Multimedia Systems (2023) 29:3891–3901
https://doi.org/10.1007/s00530-023-01166-y

SPECIAL ISSUE PAPER

View-target relation-guided unsupervised 2D image-based 3D model retrieval via transformer

Jiacheng Chang[1] · Lanyong Zhang[1] · Zhuang Shao[2]

Received: 14 April 2023 / Accepted: 12 August 2023 / Published online: 24 August 2023

Abstract

Unsupervised 2D image-based 3D model retrieval aims at retrieving images from the gallery of 3D models by the given 2D images. Despite the encouraging progress made in this task, there are still two significant limitations: (1) feature alignment of 2D images and 3D model gallery is still difficult due to the huge gap between the two modalities. (2) The important view information in the 3D model gallery was ignored by the prior arts, which led to inaccurate results. To alleviate these limitations, inspired by the success of vision transformers (ViT) in a great variety of vision tasks, in this paper, we propose an end-to-end 3D model retrieval architecture on top of ViT, termly transformer-based 3D model retrieval network (T3DRN). In addition, to take advantage of the valuable view information of 3D models, we present an attentive module in T3DRN named shared view-guided attentive module (SVAM) to guide the learning of the alignment features. The proposed method is tested on the challenging dataset, MI3DOR-1. The extensive experimental results have proved the superiority of our proposed method to state-of-the-art methods.

Keywords 3D model retrieval · Feature alignment · Vision transformer · View-guided attentive module

1 Introduction

3D technology is developing rapidly and due to the advancement of computer hardware, 3D models have been widely used in wide-ranging areas, such as 3D reconstruction [1], virtual reality [2], computer-aided medical imaging [3, 4], and 3D object detection [5]. Massive 3D models are generated by these applications, but it also poses great challenges for precise and efficient 3D model retrieval. The goal of 3D model retrieval is that given a query, a system needs to find similar 3D models in another gallery. The query can be diverse, including 2D images of different views, point clouds, etc. Since 2D images are easier to be obtained, the 2D image-based 3D model retrieval has attracted much attention in the computer vision community.

To ensure an effective retrieval performance, many efforts focused on improving the performance by training a powerful feature extractor. One typical earlier pipeline is to train a deep neural network [6–9] with a great number of labeled annotations, but at a cost of huge labor costs. Later on, inspired by the unsupervised domain adaptation (UDA) doctrine [10–12], unsupervised 2D image-based 3D model retrieval utilizes the 2D image data and trains the deep model in an unsupervised manner to transfer the knowledge learned from labeled 2D image source domain to unlabeled 3D model target domain. Specifically, [11] learned a domain-invariant classifier in the Gaussian manifold and aligned the conditional distributions of two domains in a dynamic manner. Zhang et al. [12] mapped the features of the source domain and target domain into a shared latent space to reduce geometric and distributed displacement of statistical measurement. Ganin and Lempitsky [10] used adversarial training domain to align the discrepancy between the source domain and target domain. Yue et al. [13] adopted pixel-level alignment to improve the performance of domain adaptation. Chen et al. [14] proposed a

✉ Zhuang Shao
Zhuang.Shao@warwick.ac.uk

Jiacheng Chang
dlbjhhh@hrbeu.edu.cn

Lanyong Zhang
zhanglanyong@hrbeu.edu.cn

[1] College of Intelligent Systems Science and Engineering, Harbin Engineering University, 145 Nantong Street, Nangang District, Harbin 150001, China

[2] Warwick Manufacturing Group, University of Warwick, IMC Centre, Coventry CV4 7AL, UK

cross-domain adaptation via an image-level mixture methodology to align the distribution shift of features between two domains. Peng et al. [15] concentrated on centroid alignment between features from different image topics and enforced distributed alignment for the already center-aligned features.

Despite the aforementioned encouraging efforts, the performance of unsupervised 2D image-based 3D model retrieval is still far from satisfactory. There are still two drawbacks ignored in the prior arts as follows:

To begin with, for feature representation of both 2D images and 3D models, a better backbone is always encouraged, which draws our attention to the trendy vision transformers (ViT) recently. It has proved to be a success in many relative computer vision and natural language processing (NLP) such as video event detection [16], pedestrian detection [17], person search [18, 19], and text classification [20]. ViT takes the image patch or word embedding as a sequence of tokens, and applies the self-attention mechanism to capture the internal relationships thus obtaining strong feature representation connected with downstream tasks. However, even if the wide application, the application in 2D Image-based 3D Model Retrieval is still under-explored.

Second, the view information of 3D models is valuable but always ignored by the prior works. As shown in Fig. 1, there are 12 views for each 3D model, and the different view information is important indeed during the process of feature alignment. Therefore, in this paper, we try to mine the view information and integrate it into the whole retrieval process.

To tackle these two gaps above, we propose a novel end-to-end 3D model retrieval architecture on top of ViT, dubbed Transformer-based 3D model retrieval network (T3DRN). T3DRN can effectively overcome the limitation of the discrepancy and learn domain-invariant representations. To mine more useful view information of 3D models, we also present an attentive module in T3DRN named shared view-guided attentive module (SVAM) to guide the learning process in order to better align the two modalities. Our main contributions in this paper are threefold as follows:

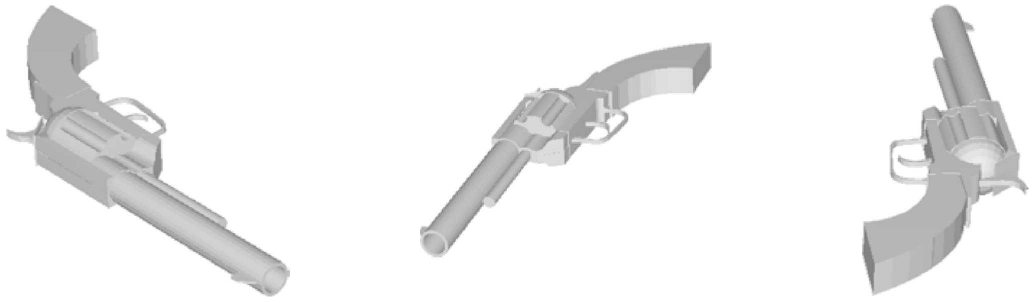
- We propose an end-to-end unsupervised 2D image-based 3D model retrieval framework on top of ViT, dubbed transformer-based 3D model retrieval network (T3DRN) with a distinctive property of mining proper view information to guide the whole retrieval process.
- A novel module, termed shared view-guided attentive module (SVAM), which can be easily integrated into T3DRN, is proposed to attend to the proper view information for the 3D model feature training.
- Qualitative and quantitative experimental results on the challenging unsupervised 2D image-based 3D model retrieval datasets show that our method outperforms the state-of-the-art methods.

The rest of this paper is organized as follows: To begin with, we review the prior works in Sect. 2. Then, in Sect. 3, we present the proposed methodology and expound on the details of T3DRN. The experimental results of our proposed method are demonstrated in Sect. 4 with both qualitative and quantitative analysis. Finally, we draw a conclusion and discuss future work in Sect. 5.

2 Related work

2.1 3D model retrieval

The purpose of 3D model retrieval is to measure the similarity between the query sample and the samples in the dataset and return the relevant 3D model according to the similarity order. The typical 3D model retrieval methods are mainly divided into model-based 3D model retrieval methods and image-based 3D model retrieval methods. Model-based methods usually extract features from 3D formats, such as point clouds and voxels. For example, [21] utilized the supervised convolutional network to encode the binary voxel grids for 3D model representation. Wu et al. [22] represented a model on a 3D voxel grid using probability distributions of



(a) View 3 (b) View5 (c) View10

Fig. 1 Examples of images of 3D models with different view information

binary variables. Qi et al. [23] utilized the neighbor points at multiple scales to capture the local structure information. In view-based methods, the 3D model is usually projected into a set of 2D view images. Su et al. [24] used a CNN to extract the feature of each view individually and then adopted max-pool operation on all view features to obtain a global model feature. Gao et al. [25] utilized patch-level features to learn the content information and spatial information within multiple views. Watanabe et al. [26] composed a compact representation of an image-based shape retrieval system employing a minimum number of views for each model. However, the above methods require a 3D model as the query, which is not convenient for users when they can only obtain an image of the 3D model. Therefore, many researchers seek retrieval approaches via image-based methods.

2.2 Domain adaptation

The object of domain adaptation is to establish knowledge transfer from the source domain to the target domain. Unsupervised domain adaptation, semi-supervised domain adaptation, and supervised domain adaptation are the three common categories of domain adaptation methods. Both source domain data and a few target domain data with labels are used during training in semi-supervised domain adaptation (SSDA) methods while all samples should be marked in supervised domain adaptation methods. Based on previous studies, [27] benefited from a subset of source domain data only containing samples highly relevant to the target domain by eliminating irrelevant source domain samples. To reduce the accuracy discrepancy across domains for better generalizations, invariant representations and invariant optimal predictors were jointly learned in [28]. Unsupervised domain adaptation (UDA) methods seek to narrow the domain gap between the source and target images without any labels of the target images and learning domain-invariant features is a common strategy for unsupervised domain adaptation. For example, [10] added an additional domain discriminator to the conventional CNN network to enforce the domain alignment. Zhou et al. [29] ensured the alignment at the class level by matching the class centroids of the source and target domains based on eliminating the offsets at the domain level. Long et al. [30] utilized the joint maximum mean difference criterion and used multiple domain-specific layers for adaptation. Multi-layer and multi-kernel maximum mean discrepancies were minimized between the source and target domain data to resolve domain transfer issues in [31]. Wang et al. [32] proposed a method that by designing a robust deep neural network, both source and target domains can be transformed to a shared feature space and the classifier trained on the source domain work well on the target domain. Later on, [33] trained a model on new target domains, while maintaining its performance on the previous target domains without forgetting. Hoyer et al. [34] proposed a method that by training with pseudo-labels fused from multiple resolutions and using an overlapping sliding window mechanism, the robustness of fine-grained pseudo-labels with respect to different contexts can be increased. The development of domain adaptation has also facilitated numerous close applications, such as fault detection [35, 36], frequency detection [37], and modulation classification [38, 39].

2.3 Transformer

Transformer was first invented in 2017 in [40]. Later on, a lot of effort was dedicated to its variants. Among them, vision transformer (ViT) [41] was a representative work that proposed transformer stacks for the image classification task. Subsequently, many ViT-related structures have been proposed. In particular, [42] developed a novel ViT backbone with a safe training strategy for UDA. Also, the encoder–decoder framework of Transformer facilitated various vision tasks, such as image captioning [43, 44], relationship or attribute detection [45, 46]. For instance, [47] applied a Transformer-based dense captioning while [48] focused on textual context and higher word learning efficiency. Recently, due to the necessity of encoding spatial information, the Swin Transformer [49] structure was proposed and also applied in many tasks [50]. The application of the transformer-based method for unsupervised 2D image-based 3D model retrieval, however, is still extremely rare. Besides, we also elaborately design a shared view-guided attentive module to better arm the transformer-based architecture.

3 Methodology

The overall framework of our proposed T3DRN is shown in Fig. 2. It is made up of three ViT-block branches with shared parameters on top of [42] and a shared view-guided attentive model(SVAM). In the training period, both 2D image data and 3D model data are sampled. The 2D image data go into the ViT block and be treated as a supervised classification task regularized by cross-entropy loss. The 3D image data follow a contrastive learning regime; it is separated into two branches, the original branch and randomly perturbed branches, respectively. These two kinds of features are also extracted by ViT blocks before going to the SVAM to interact with the view embeddings. These features are regularized

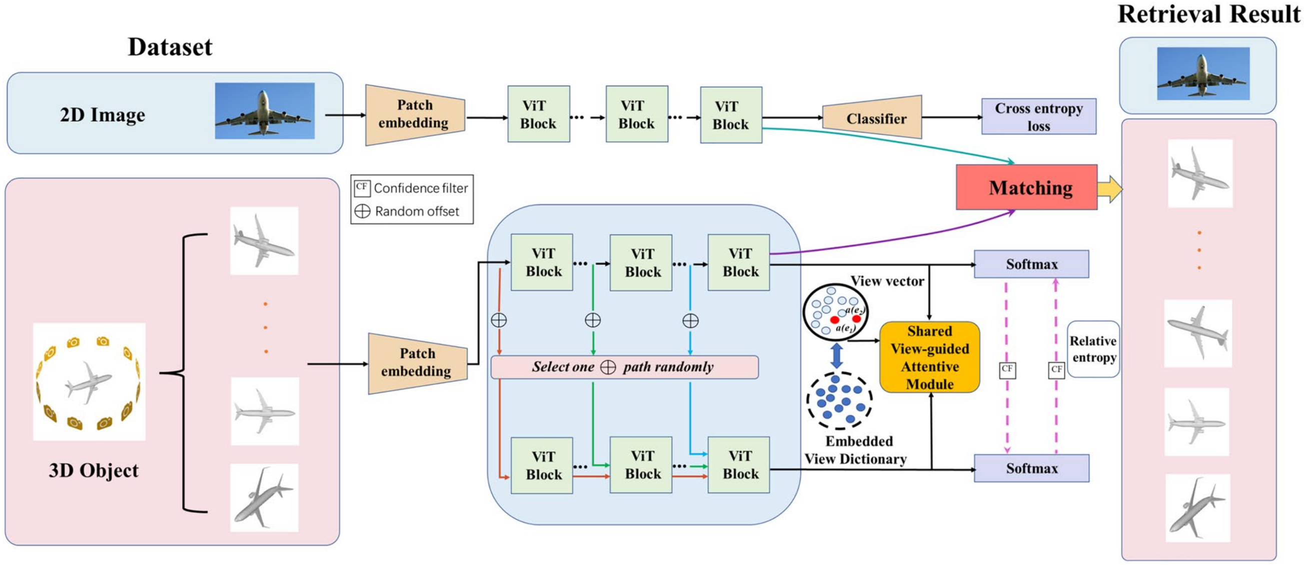


Fig. 2 Our presented T3DRN approach consists of three branches using ViT-blocks that share parameters based on [42], along with a shared view-guided attentive model (SVAM). During the training phase, both 2D image and 3D model data are sampled from the dataset. The 2D image data with labels is input to the ViT block to carry out a supervised classification task, which is regulated by the cross-entropy loss. On the other hand, the 3D image data follows a contrastive learning approach; it is divided into two branches: the original branch and randomly perturbed branches. These types of features are also processed by ViT blocks before being directed to the SVAM, where they interact with the view embeddings. After SVAM, these features are regularized using the relative entropy loss term. Additionally, the adversarial training method described in [10] is employed to align the 2D image features with the 3D model features. During the testing phase, the retrieval results are obtained using the 2D image features and the 3D model features, taken after SVAM but before the final fully connected layer.

by the relative entropy loss term after SVAM. In addition, the adversarial training strategy in [10] is adopted to align the 2D image features and 3D model features. At the test stage, the 2D image features and the 3D model features after SVAM but before the final fully connected layer are taken for retrieval results.

In the following of this section, we will first illustrate the unsupervised 2D image-based 3D model retrieval problem. After this, we introduce our proposed transformer-based 3D model retrieval network T3DRN. Next, we explain our unique shared view-guided attentive module (SVAM). Finally, we show our training and optimization details.

3.1 Problem statement

The goal of unsupervised 2D image-based 3D model retrieval is to create the cross-domain approach capable of precisely finding similar 3D models given a 2D image. In this task, the data of 2D images contain the images and their labels, denoted as $S = \{X^S, Y^S\}$, while the data of the target domain only contain the 3D models without labels, denoted as $T = \{X^t\}$. To learn better domain-invariant features without labels of 3D models, we present transformer-based 3D model retrieval network (T3DRN), which will be introduced in detail in Sect. 3.2.

3.2 Transformer-based 3D model retrieval network

We build up the transformer-based 3D model retrieval network on top of [42]. The T3DRN consists of ViT stacks (shown in Fig. 3) and a shared view-guided attentive model(SVAM), which will be elaborated in Sect. 3.3. T3DRN is composed of three ViT-block branches for 2D images, 3D model images and perturbed 3D model images, respectively. Each ViT block comprises of 4 transformer layers; the implementation of each Transformer layer is as

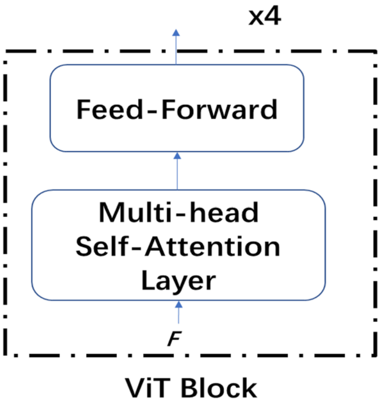


Fig. 3 The detailed structure of ViT Block

follows. For each image either from 2D image gallery or 3D model gallery, the patch embedding layer first transforms it into a token sequence including a special class token and image tokens [42] to get visual features. Then, they are added by positional encoding in [40]. We adopt the positional encoding (PE) procedure with *sin* and *cos* functions.

It should be noted that PE operation only occurs at the bottom of each ViT block. The dimension of PE is the same as the input, so PE embedding can be added directly to the input. After the visual features are added with PE, the output is denoted as F, which is input into three linear projectors to attain three different vectors Q, K, V. These three vectors are fed into the ViT block, the lth layer in a ViT block is given by:

$$\begin{aligned} V\left(F^l\right) &= \varphi(PF(\omega(F^l)), \omega(F^l)), \\ \omega(F^l) &= \begin{pmatrix} \varphi(MA(f_1^l, F^l, F^l), f_1^l) \\ \cdots \\ \varphi(MA(f_T^l, F^l, F^l), f_T^l) \end{pmatrix}, \\ \varphi(\alpha, \beta) &= LayerNorm(\alpha + \beta), \\ PF(\gamma) &= M_2^l \max(0, M_1^l \gamma + b_1^l) + b_2^l, \end{aligned} \tag{1}$$

where φ is layer normalization on residual output, PF represents the feed-forward layer which consists of two linear layers with a nonlinear activation function in between. ω is the output of assembled multi-head attention with a layer normalization by φ. M_1^l and M_2^l are the weights trained for the feed-forward layers, and b_1^l and b_2^l are bias vectors. F^l is the input of the lth encoding layer. f_t^l is given as the query to the encoding layer and l is the lth encoding layer. Note that F^0 is the aforementioned visual feature F added by positional encodings. MA is a fine-grained component called multi-head attention, which is composed of H parallel partial dot-product attention components. Its realization is as follows:

$$\begin{aligned} MA(q_i, K, V) &= \text{concat}\left(h_1, h_2, \ldots, h_H\right) W^O, \\ h_j &= A\left(W_j^q q_i, W_j^K K, W_j^V V\right), \end{aligned} \tag{2}$$

where $\{h_j | j \in [1, H]\}$ refer to the index of each independent head. W_j^q, W_j^K, W_j^V denote the linear projectors to the input q, K, V for h_j. W^O is the weight matrix for each head. It is noted that when the query comes from the decoder layer, and both the keys and values are from the encoder layer, it represents cross-module attention. In contrast, if the queries, keys, and values are all from encoder or decoder, this kind of multi-head attention is named self-attention. A is the scaled dot-product attention operation realized by the equation below.

$$A(q_i, K, V) = V \frac{\exp\left(K^T q_i / \sqrt{d}\right)}{\sum_{t=1}^{T} \exp\left(k_t^T q_i / \sqrt{d}\right)}, \tag{3}$$

where $q_i \in R^d$ is a query in all T queries that composes q_i, a group of keys $k_t \in R^d$ and values $v_t \in R^d$, where $t = 1, 2, \ldots, T$, the output of dot-product attention is the weighted sum of the v_t values. The weights are determined by the dot-products of query q_i and keys k_t. Specifically, k_t and v_t are placed into respective matrices $K = (k_1, \ldots, k_T)$ and $V = (v_1, \ldots, v_T)$. d is the dimension of q_i and $\sqrt{d}$ is to normalize the dot-product value.

In the end, with the output of l encoding layers, the encoded visual features, F^l, is the final output for the ViT blocks. It is noted that we also adopted the random perturbation strategy and the safe training mechanism in [42]. It also consists of F_{2D}^l, F_{3D}^l and F_{3Dpert}^l, the latter two parts are as the input to the SVAM.

3.3 Shared view-guided attentive model

To guide the training process appropriately with view information, we first learn an adaptive view dictionary $E = \{e_1, e_2, \ldots, e_M\}$, where M is the total views in the 3D model dataset. With F_{3D}^l and F_{3Dpert}^l, and E, we design the SVAM as follows:

$$\begin{aligned} SVAM\left(F, E'\right) &= \varphi(PF(\theta(F, E')), \theta(F, E')), \\ \theta(F, E') &= \begin{pmatrix} \varphi\left(MA(e_1', F, F), f_1\right) \\ \cdots \\ \varphi\left(MA(e_T', F, F), f_T\right) \end{pmatrix}, \end{aligned} \tag{4}$$

where F is the input features, in this scenario, it can be either F_{3D}^l or F_{3Dpert}^l. E' is the corresponding embeddings of the view labels for each feature in F. MA, φ and PF are the same with Eq. 1. In this way, we can compute the view-guided features $SVAM(F_{3D}^l, E')$ and $SVAM(F_{3Dpert}^l, E')$, denoted as F_{3D}^v and F_{3Dpert}^v. They will be added with F_{3D}^l and F_{3Dpert}^l as follow for the downstream task:

$$\begin{aligned} F_{3D} &= F_{3D}^l + \lambda F_{3D}^v, \\ F_{3Dpert} &= F_{3Dpert}^l + \lambda F_{3Dpert}^v, \end{aligned} \tag{5}$$

where λ is a balance coefficient between the original features and the view-guided features.

3.4 Training and optimization details

In this section, we show our training and optimization details. In order to enforce the 2D image data to be correctly classified, and the distributions of the representation of the 3D models to be similar with its perturbed counterparts, meanwhile confusing the data between two domains, multiple loss function items are leveraged during the Stochastic Gradient Descent [51] (SGD) at each training step in a training batch as follows:

$$L = L_{cls} + \beta L_{tgt} - L_d, \tag{6}$$

where L_{cls} is the classification binary cross-entropy loss function of the classifier for 2D image data, L_{tgt} is the KL divergence loss in [42], L_d is the domain adversarial loss in [10].

4 Experimental results and discussion

To prove the reliability of our method, we carry out experiments on the most popular dataset, MI3DOR-1 [31]. In this section, we first introduce the dataset and evaluation metrics followed by the implementation details, and then we provide the quantitative results of our method. Subsequently, we show the ablation studies. Finally, we visualize the retrieval results and conduct qualitative analysis.

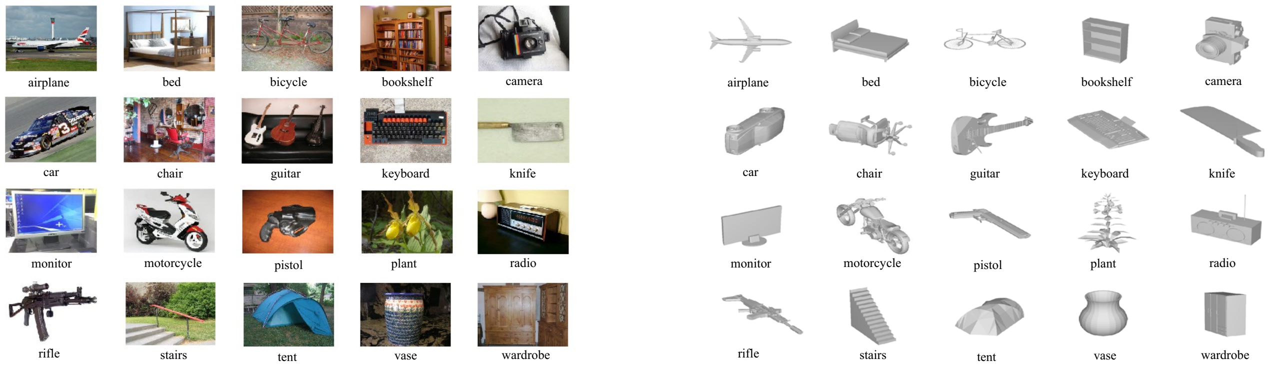
4.1 Datasets and evaluation metrics

4.1.1 Dataset

We conduct variance experiments on the most popular MI3DOR-1 dataset for 3D model retrieval. The source domain is a 2D image, and the target domain is 12 views of each 3D model. The MI3DOR-1 dataset consists of 21,000 2D real images and 7690 3D virtual models belonging to 21 categories. There are 10,500 2D images and 3845 3D models for training, while 10,500 2D images and 3845 3D models for testing. The examples of MI3DOR-1 are shown in Fig. 4.

4.1.2 Evaluation metrics

Following the same evaluation criteria of the state-of-the-art methods, seven popular evaluation metrics are selected to verify the effectiveness of our experiment, including nearest neighbor (NN), first tier (FT), second tier(ST), *F*-measure, discounted gain (DCG), average normalized modified retrieval rank (ANMRR), and area under ROC curve (AUC). The retrieval accuracy of the results returned by nearest neighbors, or the model's accuracy that is most comparable to the retrieval target, can be assessed by NN. The return values of the first *K* and 2K items, respectively, are used to define FT and ST, where *K* is the number of the relevant retrieved objects. The recall rate of retrieval results can be evaluated using these two metrics. The accuracy and the return rate of the data from the previous *K* items are evaluated jointly using the *F* measure. DCG is a statistical method that assigns relevant results to the top position with a higher weight without taking the lower results into account. The system's overall quality can be reflected by AUC value. As a rank-based measure, ANMRR evaluates the ordering information of related objects in the retrieved object. The higher value indicates better performance in terms of all evaluation criteria except ANMRR.



(a) Examples of 2D images in MI3DOR-1 dataset.

(b) Examples of 3D models in MI3DOR-1 dataset.

Fig. 4 Examples of MI3DOR-1 dataset

4.2 Implementation details

These experiments are carried out on an NVIDIA GTX 2080 Ti GPU with a memory of 11 GB. For the proposed method, β is set to 0.001 as [42], and the learning rate is set to 0.001. The image batch size is set to 32, the iteration in an epoch is 2,000, and the training epoch is 20. The perturbation coefficient α is set as 0.2, and λ in Eq. 5 is set to 0.3. The confidence threshold ϵ [42] is 0.5. For the safe training parameters, we keep the same with [42].

4.3 Quantitative results and analysis

4.3.1 Comparative methods

We mainly compare the qualitative outcomes of our proposed algorithm with a few methodologies to validate the effectiveness of domain alignment.

First comes the basic deep learning method AlexNet [52]. This method used a convolution neural network to directly extract features from 3D multiple views and 2D images and there was no transfer learning in this algorithm. Stochastic gradient descent with a batch size of 128 examples, momentum of 0.9, and weight decay of 0.0005 are adopted during the pre-training.

MEDA [11] and JGSA [12] are the traditional transfer learning methods.MEDA jointly trained the domain-invariant classifier in the Grassmann manifold and implemented dynamic alignment of cross-domain distributions in the manifold. As to training and testing settings, [12] set the manifold feature dimension $d = 20, 30, 40$ for Office+Caltech10, USPS+MNIST, and ImageNet+VOC datasets, respectively. The iteration number was set to $T = 10$. The RBF kernel was used with the bandwidth set to be the variance of inputs. The regularization parameters were set as $p = 10$, $\lambda = 10$, $\eta = 0.1$, and $\rho = 1$. As another traditional transfer learning method, JGSA made a constraint on the coupled projections of the source and target domains and projected them into a common low-dimensional subspace for less geometric and distribution shift. For the training and testing configurations, $\lambda = 11$, $\mu = 1$ were fixed in all the experiments, such that the distribution shift, subspace shift, and target variance are treated as equally important.

When it comes to deep transfer learning methods, there also exist several representative methods such as JAN [30], DLEA [29], DANN [10], HIFA [53], and MSTN [54]. To decrease the domain discrepancy in an adversarial training process, [30] proposed a method that makes use of the maximum mean difference criterion and multiple domain-specific layers, where parameter settings are $\lambda = 0$ within 100 iterations and then set it to the cross-validated value. This allowed the JDD penalty to be less sensitive to noisy signals at the early stages of the training process. Considering the feature learning and distribution alignment, DLEA implemented a domain adversarial loss and a center alignment loss. It is noted that For the discriminator D, the identical architecture was utilized. The batch size was set as 128. The entire framework was trained with the rate decay strategy (the original learning rate was 0.01) in an end-to-end manner by SGD with 0.9 momentum. DANN deployed adversarial learning into transfer learning for the first time, while introducing a method to find transferable features between different domains. To speed up the experimental procedure, the domain adaptator was stuck to the three fully connected layers ($x \rightarrow 1024 \rightarrow 1024 \rightarrow 2$), except for MNIST where a simpler ($x \rightarrow 100 \rightarrow 2$) architecture was adopted. HIFA sought a method to minimize the maximum mean feature discrepancy of the distributions of two domains for domain alignment. To conduct semantic transfer learning, MSTN minimized the Euclidean distance between category-level centers of source and target domains. It set $\gamma = \lambda$, where $\lambda = \frac{2}{1+\exp(-\gamma \cdot p)} - 1$ to suppress the noisy information brought by false labels and p is the training process.

Table 1 Comparison results with other methods on MI3DOR-1

Methods	NN	FT	ST	F	DCG	ANMRR	AUC
AlexNet [52]	0.424	0.323	0.469	0.099	0.345	0.667	–
DANN [10]	0.650	0.505	0.643	0.112	0.542	0.474	–
JAN [30]	0.446	0.343	0.495	0.085	0.363	0.647	–
JGSA [12]	0.612	0.443	0.599	0.116	0.473	0.541	–
MEDA [11]	0.430	0.344	0.501	0.046	0.361	0.646	–
MSTN [54]	0.789	0.622	0.779	0.154	0.657	0.358	0.557
DLEA [29]	0.764	0.558	0.716	0.143	0.597	0.421	–
HIFA [53]	0.778	0.618	0.768	0.151	0.654	0.362	–
Ours	**0.801**	**0.632**	**0.787**	**0.155**	**0.667**	**0.348**	**0.569**

The best results are in bold

Table 2 Experimental results of ablation studies of T3DRN and T3DRN-SVAM

Methods	NN	FT	ST	F	DCG	ANMRR	AUC
T3DRN-SVAM	0.790	0.629	0.775	0.153	0.658	0.359	0.559
T3DRN	**0.801**	**0.632**	**0.787**	**0.155**	**0.667**	**0.348**	**0.569**

Table 3 Experimental results under different balance coefficient λ ($\alpha = 0.2$ and $\epsilon = 0.5$)

λ	NN	FT	ST	F	DCG	ANMRR	AUC
0.1	0.776	0.605	0.742	0.142	0.643	0.369	0.537
0.2	0.781	0.610	0.759	0.147	0.658	0.359	0.559
0.3	**0.801**	**0.632**	**0.787**	**0.155**	**0.667**	**0.348**	**0.569**
0.4	0.794	0.625	0.776	0.153	0.651	0.360	0.553
0.5	0.790	0.621	0.761	0.144	0.645	0.367	0.541

Table 4 Experimental results under different perturbation coefficient α ($\lambda = 0.3$ and $\epsilon = 0.5$)

α	NN	FT	ST	F	DCG	ANMRR	AUC
0.2	**0.801**	**0.632**	**0.787**	**0.155**	**0.667**	**0.348**	**0.569**
0.3	0.786	0.618	0.765	0.149	0.653	0.369	0.548
0.4	0.770	0.607	0.761	0.137	0.643	0.375	0.543

4.3.2 Quantitative results with other methods and analysis

We conduct extensive experiments to compare our T3DRN approach and other baseline methods as shown in Table 1. We can make an obvious observation of a significant improvement in NN for T3DRN, reaching 0.801. Our proposed method yields a 0.023 gain of mAP against the HIFA in [53]. Also, compared with other baseline methods, i.e., DANN [10], the performance of T3DRN increases dramatically by more than 60%. The results demonstrate the superiority of T3DRN, which stems from the SVAM module and the extra proper view information captured.

4.4 Ablation studies

4.4.1 The effectiveness of SVAM

To validate the impact of our proposed SVAM module, we also conduct a wide range of ablation studies shown in Table 2. We begin with the very basic model in which SVAM is removed, denoted as the T3DRN-SVAM method. It is easy to observe a significant metric increase by around 0.01 for all of these 7 metrics when the SVAM is plugged. This improvement has proved the effectiveness of SVAM module in terms of integrating the proper view information with image features thus aligning the features of 2D images and 3D models better than the method without SVAM.

4.4.2 The effectiveness of balance coefficient

To explore the effectiveness of the different values of the balance coefficient λ ($\alpha = 0.2$ and $\epsilon = 0.5$), we also conducted relative experiment shown in Table 3. We can see that the retrieval performance slowly increases from $\lambda = 0.1$, and peaks at 0.3. This shows the importance of SVAM. However, if λ is greater than 0.3, the performance begins to decrease.

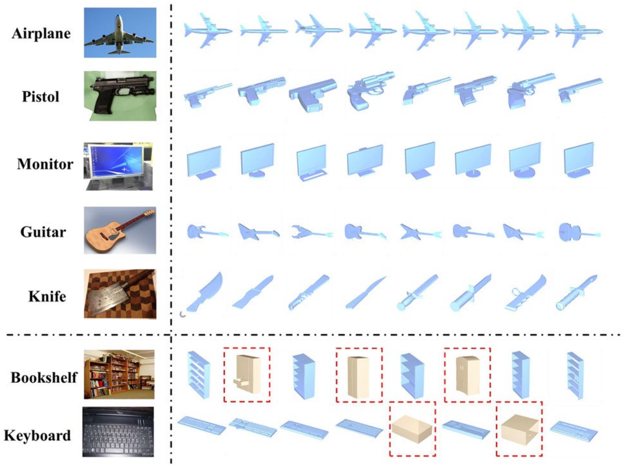


Fig. 5 Qualitative results of our proposed T3DRN

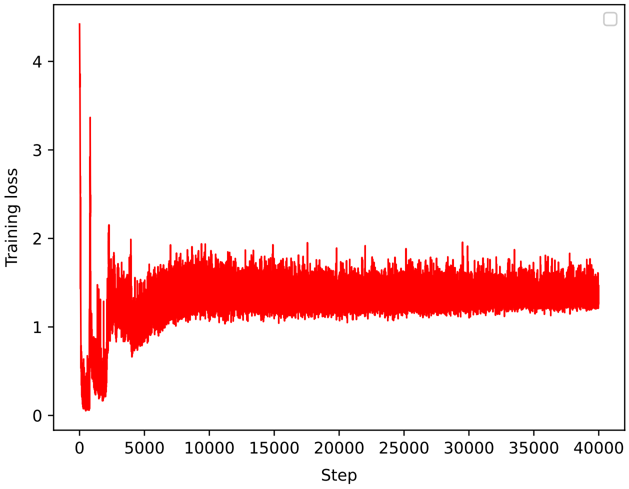


Fig. 6 Loss (objective) function value of training

4.4.3 The effectiveness of the perturbation coefficient α

To explore the effectiveness of the perturbation coefficient α, we also show the experimental results in Table 4. It can be seen that the best performance appears when α is set as 0.2. When the α value rises, the performance decreases considerably. This is because a bigger perturbation value is likely to mislead the model learning during training.

4.5 Qualitative results and analysis

4.5.1 Qualitative results of T3DRN

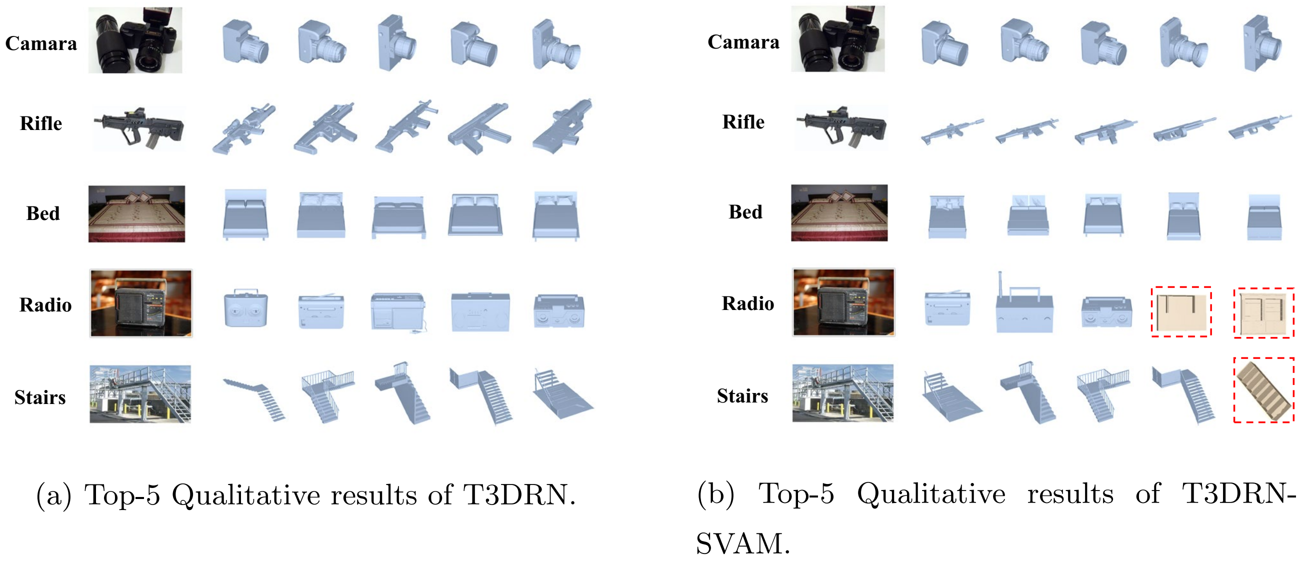
The qualitative results of T3DRN are shown in Fig. 5. It is noticeable that our proposed T3DRN method can always retrieve similar 3D models, especially for the unique classes. This should be owed to the domain-invariant features learned by T3DRN with proper view information obtained by SVAM. For some classes that are not unique, for instance, the bookshelf and keyboard, the retrieval error might happen occasionally due to the similar appearance with other classes in the dataset. (With wardrobe and tent respectively).

4.5.2 Convergence process of training

To better display the convergence process of our proposed model, we show the loss (objective) function value of our proposed model during training in Fig. 6. We can observe that initially, the objective function value highly fluctuates because the model cannot align the data of two domains. After 10k iterations, the model starts to become more stable and eventually converges at around 1.3 after 40k interactions (Fig. 6).

4.5.3 Comparative qualitative results of T3DRN and T3DRN-SVAM

The qualitative results of T3DRN are shown in Fig. 7. It is noticeable that our proposed T3DRN method shown in Fig. 7a always outperforms the T3DRN-SVAM method shown in Fig. 7b due to the appropriate view information



(a) Top-5 Qualitative results of T3DRN.

(b) Top-5 Qualitative results of T3DRN-SVAM.

Fig. 7 Comparative qualitative results of T3DRN and T3DRN-SVAM (Top-5 retrieval results)

achieved that facilitates the whole feature learning process. T3DRN-SVAM, however, retrieves a wrong 3D model (bookshelf), if it can succeed in obtaining the view information of this shelf, this mistake may have been avoided.

5 Conclusion

In this paper, a novel end-to-end transformer-based 3D model retrieval network (T3DRN) for unsupervised 2D image-based 3D model retrieval was proposed to facilitate the learning of the domain-invariant features. This T3DRN can also capture useful view information to guide the training process. To this end, we proposed another innovative unit, named shared view-guided attentive module (SVAM) to integrate the image features with guided view information. We tested this plug-and-play method on the most popular standard dataset and the results turn out that our method outperformed the state-of-the-art method by a wide margin in terms of all seven metrics.

Data availability The dataset of this paper can be found at https://github.com/tianbao-li/MI3DOR.

Conflict of interest The authors declare no conflict of interest.

References

1. Veerasamy, B., Annadurai, S.: Video compression using hybrid hexagon search and teaching–learning-based optimization technique for 3D reconstruction. Multimed. Syst. **27**, 45–59 (2021)
2. Kirya, M., Debattista, K., Chalmers, A.: Using virtual environments to facilitate refugee integration in third countries. Virtual Real. **27**(1), 97–107 (2023)
3. Liu, X., Pang, Y., Jin, R., Liu, Y., Wang, Z.: Dual-domain reconstruction network with V-Net and K-Net for fast MRI. Magn. Reson. Med. **88**(6), 2694–2708 (2022)
4. Liu, Y., Pang, Y., Liu, X., Liu, Y., Nie, J.: DIIK-Net: a full-resolution cross-domain deep interaction convolutional neural network for MR image reconstruction. Neurocomputing **517**, 213–222 (2023)
5. Gao, A., Pang, Y., Nie, J., Shao, Z., Cao, J., Guo, Y., Li, X.: ESGN: efficient stereo geometry network for fast 3D object detection. IEEE Trans. Circuits Syst. Video Technol. **2022**, 1 (2022)
6. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: deep learning on point sets for 3D classification and segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017, pp. 77–85 (2017)
7. 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, pp. 1912–1920 (2015)
8. Furuya, T., Ohbuchi, R.: Deep aggregation of local 3D geometric features for 3D model retrieval. In: Wilson, R.C., Hancock, E.R., Smith, W.A.P. (eds.) Proceedings of the British Machine Vision Conference, BMVC (2016)
9. Feng, Y., Feng, Y., You, H., Zhao, X., Gao, Y.: Meshnet: mesh neural network for 3D shape representation. In: The 33rd AAAI Conference on Artificial Intelligence, The 31st Innovative Applications of Artificial Intelligence Conference, IAAI, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI, pp. 8279–8286 (2019)
10. Ganin, Y., Lempitsky, V.S.: Unsupervised domain adaptation by backpropagation. In: Bach, F.R., Blei, D.M. (eds.) Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6–11 July 2015. JMLR Workshop and Conference Proceedings, vol. 37, pp. 1180–1189 (2015)
11. Wang, J., Feng, W., Chen, Y., Yu, H., Huang, M., Yu, P.S.: Visual domain adaptation with manifold embedded distribution alignment. In: Boll, S., Lee, K.M., Luo, J., Zhu, W., Byun, H., Chen, C.W., Lienhart, R., Mei, T. (eds.) 2018 ACM Multimedia Conference on Multimedia Conference, MM, pp. 402–410 (2018)
12. Zhang, J., Li, W., Ogunbona, P.: Joint geometrical and statistical alignment for visual domain adaptation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017, pp. 5150–5158 (2017)
13. Yue, X., Zhang, Y., Zhao, S., Sangiovanni-Vincentelli, A.L., Keutzer, K., Gong, B.: Domain randomization and pyramid consistency: simulation-to-real generalization without accessing target domain data. CoRR arXiv:abs/1909.00889 (2019)
14. Chen, Y., Ouyang, X., Zhu, K., Agam, G.: Semi-supervised domain adaptation for semantic segmentation. CoRR arXiv:abs/2110.10639 (2021)
15. Peng, D., Lei, Y., Hayat, M., Guo, Y., Li, W.: Semantic-aware domain generalized segmentation. CoRR arXiv:abs/2204.00822 (2022)
16. Liu, A.-A., Shao, Z., Wong, Y., Li, J., Su, Y.-T., Kankanhalli, M.: LSTM-based multi-label video event detection. Multimed. Tools Appl. **78**, 677–695 (2019)
17. Chu, F., Cao, J., Shao, Z., Pang, Y.: Illumination-guided transformer-based network for multispectral pedestrian detection. In: Artificial Intelligence: Second CAAI International Conference, CICAI 2022, Beijing, China, August 27–28, 2022, Revised Selected Papers, Part I, pp. 343–355 (2022). Springer, London
18. Li, Y., Yin, K., Liang, J., Tan, Z., Wang, X., Yin, G., Wang, Z.: A multitask joint framework for real-time person search. Multimed. Syst. **29**(1), 211–222 (2023)
19. Wang, J., Pang, Y., Cao, J., Sun, H., Shao, Z., Li, X.: Deep intra-image contrastive learning for weakly supervised one-step person search. Preprint arXiv:2302.04607 (2023)
20. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, Minneapolis, MN, USA, June 2–7, 2019, Volume 1 (Long and Short Papers), pp. 4171–4186 (2019). https://doi.org/10.18653/v1/n19-1423
21. Maturana, D., Scherer, S.: Voxnet: a 3D convolutional neural network for real-time object recognition. Intell. Robots Syst **2015**, 1 (2015)

22. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3D shapenets: a deep representation for volumetric shapes. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7–12, 2015, pp. 1912–1920 (2015)
23. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: deep hierarchical feature learning on point sets in a metric space. Computer Vision and Pattern Recognition, arXiv (2017)
24. 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, pp. 945–953 (2015)
25. Gao, Z., Shao, Y., Guan, W., Liu, M., Cheng, Z., Chen, S.: A novel patch convolutional neural network for view-based 3D model retrieval. Computer Vision and Pattern Recognition, arXiv (2021)
26. Watanabe, S., Takahashi, S., Wang, L.: Aggregating viewpoints for effective view-based 3D model retrieval. In: 2021 25th International Conference Information Visualisation (IV) (2021)
27. Kim, D., Seo, M., Park, J., Choi, D.: Source domain subset sampling for semi-supervised domain adaptation in semantic segmentation. CoRR arXiv:abs/2205.00312 (2022)
28. Li, B., Wang, Y., Zhang, S., Li, D., Keutzer, K., Darrell, T., Zhao, H.: Learning invariant representations and risks for semi-supervised domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1104–1113 (2021)
29. Zhou, H., Liu, A., Nie, W.: Dual-level embedding alignment network for 2D image-based 3D object retrieval. In: Amsaleg, L., Huet, B., Larson, M.A., Gravier, G., Hung, H., Ngo, C., Ooi, W.T. (eds.) Proceedings of the 27th ACM International Conference on Multimedia, MM, pp. 1667–1675 (2019)
30. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Deep transfer learning with joint adaptation networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning, ICML. Proceedings of Machine Learning Research, vol. 70, pp. 2208–2217 (2017)
31. Li, X., Zhang, W., Ding, Q., Sun, J.-Q.: Multi-layer domain adaptation method for rolling bearing fault diagnosis. Signal Process. **2019**, 1 (2019)
32. Wang, Q., Du, P., Liu, X., Yang, J., Wang, G.: Adversarial unsupervised domain adaptation for cross scenario waveform recognition. Signal Process. **2020**, 1 (2020)
33. Saporta, A., Douillard, A., Vu, T., Pérez, P., Cord, M.: Multi-head distillation for continual unsupervised domain adaptation in semantic segmentation. CoRR arXiv:abs/2204.11667 (2022)
34. Hoyer, L., Dai, D., Gool, L.V.: HRDA: context-aware high-resolution domain-adaptive semantic segmentation. CoRR arXiv:abs/2204.13132 (2022)
35. Zhao, K., Hu, J., Shao, H., Hu, J.: Federated multi-source domain adversarial adaptation framework for machinery fault diagnosis with data privacy. Reliab. Eng. Syst. Saf. **236**, 109246 (2023)
36. Zhao, K., Jia, F., Shao, H.: A novel conditional weighting transfer Wasserstein auto-encoder for rolling bearing fault diagnosis with multi-source domains. Knowl.-Based Syst. **262**, 110203 (2023)
37. Jin, B., Vai, M.I.: An adaptive ultrasonic backscattered signal processing technique for instantaneous characteristic frequency detection. Bio-Med. Mater. Eng. **24**(6), 2761–2770 (2014)
38. Zheng, Q., Zhao, P., Li, Y., Wang, H., Yang, Y.: Spectrum interference-based two-level data augmentation method in deep learning for automatic modulation classification. Neural Comput. Appl. **33**(13), 7723–7745 (2021)
39. Zheng, Q., Zhao, P., Wang, H., Elhanashi, A., Saponara, S.: Fine-grained modulation classification using multi-scale radio transformer with dual-channel representation. IEEE Commun. Lett. **26**(6), 1298–1302 (2022)
40. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 5998–6008 (2017)
41. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16 × 16 words: transformers for image recognition at scale. In: ICLR (2021)
42. Sun, T., Lu, C., Zhang, T., Ling, H.: Safe self-refinement for transformer-based domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7191–7200 (2022)
43. Liu, A.-A., Su, Y.-T., Nie, W.-Z., Kankanhalli, M.: Hierarchical clustering multi-task learning for joint human action grouping and recognition. IEEE Trans. Pattern Anal. Mach. Intell. **39**(1), 102–114 (2016)
44. Xu, N., Zhang, H., Liu, A.-A., Nie, W., Su, Y., Nie, J., Zhang, Y.: Multi-level policy and reward-based deep reinforcement learning framework for image captioning. IEEE Trans. Multimed. **22**(5), 1372–1383 (2019)
45. Liu, A.-A., Wang, Y., Xu, N., Nie, W., Nie, J., Zhang, Y.: Adaptively clustering-driven learning for visual relationship detection. IEEE Trans. Multimed. **23**, 4515–4525 (2020)
46. Ji, Z., Hu, Z., Wang, Y., Shao, Z., Pang, Y.: Reinforced pedestrian attribute recognition with group optimization reward. Image Vis. Comput. **128**, 104585 (2022)
47. Shao, Z., Han, J., Marnerides, D., Debattista, K.: Region-object relation-aware dense captioning via transformer. IEEE Trans. Neural Netw. Learn. Syst. **2022**, 1 (2022)
48. Shao, Z., Han, J., Debattista, K., Pang, Y.: Textual context-aware dense captioning with diverse words. IEEE Trans. Multimed. **2023**, 1 (2023)
49. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3202–3211 (2022)
50. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: unet-like pure transformer for medical image segmentation. In: European Conference on Computer Vision, pp. 205–218. Springer, London (2022)
51. Ruder, S.: An overview of gradient descent optimization algorithms. Preprint arXiv:1609.04747 (2016)
52. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017)
53. 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 29th International Conference on International Joint Conferences on Artificial Intelligence, pp. 839–845 (2021)
54. Xie, S., Zheng, Z., Chen, L., Chen, C.: Learning semantic representations for unsupervised domain adaptation. In: International Conference on Machine Learning, pp. 5423–5432. PMLR (2018)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.