Domain Adaptive 3D Shape Retrieval from Monocular Images

Harsh Pal
IIT Bombay
Mumbai, India
palharsh.india@gmail.com

Ritwik Khandelwal
IIT Bombay
Mumbai, India
ritwikkhandelwal09@gmail.com

Shivam Pande
IIT Bombay
Mumbai, India
pandeshivam1993@gmail.com

Biplab Banerjee
IIT Bombay
Mumbai, India
getbiplab@gmail.com

Srikrishna Karanam
Adobe Research
Bengaluru, India
karanam@adobe.com

Abstract

In this work, we address the novel and challenging problem of domain adaptive 3D shape retrieval from single 2D images (DA-IBSR). While the existing image-based 3D shape retrieval (IBSR) problem focuses on modality alignment for retrieving a matchable 3D shape from a shape repository given a 2D image query, it does not consider any distribution shift between the training and testing image-shape pairs, making the performance of off-the-shelves IBSR methods subpar. In contrast, the proposed DA-IBSR addresses the non-trivial problem of modality shift as well distribution shift across training and test sets. To address these issues, we propose an end-to-end trainable model called DAIS-NET. *Our objective is to align the images and shapes separately from both domains while simultaneously learn a shared embedding space for the 2D and 3D modalities. The former problem is addressed by separately employing maximum mean discrepancy loss across the 2D images and 3D shapes of the two domains. To address the modality alignment, we incorporate the notion of negative sample mining and employ triplet loss to bridge the gap between positive 2D-3D pairs (of same class) and increase the separation between negative 2D-3D pairs (of different class). Additionally, we employ an entropy minimization strategy to align the unlabeled target domain data in the semantic space. To evaluate our proposed approach, we define the experimental setting of DA-IBSR on the following benchmarks: SHREC'14 ↔ Pix3D and ShapeNet ↔ SHREC'14. Considering the novelty of the problem statement, we have demonstrated that the issue of domain gap is prevalent by comparing our method with the existing literature. Additionally, through extensive evaluations, we demonstrate the capability of* DAIS-NET *to successfully mitigate this domain gap in image based 3D shape retrieval.*

1. Introduction

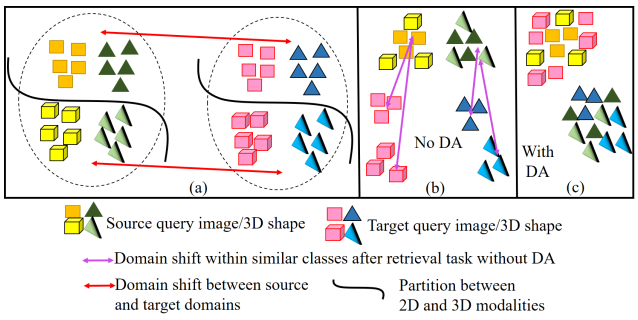

Figure 1. Graphical representation of domain adaptation in 3D shape retrieval. (a) Here, we see the query and shape for two non-overlapping domains across two classes (squares and triangles). (b). When no DA is applied for shape retrieval task, the query and shape embedding align in the shared space for source domain for same class, but not for target domain. (c) When retrieval task is carried out with DA, both the source and target domain align for the same class.

In today's era, we are witnessing a growth of available information from multiple sources, owing to unprecedented development in data sensing and data generation technology. This make it pivotal to have technology aimed at retrieving information from a practically endless sea of data. In the information technology, historically a lot of work has been focused in the field of data retrieval from text and image modalities [17, 4]. However, with advancement in computer vision domain, there has been a new research domain of retrieving 3D shapes using 2D images or sketches, primarily using deep learning based techniques. One of the

prime challenge that every 2D-3D retrieval system faces is the cross-modal differences between the two modalities [23]. Typically, most of the cross-modal retrieval problems have primarily worked to model the cross-modal differences between the 3D and 2D modalities. One way to accomplish this is by acquiring knowledge in a shared embedding space, where the matching image-3D object pair retains only the shared semantic content, discarding other non-relevant aspects for the purpose of matching [25]. Several works have been carried out in this direction. For example, [5] introduced a metric learning based shape retrieval framework from sketches. Here, the authors introduce a discriminative loss to enhance the distinction between different categories within each domain. Additionally, a correlation loss is employed to reduce the differences between sketch and 3D shape domains, aiming to minimize the domain discrepancy. Similarly, [22] presents an approach on SHREC'19 dataset, where one of the participating teams utilized triplet loss for aligning the shape and image modalities in the same metric space.

However, most of the research works in this area cater to the 2D images and 3D shapes only from a single domain, making the problem very restrictive and with limited generalization. As an example, a retrieval system trained to retrieve real world 3D objects from RGB images may not work well when a sketch image is fed to it (as a query) with the aim of retrieving a clay model of the 3D object. To mitigate these domain differences, the idea of unsupervised domain adaptation (UDA) is brought into picture. This has been graphically shown in Fig. 1. Historically, there has been an ample amount of work done in the intersection of computer vision and domain adaptation, primarily in the application of image classification [8], where the training is carried out with source domain (with labels available), while the model is then tested on a domain with disjoint distribution. Domain adaptation has its fair share of challenges, such as aligning the two domains in a shared space and establishing interclass differences within the target domain in absence of labels. Existing literature has shown to handle these issues with several metric learning based approaches, to bridge the gap between the two domains. For instance, [34] proposes a novel discriminative maximum mean discrepancy (MMD) approach that employs an intra-class trade off parameter and weighted inter-class distances to effectively control the degradation of feature discriminability. Similarly, [33] proposes a triplet loss guided by Bayesian perspective that adjusts the weights of intra-domain and inter-domain pair-wise samples, particularly focusing on hard positive and hard negative pairs, resulting in improved target pseudo labels. Additionally, enhancing the interclass differences within the classes of the target domain in absence of labels is a non-trivial problem. This problem is addressed in several works using the notion of unsupervised clustering [29, 21] and information theoretic approaches [26, 1].

However, in several existing cross-modal retrieval tasks, domain adaptation frameworks typically consider one of the modality as a separate domain, thereby working on aligning the modalities, where one of the modality may act as query while other one is to be retrieved [39]. Therefore, in our present work, we simultaneously address the problem of modality alignment between the 2D images and the 3D shapes, as well as domain alignment between the image and sketch queries and corresponding 3D shapes of the two domains. To this end, we incorporate the domain alignment through maximum mean discrepancy loss (MMD) for both 2D and 3D modalities. For modality alignment, we use the notion of triplet loss, with negative class samples from 3D domain, as well as classification loss pertaining to each modality, thus projecting the modality invariant features in the same shared space. Additionally, inspired from information theoretic approaches, we include an entropy minimization term over the probabilities of the target domain features, so as to make the class embeddings more discriminative even in absence of groundtruth labels.

Our specific contributions are enlisted as follows:

- We propose a ***pioneer work*** of image based shape retrieval (IBSR) across non-overlapping domains. This means that during training stage, we have labels and mapped query and shape information from one distribution, and we have to transfer this model for shape retrieval in another distribution for the retrieval task.
- To this end, we follow a transductive approach by using the unpaired and unlabeled images and 3D shapes from the target domain during training. We use this information to mitigate the domain differences between the source and target domains using an MMD loss.
- Additionally, we also incorporate target level feature refinement by introducing an entropy loss over the target probabilities to ensure better intraclass compactness among the target features.
- For modality alignment, we employ a synergistic combination of *cross-modal* triplet loss [35] and classification loss. The former brings the class embedding of samples of same class from 2D and 3D domains closer and increases their distance otherwise. The latter separately aligns the features of individual modality towards it corresponding label, which is shared among the two modalities.
- We have conducted several experiments on Pix3D, SHREC'14 and ShapeNet datasets, both with and without domain adaptation, and it is clearly visible that in absence of cross-domain training, the results are subpar as compared to our method.

2. Related Work

In this section, we discuss the existing literature works pertaining to 3D shape retrieval and domain adaptation.

2.1. 3D Shape Retrieval

3D shape retrieval has been a widely studied topic, with various approaches proposed to address the challenges in this domain. Shape retrieval is one of the most fundamental problems in computer vision. With recent development in deep learning techniques for feature extraction and 3D shape datasets, 3D shape retrieval from single images based shape retrieval (IBSR) has gained more attention. Mu et al. [36] proposed a novel architecture that maps two kinds of features into high-dimensional Hilbert space to decrease the gap. Deep cross-modality adaptation (DCA) employs a metric learning-based method to learn domain discriminative features and cross-modal transformation network to transfer the features of the 2D sketch to the 3D shape feature space. [42] proposed the unsupervised dual-level embedding alignment (DLEA) network, which was a first end-to-end network for this task. The gap between the two modalities is reduced by alignment at the embedding on domain and class levels. Recently, CDA [16] introduces a joint domain-class alignment module to learn a class-discriminative and domain-agnostic feature space for 2D images and 3D models. [43] proposes to learn discriminative and transferable cross-domain representation for 2D and 3D data using unsupervised adversarial domain adaptation. Despite significant prior works, using single images to retrieve the 3D shapes is still a challenging problem, and the major reason for this challenge is the problem of domain shift in both 2D and 3D features space. Solution to this problem is Domain Adaptation (DA) [2, 6, 40]. For our work, we will be focusing on Unsupervised Domain adaptation techniques.

2.2. Domain Adaptation

Domain adaptation refers to the process of adjusting a machine learning model that has been trained on one dataset (the source domain) to perform well on a different dataset (the target domain) where the data distributions may vary. The existing literature offers a wide range of approaches for domain adaptation, including techniques such as subspace alignment, pseudo-labeling, and adversarial methods, among others [20]. The idea of DA is not only limited to primary vision domain, but also extends to other derived applications, such as remote sensing [24] and medical imaging [14]. Another commonly used approach is Domain-Adversarial Neural Networks (DANN)[9], which involves incorporating a domain classifier into the deep neural network to enable it to differentiate between source and target domain data. [30] introduced the notion of generalised adversarial learning in the domain adaptation framework in discriminative setting. The approach presents a step-wise adaptation framework, where a pretrained network is taught to adapt to the target domain through adversarial learning.

CyCADA[15] is an approach that utilizes cycle-consistent adversarial learning to align the feature distributions of the source and target domains. CDTrans[38] employs cross-attention and two-way center-aware labeling in Transformers[31] to achieve domain alignment, making it robust against noisy label pairs. [32] proposes a novel method of domain adaptation based on contrastive learning with pseudo-labels. Here, given an anchor image from source domain, the distance of the target sample with same pseudo-label as that of anchor is minimised with the anchor sample, while pseudo sample generation follows a clustering based approach. Complementary to adversarial learning, there are several metric learning approaches to address UDA. [18] proposed an optimal transport based approach of UDA, where Wasserstein distance metric was used to bridge the gap between the source and the target domains. [41] incorporated a modified form of triplet loss in the UDA along with the notion of penalising the distance of second highest probability from the decision boundary. Recently, there has been interest in exploring vision-language models to tackle the domain adaptation task, given their improved feature space. The current method in this domain, DAPL[10], relies on ad-hoc prompting to learn disentangled domain and category representations.

2.3. How are we different?

The novelty of our work lies in tackling the problem of image based 3D shape retrieval in cross-domain setting, i.e. the annotated and paired dataset (primarily used for inductive training) is vastly different in distribution aspect than the test data. This makes our work a foundation research in this domain. To this end, we have segregated the the cross-domain retrieval task into multiple smaller sub-tasks, and addressed them accordingly. We have mitigated the domain differences between the participating sketch and image domains (from 2D modalities) and the corresponding 3D shapes using maximum mean discrepancy (MMD) loss. To tackle the task of modality alignment between shapes and images, we have incorporated supervised learning, where the paired source domain features of the two modalities are trained to classify the respective samples in the same class. Additionally, to further increase the inter-class difference, we incorporate triplet loss, where we harness the notion of negative sample mining and increase the distance between samples of dissimilar classes. Finally, to bolster the intraclass compactness among the target samples, we incorporate entropy among their class probabilities. In the next section, we will give the description of our proposed method along with the relevant mathematical background.

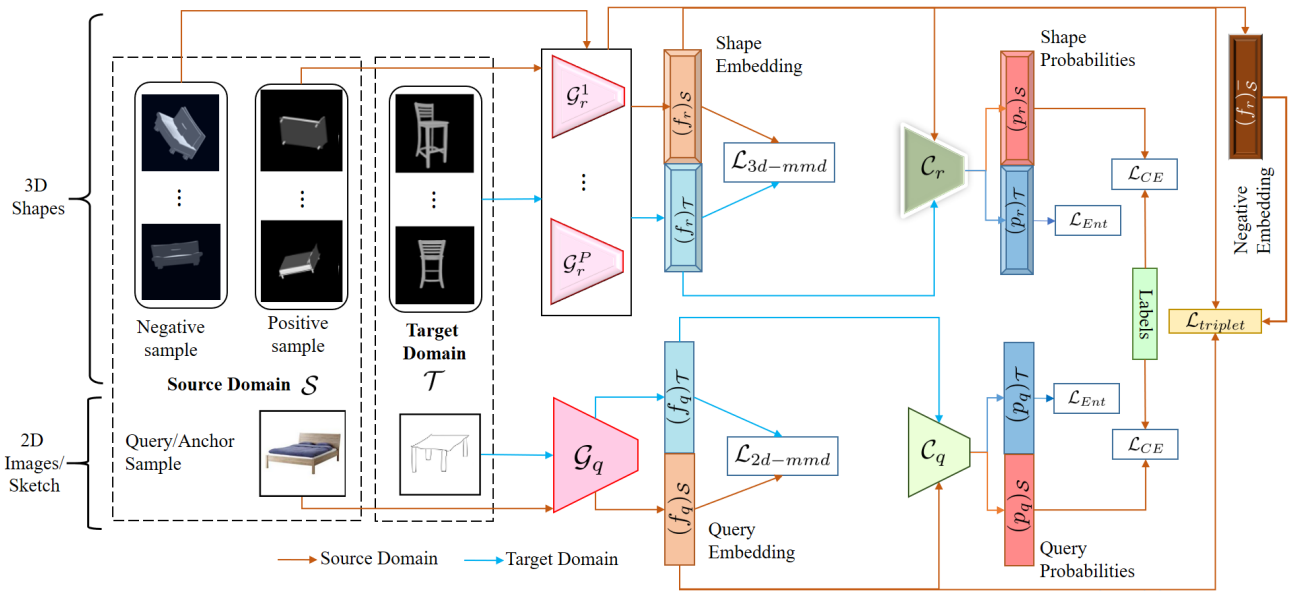

Figure 2. Schematic of proposed DAIS-NET from cross-domain 3D shape retrieval. Our model addresses the different aspects of modality alignment and domain alignment in transductive setting to perform shape retrieval from target domain, when trained with labels and pairs from source domain and unpaired and unlabelled samples from target domain. Initially, we pass the 2D images (from $\mathcal{S}$ and $\mathcal{T}$) through the feature extractors $\mathcal{G}_q$, thus getting embedding $(f_q)_\mathcal{S}$ and $(f_q)_\mathcal{T}$. Simultaneously, 3D shapes are passed through $\mathcal{G}_r$ to get embeddings $(f_r)_\mathcal{S}$, $(f_r)_\mathcal{T}$ and $(f_r)_\mathcal{S}^-$ (the last one being the embedding of the negative sample). For modality alignment, we primarily rely on the triplet loss, which minimizes the Euclidean distance between $(f_q)_\mathcal{S}$ and $(f_r)_\mathcal{S}$, while maximizes the same between $(f_q)_\mathcal{S}$ and $(f_r)_\mathcal{S}^-$. Additionally, the features $(f_q)_\mathcal{S}$ and $(f_r)_\mathcal{S}$ are passed through the classifiers $\mathcal{C}_q$ and $\mathcal{C}_r$ to get the corresponding probabilities, which are compared against the available groundtruth. For domain alignment, we minimize the maximum mean discrepancy loss between the 2D ($(f_q)_\mathcal{S}$ and $(f_q)_\mathcal{T}$) and 3D ($(f_r)_\mathcal{S}$ and $(f_r)_\mathcal{T}$) embeddings of the two domains each. Furthermore, for better intraclass discriminativeness in $\mathcal{T}$, we employ an entropy loss for both the modalities.

3. DAIS-NET for IBSR

The details and mathematical background of DAIS-NET will be discussed in this section. The overall model can be seen in Fig. 2.

3.1. Preliminaries

We are given the dataset $\mathcal{D} = \{\mathcal{S}, \mathcal{T}\}$ comprising the source domain $\mathcal{S}$ and the target domain $\mathcal{T}$, such that, $\mathcal{S} = \{\mathbf{s}_I^j, \mathbf{s}_M^j, y_\mathcal{S}^j\}_{j=1}^{n_\mathcal{S}}$ and $\mathcal{T} = \{\mathbf{t}_I^k, \mathbf{t}_M^k, y_\mathcal{T}^k\}_{k=1}^{n_\mathcal{T}}$ Here, we have the input image pairs, denoted as $\mathbf{s}_I^j$ and $\mathbf{t}_I^j$, in the respective domains and the corresponding 3D shapes are represented as $\mathbf{s}_M^j$ and $\mathbf{t}_M^j$, respectively. Since, the problem setting is of domain adaptation, the The dimensions of the image pairs are given by $\{\mathbf{s}_I^j, \mathbf{t}_I^j\} \in \mathbb{R}^{M \times N \times 3}$, where M and N refer to the spatial dimensions of the three-channel RGB images, while the 3D shapes are represented by $\{\mathbf{s}_M^j, \mathbf{t}_M^j\} \in \mathbb{R}^{P \times M \times N \times 3}$, with P denoting the number of image frames used to represent a 3D object or shape. Additionally, $y_\mathcal{S}^j$ and $y_\mathcal{T}^j$ depict the labels associated with the source and target domains, respectively. The variables $n_\mathcal{S}$ and $n_\mathcal{T}$ indicate the number of training pairs available for the source and target domains.

3.2. Problem Definition

Given the image samples $\mathbf{s}_I^j$ and corresponding class labels $y_\mathcal{S}^j$, our objective is to retrieve the 3D shapes from the provided dataset. The scope of this research extends to a cross-domain setting, wherein the retrieval model is primarily trained on the data originating from the source domain $\mathcal{S}$ and subsequently deployed in another domain denoted as $\mathcal{T}$. To address the challenge of domain discrepancy, we adopt a transductive learning approach, leveraging unpaired and unlabeled query 2D images and 3D shapes from the target domain during the training phase to facilitate domain alignment.

3.3. Modality Specific Feature Extractors

We define separate feature extractors for the query images and 3D shapes, given as $\mathcal{G}_q$ and $\mathcal{G}_r$ respectively. For 3D shapes, owing to the multiple number of frames, $\mathcal{G}_r$ is based on a multiview CNN [27]. Being the transductive setting, the samples from both $\mathcal{S}$ and $\mathcal{T}$ are passed through them to give the corresponding embedding vectors, given as:

$$(f_q^j)_{\mathcal{S}} = \mathcal{G}_q(\mathbf{s}_I^j), (f_q^j)_{\mathcal{T}} = \mathcal{G}_q(\mathbf{t}_I^j) \quad (1)$$

$$(f_r^j)_{\mathcal{S}} = \mathcal{G}_r(\mathbf{s}_M^j), (f_r^j)_{\mathcal{T}} = \mathcal{G}_r(\mathbf{t}_M^j) \quad (2)$$

The embeddings are then sent to the subsequent modules for further processing.

3.4. Cross-Domain Alignment

In order to bring the source domain closer to the target domain, maximum mean discrepancy (MMD) [13] loss is considered. MMD serves as a powerful non-parametric measure to assess the similarity between distributions using two distinct datasets. The equations for the same are seen below.

$$\mathcal{L}^2_{mmd_{2D}}(\mathcal{S}, \mathcal{T}) = \Big|\Big| \frac{1}{b_1}\phi((f_q^j)_{\mathcal{S}}) - \frac{1}{b_2}\phi((f_q^j)_{\mathcal{T}}) \Big|\Big|^2_{\mathcal{H}} \quad (3)$$

$$\mathcal{L}^2_{mmd_{3D}}(\mathcal{S}, \mathcal{T}) = \Big|\Big| \frac{1}{b_1}\phi((f_r^j)_{\mathcal{S}}) - \frac{1}{b_2}\phi((f_r^j)^k_{\mathcal{T}}) \Big|\Big|^2_{\mathcal{H}} \quad (4)$$

Here, $\phi(.)$ represents the kernel function, which is assumed to be a Gaussian radial basis function (RBF). The kernel projects the features to $\mathcal{H}$, the reproducing kernel Hilbert space (RKHS).

3.5. Modality Alignment using Negative Samples

The major challenge in IBSR is the alignment of the participating modalities. To this end, we have leveraged the idea of triplet loss, where we not only bring the samples from 2D and 3D modalities of the two same classes closer to each other, but also use negative samples (samples from different class) from 3D modality and increase its distance from the 2D sample. Initially, the negative sample (denoted as $(\mathbf{s}_M^j)^-$). We first pass this through the feature extractor $\mathcal{G}_r$, and get the embedding $(f_r^j)^-$. The triplet loss is thus represented as:

$$\mathcal{L}_{triplet} = \sum_{i=1}^{n_{\mathcal{S}}} max\Big\{ \Big(\Big|\Big| (f_q^j)_{\mathcal{S}} - (f_r^j)_{\mathcal{S}} \Big|\Big|_2^2 - \Big|\Big| (f_q^j)_{\mathcal{S}} - (f_r^j)^-_{\mathcal{S}} \Big|\Big|_2^2 + \alpha \Big), 0 \Big\} \quad (5)$$

Here, α is the margin term in the triplet loss, which defines the threshold difference between the distances between positive and negative pairs.

3.6. Modality Specific Classification Module

The problem of shape retrieval requires the query image to retrieve the shape of the corresponding class. Thus, to ascertain that the feature learning at source level is discriminative enough at class level, we incorporate a shared classifier at domain level for the two modalities ($\mathcal{C}_q$ and $\mathcal{C}_r$) respectively for query and shape embeddings. The classifiers output the softmax probabilities, for features corresponding to query image and retrieved shape for each domain given as $(p_q^j)_{\mathcal{S}}$ and $(p_r^j)_{\mathcal{S}}$, and $(p_q^j)_{\mathcal{T}}$ and $(p_r^j)_{\mathcal{T}}$ in Eq. 6 and 7.

$$(p_q^j)_{\mathcal{S}} = \mathcal{C}_q((f_q^j)_{\mathcal{S}}), (p_q^j)_{\mathcal{T}} = \mathcal{C}_q((f_q^j)_{\mathcal{T}}) \quad (6)$$

$$(p_r^j)_{\mathcal{S}} = \mathcal{C}_r((f_r^j)_{\mathcal{S}}), (p_r^j)_{\mathcal{T}} = \mathcal{C}_r((f_r^j)_{\mathcal{T}}) \quad (7)$$

The modality specific softmax probabilities from source domain are subjected to cross-entropy loss against the available groundtruth samples, while those corresponding to target domain are used for unsupervised feature refinement (see section 3.6.1).

$$\mathcal{L}_{CE_q} = -\sum_{j}^{n_{\mathcal{S}}} y^j \log(p_q^j)_{\mathcal{S}} \quad (8)$$

$$\mathcal{L}_{CE_r} = -\sum_{j}^{n_{\mathcal{S}}} y^j \log(p_r^j)_{\mathcal{S}} \quad (9)$$

3.6.1 Entropy guided feature extraction

In this research work, the problem of DA is addressed in a transductive setting. In such a case, we have no label information corresponding to the target domain. Thus, in order to the induce the discriminativeness within the sample, we introduce the idea of entropy minimization in unsupervised setting, for both the modalities in the target domain. This forces the features of samples of similar class to be near to each other in the shared space, as shown in:

$$\mathcal{L}_{E_q} = -\sum_{j}^{n_{\mathcal{T}}} (p_q^j)_{\mathcal{T}} \log(p_q^j)_{\mathcal{T}} \quad (10)$$

$$\mathcal{L}_{E_r} = -\sum_{j}^{n_{\mathcal{T}}} (p_q^j)_{\mathcal{T}} \log(p_r^j)_{\mathcal{T}} \quad (11)$$

3.7. Training and Inference

The model is trained on the combined loss function given as:

$$\mathcal{L}_{final} = (\lambda_{2D}\mathcal{L}_{mmd_{2D}} + \lambda_{3D}\mathcal{L}_{mmd_{3D}}) + (\lambda_{c_q}\mathcal{L}_{c_q} + \lambda_{c_r}\mathcal{L}_{c_r}) + (\lambda_{triplet}\mathcal{L}_{triplet}) + (\lambda_{E_q}\mathcal{L}_{E_q} + \lambda_{E_r}\mathcal{L}_{E_r}) \quad (12)$$

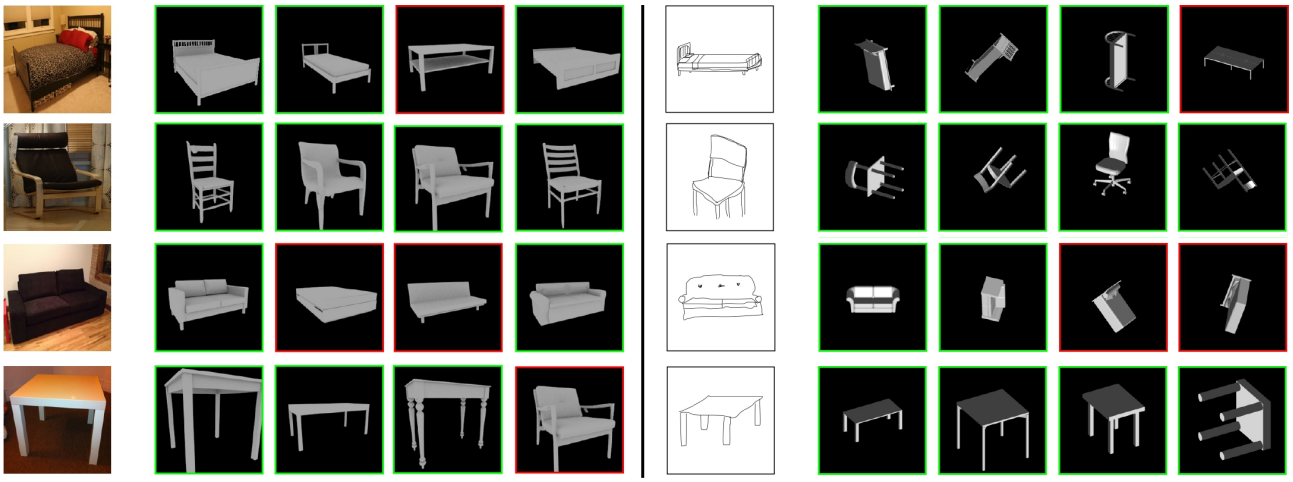

Figure 3. Cross domain retrieval of 3D shapes from (a) Pix3D dataset (when trained on SHREC'14 as source) and (b) SHREC'14 dataset (when trained on Pix3D as source).

Here, all the λ terms represent the Lagrangian multipliers for the corresponding loss functions.

After the training is complete, we take the corresponding feature extractors for the 2D and 3D modalities from the target set $\mathcal{T}$, and pass the query images and shapes from the retrieval dataset through them. Then the similarity between the embedding of query and the shapes is calculated using L2 norm, and the images are retrieved in the ascending order of the magnitude of the L2-norm. In order to assess the performance of shape retrieval, we use mean average precision (mAP) [19], calculated over K number of retrieved samples.

4. Experiments

In this section, we discuss the datasets and experimental protocols followed in the research work.

4.1. Dataset

We used three different benchmark datasets for our experimentation, SHREC'14[12], Pix3D[28] and ShapeNet[3]. SHREC'14 contains 13,680 hand-drawn sketches and 8,987 3D shapes. This dataset also contains 80 sketches for each category, 50 for training and 30 for testing. We combined both the split into one set. Pix3D consists of 10,069 real-world images and 395 unique 3D models. ShapeNet consists of 51,300 3D shape models, the query images for ShapeNet were taken from ImageNet[7], which consists of 1.3 million images across 1000 categories. We chose 4 common classes across these 4 datasets for our experiments, viz. bed, chair, sofa and table.

4.2. Training Protocols

The problem of domain adaptation has been addressed by creating pairs of SHREC'14 and Pix3D, and SHREC'14 and ShapeNet datasets. For both the dataset pairs, one of the dataset is considered as the source domain while the other is considered as target domain, thus giving 4 dataset pairs to test our methods. In order to ascertain the effectiveness of our method, in the proposed architecture, the initial feature extraction is carried out using pretrained CNN based vision models, which are finetuned while training. Specifically, for shape features, we rely on multi-view CNN architecture. We primarily try two backbones, namely ResNet18 and ResNet34, both for query and shape features for all the datasets. During training and inference, all the images are resized to 224×224 before feeding into the models. The training is carried out on Nvidia DGX 80GB with the Adam optimizer, with an initial learning rate of 0.0001. The number of training epochs are fixed to 30 for all the datasets. Additionally, in the final loss function, the value of λ_{E_q} and λ_{E_r} is set to 0.1, while all the other λ values are set to 1.

5. Results and Discussions

Table 1. mAP analysis using our proposed approach on SHREC'14 and Pix3D datasets. In the experiments, we have alternately used SHREC'14 and Pix3D as source and target domains.

Method	SHREC → Pix3D	Pix3D → SHREC
IBSR [23]	0.14	0.24
DD-GAN [37]	0.42	0.25
DA - MSE Loss (ResNet18)	0.34	0.22
DA - Triplet Loss (ResNet18)	**0.70**	**0.36**
DA - MSE Loss (ResNet34)	0.45	0.30
DA - Triplet Loss (ResNet34)	**0.75**	**0.42**

Table 2. mAP analysis using our proposed approach on SHREC'14 and ShapeNet datasets. In the experiments, we have alternately used SHREC'14 and ShapeNet as source and target domains. The correctly retrieved object are placed in a green box while the incorrect ones are placed in the red box.

Method	SHREC → ShapeNet	ShapeNet → SHREC
IBSR [23]	0.27	0.24
DD-GAN [37]	0.40	0.27
DA - MSE Loss (ResNet18)	0.28	0.30
DA - Triplet Loss (ResNet18)	**0.62**	**0.37**
DA - MSE Loss (ResNet34)	0.30	0.33
DA - Triplet Loss (ResNet34)	**0.64**	**0.41**

The results of our proposed approach are presented in Tables 1 and 2 respectively for SHREC'14/Pix3D and SHREC'14/ShapeNet datasets. The first two rows in both the tables show the results of cross-domain retrieval using methods that were trained on single domain setting (IBSR [23] and DD-GAN [37]). We find that DD-GAN outperforms IBSR owing to the effort to generate modality invariant features by the former, while IBSR only focuses on color invariant features. However, when compared against DAIS-NET, it is visible that the mAP is much better in comparison to when DA is not applied. Furthermore, it is also visible in both the tables, that the performance is better when ResNet34 is used as a feature extractor instead of ResNet18. This is self-evident from the fact that ResNet34 has richer feature representation of the two and finetuning it does not degrades its performance in retrieval task. In addition, to highlight the effectiveness of triplet loss for modality alignment, we have compared it against mean squared error loss (MSE). It is clearly visible that triplet loss outperforms the MSE loss by a large margin, in both the dataset pairs and for both ResNet18 and ResNet34 feature extractors.

In addition, it is also visible in both the tables that when SHREC dataset (consisting of sketches) is chosen as source domain, the retrieval performance on Pix3D and ShapeNet (which are image based datasets) is much better in contrast to the case when Pix3D or ShapeNet chosen as source domain and test on SHREC. This is due to the fact that CNN based feature extractors inherently suffer from a shape bias [11]. This means, when a model that is trained on sketches is tested for shape retrieval using images, it would require to match the shape and outlines of the object only. However, in the reverse case, the model is trained to expect the texture and colour information as well, which is missing in the sketches. This would make it difficult for the model to transfer the knowledge from image to sketch domain for shape retrieval.

Moreover, we have also shown the qualitative performance of shape retrieval on top-4 objects DAIS-NET in Fig. 3 on Pix3D and SHREC'14 dataset, alternately taken as source and target domains. The green box shows that the object is retrieved from the same class as that of the query image/sketch, while the red box shows otherwise. We see that in most of the cases, the objects are correctly retrieved. However, there are still a few instances of incorrectly retrieved objects, owing to the problem of domain difference in the shape retrieval problems.

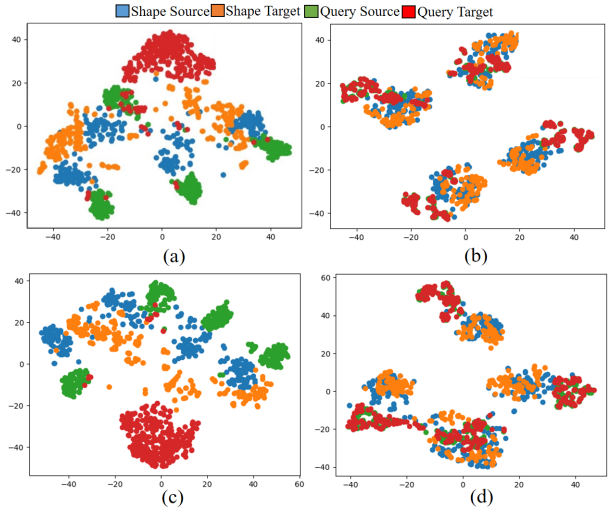

Figure 4. TSNE plots when the training is carried out (a) ShapeNet → Pix3D without DA (b) ShapeNet → Pix3D With DA (c) Pix3D → ShapeNet without DA (d) Pix3D → ShapeNet with DA

To visualize the discrimination ability by our proposed approach, we have presented the TSNE plot showing the retrieval results without DA and with DA. In the Fig. 4 (a) and (b), we have considered that domain $\mathcal{S}$ is sketch-shape pair, while domain $\mathcal{T}$ is image-shape pair, from pix3D and SHREC datasets. The vice versa case is demonstrated in Fig. 4 (c) and (d). We can clearly see that when DA is applied (Fig. 4 (b) and (d)), we end up with more refine clusters, where the same classes from both the domains and modalities overlap and 4 distinct clusters corresponding to each class are created. However, when DA is not applied, the clusters corresponding to same classes for same modality do not overlap (Fig. 4 (a) and (c)), which shows the significance of domain differences and the need of applying domain adaptation. If we compare Fig. 4 (b) and (d), we can also observe that the clusters are more well-defined in the the former than the latter. This could be attributed to the fact that when training is carried out with sketches as queries, then the evaluation with images as queries is easier, instead of when the order is reversed. This is because in the former case, the model is able to learn the shape information from the sketches, which could be deciphered in case of images. However, in the latter case, the model learns the colour and texture information as well from the images for retrieval task, which is entirely absent in the images, leading to poor discrimination and thus, overlapping clusters.

Table 3. Ablation study over the different combination of domain adaptive and modality alignment losses in DAIS-NET on SHREC'14→Pix3D dataset pair.

Loss combination	mAP
$\mathcal{L}_{mmd_{2D}} + \mathcal{L}_{mmd_{3D}} + \mathcal{L}_{CE}$	0.47
$\mathcal{L}_{mmd_{3D}} + \mathcal{L}_{triplet} + \mathcal{L}_{CE}$	0.48
$\mathcal{L}_{mmd_{2D}} + \mathcal{L}_{triplet} + \mathcal{L}_{CE}$	0.57
$\mathcal{L}_{mmd_{2D}} + \mathcal{L}_{mmd_{3D}} + \mathcal{L}_{triplet}$	0.66
$\mathcal{L}_{mmd_{2D}} + \mathcal{L}_{mmd_{3D}} + \mathcal{L}_{triplet} + \mathcal{L}_{CE}$	0.71
$\mathcal{L}_{mmd_{2D}} + \mathcal{L}_{mmd_{3D}} + \mathcal{L}_{triplet} + \mathcal{L}_{CE} + \mathcal{L}_{E}$	**0.75**

We have performed an ablation study on the losses, the ablation study of which is shown in Table 3 for SHREC'14 and Pix3D dataset (with the former one being the source domain) with ResNet34 as the backbone model. For simplicity, the entropy and classification losses for both the modalities are represented as $\mathcal{L}_{CE}$ and $\mathcal{L}_{E}$. It is clearly visible that when all the losses are involved, mAP is the highest (0.75), thus proving that all the losses complement each other and contribute together in learning a shared cross-modal and cross-domain representation. Additionally, we also observe the contribution of entropy based loss over the target domain, which gives and improvement of 0.04 (0.71 $\rightarrow$ 0.75), than when its is not used. Moreover, we get a very low mAP performance in the first row (0.47), when neither of triplet loss and entropy are used, which further prove their contribution in our approach. Furthermore, from the second and third rows, we see that the performance is high when MMD loss is applied only in the 2D modality (query) than in the 3D modality, which shows that domain alignment is a more serious issue in the former. From the fourth row, we can see that when classification losses are removed, the mAP loss decreases which shows their effectiveness in modality alignment.

Table 4. Number of parameters for the proposed and existing methods. 'M' denotes 'millions'.

Method	IBSR	DDGAN	Ours (ResNet18)	Ours (ResNet34)
Number of parameters	35.06 M	33.43 M	**28.92 M**	39.03 M

In Table 4, we compare the computation complexity of our methods against the existing methods of shape retrieval. It is visible that even with 28.92 million parameters on ResNet18 based model, our method has the ability to surpass the existing methods in the task of image and sketch based shape retrieval.

6. Conclusions and Future Scope

In this work, we propose a pioneering approach of multimodal domain adaptation for image-based 3D shape retrieval. Our method tackles the challenges of aligning two modalities (2D images/sketches and 3D shapes) and aligning disjoint domains containing the 2D queries and 3D shapes. To address the modality alignment challenge, we introduce the notion of negative sample mining and in the source domain and create the negative samples from 3D shapes from the classes different from that of the query image. Then, we employ the triplet loss where the distance between the samples of same class is minimised and that of two different classes is maximised between 2D image and 3D shape. Furthermore, we incorporate modality-specific classifiers to align the 3D shapes and 2D images in the source domain. To address domain alignment, we utilize the maximum mean discrepancy for each modality across the two domains. Additionally, we introduce an entropy loss to establish discriminativeness among the target features based on class probabilities of the target embeddings for both modalities. Our model is evaluated on two pairs of image-shape datasets, namely SHREC'14-Pix3D and SHREC'14-ShapeNet. The experimental results demonstrate the presence of the domain adaptation problem, and our proposed approach effectively addresses it, as evidenced by both qualitative and quantitative evaluations. As future work, we plan to extend our approach to zero-shot domain adaptation and explore domain generalization techniques to handle multiple domains simultaneously.

References

[1] Gholamali Aminian, Mahed Abroshan, Mohammad Mahdi Khalili, Laura Toni, and Miguel Rodrigues. An information-theoretical approach to semi-supervised learning under covariate-shift. In *International Conference on Artificial Intelligence and Statistics*, pages 7433–7449. PMLR, 2022.

[2] Mahsa Baktashmotlagh, Mehrtash T Harandi, Brian C Lovell, and Mathieu Salzmann. Unsupervised domain adaptation by domain invariant projection. In *Proceedings of the IEEE international conference on computer vision*, pages 769–776, 2013.

[3] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[4] Ushasi Chaudhuri, Biplab Banerjee, Avik Bhattacharya, and Mihai Datcu. CrossATNet-A novel cross-attention based framework for sketch-based image retrieval. *Image and Vision Computing*, 104:104003, 2020.

[5] Guoxian Dai, Jin Xie, Fan Zhu, and Yi Fang. Deep correlated metric learning for sketch-based 3D shape retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[6] Hal Daumé III, Abhishek Kumar, and Avishek Saha. Frustratingly easy semi-supervised domain adaptation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 53–59, 2010.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image

database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[8] Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R Arabnia. A brief review of domain adaptation. *Advances in Data Science and Information Engineering: Proceedings from ICDATA 2020 and IKE 2020*, pages 877–894, 2021.

[9] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[10] Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. Domain adaptation via prompt learning. *arXiv preprint arXiv:2202.06687*, 2022.

[11] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

[12] Afzal Godil and Chunyuan Li. SHREC'14 track: Extended large scale sketch-based 3d shape retrieval. The Seventh Eurographics Workshop on 3D Object Retrieval (3DOR 2014)., Strasbourg, -1, 2014-06-12 2014.

[13] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[14] Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2021.

[15] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. Pmlr, 2018.

[16] Nian Hu, Heyu Zhou, An-An Liu, Xiangdong Huang, Shenyuan Zhang, Guoqing Jin, Junbo Guo, and Xuanya Li. Collaborative distribution alignment for 2D image-based 3D shape retrieval. *Journal of Visual Communication and Image Representation*, 83:103426, 2022.

[17] Donna Karolchik, Angela S Hinrichs, Terrence S Furey, Krishna M Roskin, Charles W Sugnet, David Haussler, and W James Kent. The ucsc table browser data retrieval tool. *Nucleic acids research*, 32(suppl_1):D493–D496, 2004.

[18] Tanguy Kerdoncuff, Rémi Emonet, and Marc Sebban. Metric learning in optimal transport for domain adaptation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 2162–2168, 2021.

[19] Kazuaki Kishida. *Property of average precision and its generalization: An examination of evaluation indicator for information retrieval experiments*. National Institute of Informatics Tokyo, Japan, 2005.

[20] Wouter M Kouw and Marco Loog. A review of domain adaptation without target labels. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):766–785, 2019.

[21] Guangrui Li, Guoliang Kang, Yi Zhu, Yunchao Wei, and Yi Yang. Domain consensus clustering for universal domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9757–9766, 2021.

[22] Wenhui Li, Anan Liu, Weizhi Nie, Dan Song, Yuqian Li, Zjenja Doubrovski, Jo Geraedts, Zishun Liu, Yunsheng Ma, et al. Shrec 2019-monocular image based 3d model retrieval. In *Proc. 12th Eurographics Workshop 3D Object Retr.*, pages 1–8, 2019.

[23] Ming-Xian Lin, Jie Yang, He Wang, Yu-Kun Lai, Rongfei Jia, Binqiang Zhao, and Lin Gao. Single image 3d shape retrieval via cross-modal instance and category contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11405–11415, 2021.

[24] Shivam Pande, Biplab Banerjee, and Aleksandra Pižurica. Class reconstruction driven adversarial domain adaptation for hyperspectral image classification. In *Pattern Recognition and Image Analysis: 9th Iberian Conference, IbPRIA 2019, Madrid, Spain, July 1–4, 2019, Proceedings, Part I 9*, pages 472–484. Springer, 2019.

[25] Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. Stylemeup: Towards style-agnostic sketch-based image retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8504–8513, 2021.

[26] Yuan Shi and Fei Sha. Information-theoretical learning of discriminative clusters for unsupervised domain adaptation. *arXiv preprint arXiv:1206.6438*, 2012.

[27] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 945–953, 2015.

[28] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3d: Dataset and methods for single-image 3D shape modeling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[29] Hui Tang, Ke Chen, and Kui Jia. Unsupervised domain adaptation via structurally regularized deep clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8725–8735, 2020.

[30] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[32] Rui Wang, Zuxuan Wu, Zejia Weng, Jingjing Chen, Guo-Jun Qi, and Yu-Gang Jiang. Cross-domain contrastive learning for unsupervised domain adaptation. *IEEE Transactions on Multimedia*, 2022.

[33] Shanshan Wang, Lei Zhang, Pichao Wang, MengZhu Wang, and Xingyi Zhang. Bp-triplet net for unsupervised domain adaptation: A bayesian perspective. *Pattern Recognition*, 133:108993, 2023.

[34] Wei Wang, Haojie Li, Zhengming Ding, Feiping Nie, Junyang Chen, Xiao Dong, and Zhihui Wang. Rethinking maximum mean discrepancy for visual domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[35] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research*, 10(2), 2009.

[36] Zizhao Wu, Yunhui Zhang, Ming Zeng, Feiwei Qin, and Yigang Wang. Joint analysis of shapes and images via deep domain adaptation. *Computers & Graphics*, 70:140–147, 2018.

[37] Rui Xu, Zongyan Han, Le Hui, Jianjun Qian, and Jin Xie. Domain disentangled generative adversarial network for zero-shot sketch-based 3d shape retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2902–2910, 2022.

[38] Tongkun Xu, Weihua Chen, Pichao Wang, Fan Wang, Hao Li, and Rong Jin. CDTrans: Cross-domain transformer for unsupervised domain adaptation. *arXiv preprint arXiv:2109.06165*, 2021.

[39] Fan Yang, Yang Wu, Zheng Wang, Xiang Li, Sakriani Sakti, and Satoshi Nakamura. Instance-level heterogeneous domain adaptation for limited-labeled sketch-to-photo retrieval. *IEEE Transactions on Multimedia*, 23:2347–2360, 2020.

[40] Ting Yao, Yingwei Pan, Chong-Wah Ngo, Houqiang Li, and Tao Mei. Semi-supervised domain adaptation with subspace learning for visual recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2142–2150, 2015.

[41] Yueming Yin, Zhen Yang, Haifeng Hu, and Xiaofu Wu. Metric-learning-assisted domain adaptation. *Neurocomputing*, 454:268–279, 2021.

[42] Heyu Zhou, An-An Liu, and Weizhi Nie. Dual-level embedding alignment network for 2D image-based 3D object retrieval. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1667–1675, 2019.

[43] Yaqian Zhou, Yu Liu, Heyu Zhou, Zhiyong Cheng, Xuanya Li, and An-An Liu. Learning transferable and discriminative representations for 2D image-based 3D model retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):7147–7159, 2022.