Neural Networks (MVCNN), which fused the multi-view features by a max-pooling layer across different views. Unlike the view-to-shape setting in MVCNN, [Feng *et al.*, 2018] proposed a view-group-shape architecture, group-view convolutional neural networks (GVCNN), which divided views into several clusters based on visual similarity and then fused their features in both group-level and shape-level. Model-based methods usually directly extract the features from origin 3D data, such as point cloud and voxel. [Qi *et al.*, 2017a] proposed PointNet, which first applied CNN to deal with unordered point sets. However, PointNet can not capture the local structures of the point sets, which resulted in the less discriminative of the feature representation. To resolve this problem, [Qi *et al.*, 2017b] further proposed PointNet++, which utilized the neighbor points at multi-scale to capture the local structure information. [Maturana and Scherer, 2015] proposed VoxNet. This method utilized supervised convolutional network to encode the binary voxel grids for 3D shape representation.

2.2 Domain Adaptation

Domain adaptation aims to narrow the gap between two data distributions, usually with unlabeled target data and labeled source data. A common strategy for domain adaptation is to learn domain-invariant features. For example, [Long *et*

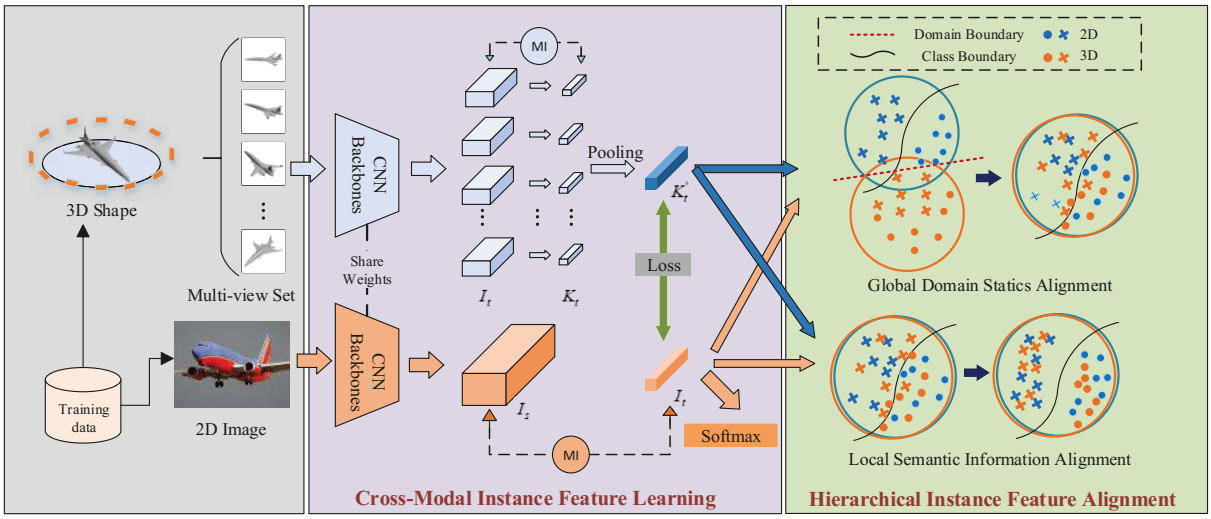


Figure 2: Illustration of HIFA which mainly contains the cross-modal feature learning and hierarchical instance feature alignment procedures.

al., 2013] proposed joint distribution adaptation (JAN) network, which aimed to minimize the joint MMD on the specific domain adaptation layer across two domains to learn the domain-invariant features. [Ganin and Lempitsky, 2015] proposed domain adversarial neural network (DANN). This method imposed an additional domain discriminator to the traditional CNN network to enforce the domain alignment. Based on DANN, which only narrowed the domain shift in the global distribution statistics level, [Xie *et al.*, 2018] proposed to align the feature centroids of the same class from different domains to preserve the semantic consistency in local class level. [Wang *et al.*, 2018] focused on learning a domain-invariant classifier instead of feature in grassmann manifold with structural risk minimization to perform distribution alignment. [Xu *et al.*, 2019] enabled the CNN network to learn the transferable features across two domains by imposing the large identical norm restriction on feature-norm expectations. Unlike these feature adaptation methods, [Zhang *et al.*, 2017] proposed joint geometrical and statistical alignment (JGSA), which projected the source and target data into low-dimensional subspaces with two coupled map functions. In the subspace, both the geometrical and statistical shift across two domains were reduced.

3 Method

3.1 Problem Definition

In unsupervised 2D image-based 3D shape retrieval, we have access to the labeled source domain (2D) $\mathcal{S} = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{n_s}$ where $y_i^s \in \mathcal{Y} = \{1, \ldots, C\}$ and unlabeld target domain (3D) $\mathcal{T} = \{\mathbf{x}_j^t\}_{j=1}^{n_t}$. $\mathcal{S}$ and $\mathcal{T}$ are assumed to be different but share the same label distribution. The key problem of this task is to learn a feature function G that projects 2D images and 3D shapes into a common embedding space. In our paper, G denotes all the layers up to fc_7 including mutual information estimation operation and F represents the source classifier.

3.2 Cross-Modal Instance Feature Learning

This step mainly consists of two modules, the basic CNN network for visual feature extraction on both 2D and 3D domains and a mutual information maximization module for discriminative feature learning.

Visual Feature Extraction

Given a 3D shape, we can set the preset virtual cameras around it to obtain the multi-view image set $V = \{v_i\}_{i=1}^{N}$ (N represents the view number) by the Phong reflection model [Phong, 1975]. Similar to the most related works, we set the view number as 12 and the camera array setting follows [Su *et al.*, 2015]. To extract the 2D image and multi-view features, a shared 2D CNN is employed. The backbone CNN in our work is AlexNet [Krizhevsky *et al.*, 2012] pre-trained on ImageNet. The original AlexNet has 8 layers, which contains 5 convolutional layers ($conv1$-5) and three full-connected layers (fc_6-fc_8). Then we can obtain low-level image feature map $I_s \in \mathbb{R}^{W \times H \times C}$ and multi-view embedding tensor $I_t = \left[i_t^1, i_t^2, \ldots, i_t^N\right]^T \in \mathbb{R}^{N \times W \times H \times C}$ from the output of $conv5$, the high-level image embedding $K_s \in \mathbb{R}^D$ and multi-view image embedding matrix $K_t = \left[k_t^1, k_t^2, \ldots, k_t^N\right]^T \in \mathbb{R}^{N \times D}$ from the output of fc_7. Note that W, H, C, D are $6, 6, 256, 256$ in our experiment. Finally, we impose the max-pooling operation on K_t across different views to obtain the final 3D shape descriptor $K_t^{'} \in \mathbb{R}^D$ for retrieval.

Mutual Information Maximization

The traditional deep learning works pay more attention to the input features and the associated labels, in which the features focus on preserving the commonality of categories. However, the common-class features are not suitable for 3D shape retrieval task. In other words, the intra-class distance should be smaller while the inter-class distance should be big enough in two domains. To meet these requirements, we maximize the mutual information (MI) between the low-level convolution features (I_s, I_t) and the high-level embeddings (K_s, K_t).

Based on [Belghazi *et al.*, 2018], the MI estimation can be computed as the Kullback-Leibler divergence between the joint $\mathbb{P}_{MN}$, and the product of the marginals $\mathbb{P}_M \otimes \mathbb{P}_N$, *i.e.* $I(M;N) = D_{KL}(\mathbb{P}_{MN} \| \mathbb{P}_M \otimes \mathbb{P}_N), M \in \{I_s, I_t\}, N \in \{K_s, K_t\}$. However, the unbounded upper limit of D_{KL} may result in the infinite result. To address this problem, [Belghazi *et al.*, 2018; Hjelm *et al.*, 2019] propose to represent the KL-divergence with Donsker-Varadhan representation [Donsker and Varadhan, 1975] to obtain the lower-bound. Specifically, the MI estimator can be rewritten as:

$$D_{KL}(\mathbb{P}\|\mathbb{Q}) \geq \sup_{T \in \mathcal{F}} \mathbb{E}_{\mathbb{P}}[T] - \log\left(\mathbb{E}_{\mathbb{Q}}\left[e^{T}\right]\right) \tag{1}$$

where T can be implemented as a discriminator function with parameter θ_k in deep neural network. Therefore, $I_\theta(M;N)$ can be computed by:

$$\begin{aligned} I_{\theta_k}(M;N) :=& \mathbb{E}_{\mathbb{P}_{MN}}\left[T_{\theta_k}(M,N)\right] - \\ & \log\left(\mathbb{E}_{\mathbb{P}_M \otimes \mathbb{P}_N}\left[e^{T_{\theta_k}(M,N)}\right]\right) \end{aligned} \tag{2}$$

During the training procedure, the MI estimator implemented by a discriminator can approximate the mutual information with arbitrary accuracy gradually by the gradient back propagation. In our architecture, the input of MI estimator is the combination of the low-level features I_s/I_t and high-level features K_s/K_t. "Fake samples" are constructed by combining the same low-level features with high-level features obtained from another image/shape. Note that the 2D and 3D domains share an identical MI estimator. Then the discriminator is trained to distinguish the "real or fake" samples, and we can calculate the binary cross-entropy loss $\mathcal{L}_{MI}$, which can be regarded as the estimation of MI.

3.3 Hierarchical Instance Feature Alignment

This step mainly consists of two modules, global domain statistics alignment and local semantic information alignment for reducing the domain shift.

Global Domain Statistics Alignment

Although we can obtain more discriminative feature for individual instance (2D image or 3D shape) with the help of MI estimation, the gap between two domains has not been reduced. As shown in Figure. 2, due to the large global domain and local semantic variations, the overlap of the feature distributions is small. Inspired by the Misaligned-Feature-Norm Hypothesis [Xu *et al.*, 2019] that the domain shift is decided by the misaligned feature-norm expectations and consequently the equal value of feature-norm expectations in two domains can reduce the domain shift, we propose to enforce the global domain statistics alignment between the 2D and 3D domains by the Maximum Mean Feature Norm Discrepancy (MMFND),

$$\begin{aligned} \mathrm{MMFND}\left[G, F, \mathcal{D}_s, \mathcal{D}_t\right] := \sup_{G,F} \Bigg(& \frac{1}{n_s} \sum_{x_i \in \mathcal{D}_s} \left\| F_{L-1}\left(G\left(x_i\right)\right)\right\|_2 \\ & - \frac{1}{n_t} \sum_{x_i \in \mathcal{D}_t} \left\| F_{L-1}\left(G\left(x_i\right)\right)\right\|_2 \Bigg) \end{aligned} \tag{3}$$

where $\mathcal{H}$ denotes the combination of all the potential feature extraction functions composited by the L_2-norm, i.e., $h(x) = (\|\cdot\|_2 \circ G)(x)$. Intuitively, the function class $\mathcal{H}$ without any restriction may result in the great deviation of the upper bound from zero. To avoid it, we impose the identical large norm restriction R on the feature-norm expectations on both 2D and 3D domains. Consequently, the domain gap measured by MMFND will gradually vanish to zero and we can obtain domain-invariant features for retrieval. Thus, the global domain confusion loss $\mathcal{L}_G$ can be written as:

$$\mathcal{L}_G = L_d\left(\frac{1}{n_s} \sum_{x_i^s \in \mathcal{S}} h\left(x_i^s\right), R\right) + L_d\left(\frac{1}{n_t} \sum_{x_i^t \in \mathcal{T}} h\left(x_i^t\right), R\right) \tag{4}$$

where $L_d(.)$ represents L_2-norm penalty. We empirically set R as 25 by following [Xu *et al.*, 2019] since the lower value is prone to achieve lower accuracy on target domain while a sufficiently large R may lead to the gradient explosion.

Local Semantic Information Alignment

As shown in Figure. 2, the global domain statistics alignment does not imply a local semantic class-to-class alignment. For example, the visual feature of 3D vase may be mapped nearby the embedding of 2D cup while satisfying the global domain statistics alignment since the unlabeled 3D domain adds the semantic ambiguity. Conversely, the 2D domain can preserve the semantic consistency by the classification loss,

$$\mathcal{L}_C = \frac{1}{n_s} \sum_{i=1}^{n_s} L_c\left(F \circ G\left(\mathbf{x}_i^s\right), \mathbf{y}_i^s\right) \tag{5}$$

where L_c represents the classification cross-entropy loss. Thus, it is necessary to pursue the local semantic information alignment under the absence of true 3D labels. We solve this problem by minimizing the distance between two centroids of the same class from different domains. Unfortunately, we have no access to label information from 3D domain. To circumvent the impossibility of local semantic alignment, we assign pseudo labels to 3D samples with the classifier F. Finally, the local semantic information alignment can be narrowed by minimizing,

$$\mathcal{L}_L = \sum_{k=1}^{K} \left\| C_S^k - C_T^k \right\|^2 \tag{6}$$

where $\mathcal{L}_L$ is the local semantic loss, K denotes the class number and C_S^k, C_T^k are the tth class feature centroid on 2D and 3D domains respectively. We can obtain $2K$ centroids in total. Obviously, there must be some false pseudo-labeled samples and they may have a negative effect on adaptation. Fortunately, when computing the centroids, all pseudo-labels (false or true) samples are used together and the negative effect brought by fake samples can be neutralized by true pseudo-labeld samples to some extent.

The total objective loss can be written as:

$$\mathcal{L}_{Total} = \mathcal{L}_C + \lambda \mathcal{L}_{MI} + \beta \mathcal{L}_G + \gamma \mathcal{L}_L \tag{7}$$

where λ, β and γ are hyper-parameters to trade off different loss functions. We set λ, β as $0.05, 1$ and $\gamma = \frac{2}{1+\exp(-10 \cdot p)} - 1$, respectively, where p denotes the training progress varying

from 0 to 1. This setting can further reduce the negative effect caused by pseudo-labeled 3D shapes since the pseudo label confidence is low in the early training process. As the discrepancy between the two domains decreases, the pseudo label confidence will increase and consequently the local semantic loss will play a more important role in enforcing the fine-grained class-to-class alignment.

4 Experimental Settings

4.1 Datasets and Evaluation Criteria

MI3DOR. The MI3DOR dataset [Zhou *et al.*, 2019a] contains 21 categories with 21,000 images and 7,690 3D shapes. There are 10,500 images and 3,842 3D shapes for training, while 10,500 images and 3,848 3D shapes for testing.

MI3DOR-2. The MI3DOR-2 dataset [Zhou *et al.*, 2019a] is composed of 40 categories with 19,694 images and 3,982 3D shapes. MI3DOR-2 dataset is divided into two parts, 19,294 images/3,182 3D shapes are used as the training set and 400 images/800 3D shapes are used for testing.

Evaluation Criteria. For fair comparisons, we employ the common retrieval evaluation criteria as [Zhou *et al.*, 2019a]. It includes the Nearest Neighbor (NN), First Tier (FT), Second Tier (ST), F_measure (F), Discounted Cumulative Gain (DCG) and Average Normalized Retrieval Rank (ANMRR). The value of these criteria ranges from 0 to 1 and the higher value means well performance other than ANMRR.

4.2 Compared Methods

We mainly compare HIFA with several representative methods: 1) basic deep learning method, AlexNet [Krizhevsky *et al.*, 2012], 2) traditional transfer learning method, MEDA [Wang *et al.*, 2018] and JGSA [Zhang *et al.*, 2017]. 3) deep transfer learning method, JAN [Long *et al.*, 2017], RevGard [Ganin and Lempitsky, 2015] and DLEA [Zhou *et al.*, 2019a]. Since the most mentioned methods (other than DELA) are proposed for 2D visual domain adaptation, the max-pool pooling operation are adopted to fuse multi-view features.

5 Experimental Results

5.1 Comparison with the State-of-the-art Methods

Table. 1 and Table. 2 present the experiment results of HIFA and the comparisons among representative methods on MI3DOR and MI3DOR-2, respectively. It's obvious that HIFA can achieve the best retrieval performance. Specifically, on MI3DOR, HIFA is superior to the comparisons with the gain of 1.83%-83.49%, 10.75%-91.33%, 7.26%-63.75%, 5.59%-52.53%, 9.55%-89.56% with respect to NN, FT, ST, F, DCG, and with the decline of 14.01%-45.73% according to ANMRR. On MI3DOR-2, HIFA is superior to the comparisons with the gain of 3.45%-39.96%, 2.70%-60.56%, 4.26%-45.49%, 2.00%-60.56%, 0.84%-56.14% with respect to NN, FT, ST, F, DCG, and with the decline of 2.59%-34.34% according to ANMRR. Moreover, we can notice the following observations:

	NN	FT	ST	F	DCG	ANMRR
AlexNet	0.424	0.323	0.469	0.099	0.345	0.667
MEDA	0.430	0.344	0.501	0.046	0.361	0.646
JGSA	0.612	0.443	0.599	0.116	0.473	0.541
JAN	0.446	0.343	0.495	0.085	0.364	0.647
RevGard	0.650	0.505	0.643	0.112	0.542	0.474
DLEA	0.764	0.558	0.716	0.143	0.597	0.421
HIFA	**0.778**	**0.618**	**0.768**	**0.151**	**0.654**	**0.362**

Table 1: Retrieval Comparisons on MI3DOR

	NN	FT	ST	F	DCG	ANMRR
AlexNet	0.518	0.355	0.488	0.355	0.383	0.629
MEDA	0.570	0.392	0.523	0.392	0.425	0.590
JGSA	0.585	0.405	0.533	0.405	0.433	0.577
JAN	0.608	0.501	0.646	0.501	0.527	0.484
RevGard	0.623	0.467	0.614	0.467	0.503	0.514
DLEA	0.700	0.555	0.681	0.555	0.593	0.424
HIFA	**0.725**	**0.570**	**0.710**	**0.570**	**0.598**	**0.413**

Table 2: Retrieval Comparisons on MI3DOR-2

HIFA vs. DLEA. DLEA is the most recent method achieving the best performance for this task since it can jointly realize the feature learning and distribution alignment by the domain discriminator and centroid alignment operation. Compared with DLEA, HIFA achieves the gain of 1.83%/3.45%, 10.75%/2.70%, 7.26%/4.26%, 5.59%/2.00% with respect to NN, FT, ST, F, DCG, and the decline of 14.01%/2.59% with respect to ANMRR on MI3DOR and MI3DOR-2, respectively. This result can demonstrate that the mutual information maximization module can preserve as much visual characteristic as possible of individual instance (2D image and 3D shape) for retrieval. Comparatively, DLEA ignores the instance visual information and only focuses on the common-class features, which will weaken the retrieval performance.

HIFA vs. others. Compared with other methods, HIFA has two main advantages. First, HIFA can jointly eliminate the global domain and local semantic variations by the hierarchical instance feature alignment module. Comparatively, previous methods either enforce the domain alignment or project the features into the identical subspace while neglecting the semantic consistency in two domains. Second, HIFA can learn the discriminative instance features for retrieval. Comparatively, the other methods only concentrate on narrowing the gap between the 2D and 3D domains while neglecting the significance of feature representation for retrieval.

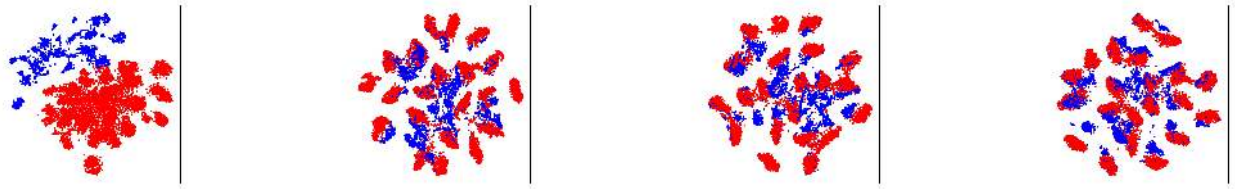
5.2 Discussion

In this section, we explore the retrieval influence caused by the global domain statistics alignment (-G), local semantic information alignment (-L) and mutual information (MI) maximization modules. Besides, we also visualize the features to further analyze the proposed method.

Effect of Global Domain Statistics Alignment. As shown in Table. 3, HIFA-G can achieve the gain of 62.03%/26.45%, 71.16%/46.76%, 58.21%/36.07%, 33.33%/46.76, 72.17%/41.25% with respect to NN, FT, ST, F, DCG, and the decline of 37.18%/26.07% on ANMRR comparing against AlexNet on MI3DOR and MI3DOR-2,

Method	MI3DOR						MI3DOR-2					
	NN	FT	ST	F	DCG	ANMRR	NN	FT	ST	F	DCG	ANMRR
AlexNet	0.424	0.323	0.469	0.099	0.345	0.667	0.518	0.355	0.488	0.355	0.383	0.629
HIFA-G	0.687	0.569	0.742	0.132	0.594	0.419	0.655	0.521	0.664	0.521	0.541	0.465
HIFA-GL	0.753	0.600	0.752	0.148	0.637	0.379	0.673	0.547	0.686	0.547	0.573	0.437
HIFA-GL (W/ MI)	**0.778**	**0.618**	**0.768**	**0.151**	**0.654**	**0.362**	**0.725**	**0.570**	**0.710**	**0.570**	**0.598**	**0.413**

Table 3: Performances with respect to different architectures on MI3DOR and MI3DOR-2.



(a) AlexNet (b) HIFA-G (c) HIFA-GL (d) HIFA-GL(W/MI)

Figure 3: Feature visualization on MI3DOR.

respectively. The basic deep learning method AlexNet without feature alignment can not narrow the domain discrepancy caused by the global domain and local semantic variations, and consequently the gap between the 2D and 3D domain is large. Comparatively, HIFA-G imposes the large norm constraint on feature-norm expectations to bridge the significant 2D-to-3D domain gap.

Effect of Local Semantic Information Alignment. As shown in Table. 3, HIFA-GL can achieve the gain of 9.61%/2.75%, 5.45%/4.99%, 1.35%/3.31%, 12.12%/4.99%, 7.24%/3.20% with respect to NN, FT, ST, F, DCG, and the decline of 9.55%/6.02% on ANMRR comparing against HIFA-G on MI3DOR and MI3DOR-2, respectively. Global domain statistics alignment can contribute to generate domain-invariant features for this task. However, domain-invariance doesn't mean semantic consistency and the class-level difference still exists. Comparatively, HIFA-GL can narrow the local semantic variations by minimizing the distance of class centroids from two domains.

Effect of Mutual Information Maximization. As shown in Table. 3, HIFA-GL (W/ MI) can achieve the gain of 3.32%/7.73%, 3.00%/4.20%, 2.13%/3.50%, 2.03%/4.20%, 2.07%/4.36%, 9.55%/0.84% with respect to NN, FT, ST, F, DCG, and the decline of 1.70%/5.49% on ANMRR comparing against HIFA-GL on MI3DOR and MI3DOR-2, respectively. It's not enough to achieve the well performance only by enforcing the global and local alignment, since the significant visual discriminative information of individual instance has been neglected. Therefore, HIFA with MI can further improve retrieval performance.

Feature Visualization. As shown in Figure. 3, we visualized the features obtained from different settings on MI3DOR dataset by t-SNE [Maaten and Hinton, 2008]. Note that the red and blue points represent the features from the 2D image and 3D shape domains, respectively. A transferable and discriminative feature mapping should mix up the red and blue points, and meanwhile the points can be easier to recognize its cluster. Other than AlexNet, which does nothing to enforce the alignment across two domains, other approaches can learn the domain-invariant features for this task since the global domain statistics alignment module has been adopted. Besides, the features learned by HIFA-GL and HIFA-GL (W/MI) are more suitable for retrieval task since the identical class across two domains aligns better and the gap between the red and blue points is smaller. Representations learned by HIFA-GL (W/ML) behave better than HIFA-GL and features in different classes from two domains are dispersed relatively instead of mixing up. This finding tells us that the mutual information maximization can further minimize the intra-class discrepancy and maximize the inter-class margin.

6 Conclusion

We propose a novel unsupervised hierarchical instance feature alignment network for 2D image-based 3D shape retrieval. It can jointly realize the cross-modal instance feature learning and hierarchical instance feature alignment. Different from previous methods, which only concentrate on the common-class features learning and global domain alignment, our work first introduce MI maximization to obtain discriminative instance feature and enforce the embedding alignment in both global and local levels. Experimental analysis can demonstrate the superiority of the proposed method.

Acknowledgments

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

References

[Belghazi *et al.*, 2018] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, R. Devon Hjelm, and Aaron C. Courville. Mutual information neural estimation. In *ICML*, pages 530–539, 2018.

[Borgwardt *et al.*, 2006] Karsten M. Borgwardt, Arthur Gretton, Malte J. Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alexander J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. In *ISMB*, pages 49–57, 2006.

[Cheng *et al.*, 2018] Zhiyong Cheng, Ying Ding, Xiangnan He, Lei Zhu, Xuemeng Song, and Mohan S. Kankanhalli. Aˆ3ncf: An adaptive aspect attention model for rating prediction. In *IJCAI*, pages 3748–3754, 2018.

[Donsker and Varadhan, 1975] Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time, i. *Math. Pures. Appl.*, 28:1–47, 1975.

[Feng *et al.*, 2018] Yifan Feng, Zizhao Zhang, Xibin Zhao, Rongrong Ji, and Yue Gao. GVCNN: group-view convolutional neural networks for 3d shape recognition. In *CVPR*, pages 264–272, 2018.

[Ganin and Lempitsky, 2015] Yaroslav Ganin and Victor S. Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, pages 1180–1189, 2015.

[Hjelm *et al.*, 2019] R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR*, 2019.

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

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1106–1114, 2012.

[Li and et al., 2012] Bo Li and Tobias Schreck et al. Shrec'12 track: Sketch-based 3d shape retrieval. In *3DOR*, pages 109–118, 2012.

[Liu *et al.*, 2018] Anan Liu, Shu Xiang, Wenhui Li, Weizhi Nie, and Yuting Su. Cross-domain 3d model retrieval via visual domain adaption. In *IJCAI*, pages 828–834, 2018.

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

[Long *et al.*, 2017] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Deep transfer learning with joint adaptation networks. In *ICML*, pages 2208–2217, 2017.

[Lu *et al.*, 2019a] Xu Lu, Lei Zhu, Zhiyong Cheng, Jingjing Li, Xiushan Nie, and Huaxiang Zhang. Flexible online multi-modal hashing for large-scale multimedia retrieval. In *MM*, pages 1129–1137, 2019.

[Lu *et al.*, 2019b] Xu Lu, Lei Zhu, Zhiyong Cheng, Liqiang Nie, and Huaxiang Zhang. Online multi-modal hashing with dynamic query-adaption. In *SIGIR*, pages 715–724, 2019.

[Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

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

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

[Qi *et al.*, 2017a] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, pages 77–85, 2017.

[Qi *et al.*, 2017b] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NIPS*, pages 5105–5114, 2017.

[Su *et al.*, 2015] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik G. Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *ICCV*, pages 945–953, 2015.

[Wang *et al.*, 2018] Jindong Wang, Wenjie Feng, Yiqiang Chen, Han Yu, Meiyu Huang, and Philip S. Yu. Visual domain adaptation with manifold embedded distribution alignment. In *ACM MM*, pages 402–410, 2018.

[Xie *et al.*, 2018] Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. Learning semantic representations for unsupervised domain adaptation. In *ICML*, pages 5423–5432, 2018.

[Xu *et al.*, 2019] Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *ICCV*, 2019.

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

[Zhou *et al.*, 2019a] Heyu Zhou, An-An Liu, and Weizhi Nie. Dual-level embedding alignment network for 2d image-based 3d object retrieval. In *ACM MM*, pages 1667–1675, 2019.

[Zhou *et al.*, 2019b] Heyu Zhou, An-An Liu, Weizhi Nie, and Jie Nie. Multi-view saliency guided deep neural network for 3d object retrieval and classification. *IEEE Trans. Multimedia*, 2019.