文章编号:1672-3961(2025)02-0071-07 DOI:10.6040/j.issn.1672-3961.0.2024.164

基于实例判别与特征增强的单图三维模型检索

刁振宇[1,2],韩小凡[1,2],张承宇[1,2],聂慧佳[1,2],赵秀阳[1,2],牛冬梅[1,2*]
(1. 山东省泛在智能计算重点实验室(筹), 山东 济南 250022; 2.济南大学信息科学与工程学院, 山东 济南 250022)

摘要:为减小图像检索三维模型算法中图像域和模型域间的模态差距,提出一种由 4 个模块组成的神经网络算法模型。数据交换模块通过一定概率交换图像和三维模型数据,使图像域网络具有模型域特征学习能力,模型域网络具有图像域特征学习能力,初步减小模态差距。特征对齐模块有实例样本判别损失函数和图像模型配对损失函数,进一步对齐图像域和模型域。实例判别损失函数将每个实例视为独立个体类,对其进行分类,使相同实例的图像和三维模型的特征相似。图像模型配对模块旨在拉近相同实例的图像和三维模型,推远不同实例的图像和三维模型。基于对比学习在图像域中增加特征增强模块,提高图像域内特征区分性。试验结果表明,提出的算法在 3 个常见数据集 Pix3D、CompCars 和 StanfordCars 上取得良好效果,检索精度较现有经典方法提高 4.5%。实现图像域和三维模型域对齐,减小模态差距,提高图像检索三维模型精度。

关键词:三维模型检索;度量学习;对比学习;多模态;跨模态检索
中图分类号:TP183 **文献标志码**:A
引用格式:刁振宇,韩小凡,张承宇,等.基于实例判别与特征增强的单图三维模型检索[J].山东大学学报(工学版),2025,55(2):71-77.
DIAO Zhenyu, HAN Xiaofan, ZHANG Chengyu, et al. Single image 3D model retrieval based on instance discrimination and feature enhancement[J]. Journal of Shandong University (Engineering Science), 2025, 55(2):71-77.

Single image 3D model retrieval based on instance discrimination and feature enhancement

DIAO Zhenyu[1,2], HAN Xiaofan[1,2], ZHANG Chengyu[1,2], NIE Huijia[1,2], ZHAO Xiuyang[1,2], NIU Dongmei[1,2*]
(1. Shandong Provincial Key Laboratory of Ubiquitous Intelligent Computing, Jinan 250022, Shandong, China; 2. School of Information Science and Engineering, University of Jinan, Jinan 250022, Shandong, China)

Abstract: To reduce the modal gap between the image domain and the model domain in 3D model retrieval algorithms, a neural network algorithm model consisting of four modules was proposed. The data exchange module exchanged image and 3D model data with a certain probability, allowing the image domain network to learn model domain features and the model domain network to learn image domain features, thus initially reducing the modal gap. The feature alignment module included an instance sample discrimination loss function and an image-model pairing loss function, which further aligned the image domain and model domain. The instance discrimination loss function treated each instance as an independent class and classified it, making the features of the same instance′s images and 3D models similar. The image-model pairing module aimed to bring closer the images and 3D models of the same instance and push apart the images and 3D models of different instances. Based on contrastive learning, a feature enhancement module was added to the image domain to improve feature discrimination within the image domain. The experimental results showed that the proposed algorithm achieved good results on three common datasets: Pix3D, CompCars, and StanfordCars, improving retrieval accuracy by up to 4.5% compared to existing classical methods. This aligned the image domain and the 3D model domain, reduced the modal gap, and improved the accuracy of image retrieval of 3D models.

Keywords: 3D model retrieval; metric learning; contrastive learning; multimodal; cross modal retrieval

收稿日期:2024-07-14
基金项目:国家自然科学基金资助项目(62102163);山东省高等学校青年创新团队发展计划资助项目;山东省科技型中小企业创新能力提升工程资助项目(2023TSGCO244)
第一作者简介:刁振宇(1998—),男,山东枣庄人,硕士研究生,主要研究方向为三维模型表示、三维模型检索。E-mail:dzy10242023@163.com
*通信作者简介:牛冬梅(1988—),女,山东泰安人,副教授,硕士生导师,博士,主要研究方向为三维模型处理。
E-mal:ise_niudm@ujn.edu.cn

0 引言

随着三维模型建模技术的成熟，与之相关的技术也得到了长足发展和应用[1-2]。基于图像三维模型检索是计算机视觉领域的一个重要研究方向，在现实生活中应用广泛，涉及生物医学、建筑工程、计算机辅助设计以及虚拟现实等多个领域。

图像和三维模型属于两个不同数据分布，二者之间存在巨大领域差距。算法核心在于如何有效弥合图像域与三维模型域之间语义鸿沟。为降低域对齐复杂性，文献[3-6]提出在图像域和三维模型域之间生成中间域的方法；文献[3]对属于同一类别的源域样本和目标域样本特征进行平均得到中间域，通过将源域和目标域样本特征向中间域拉近的方法减小域对齐难度；文献[4]利用傅里叶变换生成中间域，使用渐进性对抗自适应策略减小中间域影响，减小两个域之间差距。在损失函数方面，文献[7-8]使用三元组损失函数，通过比较锚定样本、正样本和负样本之间相似度，帮助模型学习更好特征表示；文献[9]提出使用对比损失函数，通过类别和实例两个级别对比损失实现域对齐。

以上方法取得良好效果，存在着一些问题：(1)无论三元组损失函数还是对比损失函数，在性能上仍有欠缺，这是由损失函数限制所致。三元组损失函数本身存在局限，需解决最难负样本挖掘问题。对比损失函数不需挖掘最难分负样本，文献[10]指出，要使对比学习效果良好，训练批量大小应尽可能大。三维模型表示本身是资源密集型的，无法保证批量大小。(2)现有方法大多关注于减小领域差距，忽视图像域本身特征判别性对算法影响。为解决上述问题，提出一种新的算法框架，其中包括数据交换和颜色转换、特征提取、特征对齐和增强模块。数据交换通过在查询图像和三维模型数据之间以一定概率进行数据交换，实现特征融合。特征对齐包括实例级样本判别和图像模型配对损失函数。实例级样本判别通过将每个实例视为独立个体类，对其进行分类，确保两个域投影到相同特征空间，使得相同实例特征更为相似。图像模型配对损失函数用于计算图像和模型之间相似性，希望提高同一实例图像和三维模型相似性，减小不同实例图像和三维模型相似性。受文献[11]启发，基于对比学习，本研究在图像域内拉近同一实例图像特征向量相似性，同时推远不同实例图像特征向量相似性，增强学习到图像域特征判别能力。

1 相关工作

三维模型检索是从三维模型库中找到与待查询样本相匹配三维模型。根据查询样本类型不同，这些方法可以分为使用模型检索三维模和使用图像检索三维模型的方法。

近年来，模型检索三维模型方法受到广泛关注，根据三维模型表示方式不同，这些方法又可分为基于体素[12-15]、基于点云[16-19]和基于视图[20-23]的三维模型检索。基于体素的三维模型检索是将三维模型表示成大小固定立方块，基于此进行检索的方法。基于点云的三维模型检索方法是将三维模型处理成一系列点用以学习表示。一个经典算法是 PointNet[16]，专门用于从无序点云数据集中提取特征，在三维模型的分类和检索任务中表现出色并且具有很强稳定性。基于视图的三维模型检索是一个重要研究领域，主要将三维模型渲染为多个视图，将这些视图作为输入。代表性方法是 MVCNN[23]，将三维模型渲染为多视图输入神经网络，获得特征响应图，对这些响应图进行最大池化操作，取得显著效果。

图像检索三维模型方法可分为类别级和实例级的三维模型检索方法。图像类别级的三维模型检索主要是检索出与查询图像属于同一类别的三维模型，不需要考虑具体三维模型。实例级的三维模型检索方法需更为细致搜索，需找到与查询图像最匹配三维模型。一些类别级的三维模型检索方法在图像和三维模型充分标注好情况下进行。例如，文献[24]利用二维图像位置信息表示三维模型，通过利用二维图像和三维模型间视觉相关性，有效减小领域之间差距。该方法还提出拉近同一域内和不同域间相同类别和推远不同类别样本之间距离模块，有效减小领域之间差距。这类方法需要大量人工标记三维模型，使得应用受限。一些研究者提出在三维模型没有标签情况下进行检索。这类方法往往考虑的是知识迁移，通过将学好的图像域网络知识转移到三维模型域网络实现检索三维模型。例如文献[25]使用条件生成对抗性网络实现特征自适应，文献[26]进一步通过对抗性学习和类中心来加强跨域特征空间域级别和类级别对齐。实例级的三维模型检索方法同样受到广泛关注。在文献[27]中研究者提出一种二维图像的三维模型姿态估计方法，使用姿态信息渲染三维模型，通过这种方式可以提高检索准确性。文献[7]

为减少域差距,考虑纹理信息,通过生成对抗方式给渲染图像合成三维模型视图虚拟纹理。文献[9]提出一种基于对比学习算法框架,通过两个域中相同实例图像和三维模型拉近,不同实例图像和三维模型推远,达到在类别和实例级别进行向量映射,减少域间差距。

2　问题研究

本章对提出方法进行解释,图 1 为算法流程图。算法包括数据交换和颜色变换、实例判别以及特征增强模块。

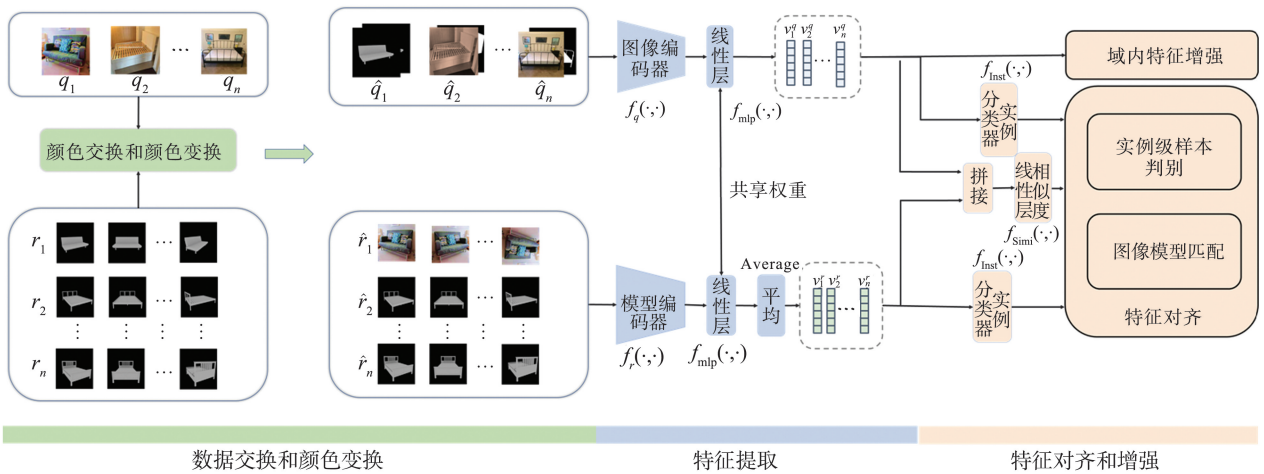

图 1　算法流程图

Fig.1　Algorithm flowchart

2.1　问题定义

假设数据集 S 包含 H 个实例,其中第 i 个实例 S_i 包含一个查询图像 q_i、一个被渲染成 M 个多视角图像的三维模型 r_i 和实例的个体标签 y_i,有数据定义为

$$\begin{cases} S=\{s_i\}_{i=1}^{H} \\ s_i=(q_i,r_i,y_i), \\ r_i=\{t_i^m\}_{m=1}^{M} \end{cases} \tag{1}$$

式中:$\{t_i^m\}_{m=1}^{M}$ 为第 i 个实例的三维模型 r_i 的 M 个多视角渲染视图;实例个体标签 y_i 的标签数和实例的数量相等,每个实例都是一个独立类别。

2.2　数据交换和颜色变换

算法采用文献[9]的随机颜色变换和 mask 图的操作。记 q_i 经颜色变换后为 $\bar{q}_i$, $\bar{q}_i$ 的 mask 图像为 k_i,数据变为

$$\begin{cases} S=\{s_i\}_{i=1}^{|B|} \\ s_i=(\bar{q}_i,r_i,y_i,k_i), \\ r_i=\{t_i^m\}_{m=1}^{M} \end{cases} \tag{2}$$

式中 B 为数据集 S 的小批次数据集。

交换模块主要目的有两个:(1)通过一定概率将图像域和模型域数据进行交换使得图像域编码器网络具有学习模型域数据能力;模型域编码器网络具有学习图像域数据能力。(2)通过学习不同分布数据可以防止图像域和模型域编码器网络过拟合。

设定超参数 $P\in[0,1]$,当随机变量 p 小于 P 时进行交换操作。$\bar{q}_i$ 先经过 M 种不同图像增强方式得到 M 张增强后图像 $\{\bar{q}_i^m\}_{m=1}^{M}$,第 i 个实例三维模型随机选取第 j 个渲染视图 t_i^j 作为交换视图,有

$$\hat{q}_i=\begin{cases} t_i^j, & p<P \\ \bar{q}_i, & p\geqslant P \end{cases}, \tag{3}$$

$$\hat{r}_i=\{\hat{t}_i^m\}_{m=1}^{M}=\begin{cases} \{\bar{q}_i^m\}_{m=1}^{M}, & p<P \\ \{t_i^m\}_{m=1}^{M}, & p\geqslant P \end{cases}, \tag{4}$$

式中,$\hat{q}_i$、$\hat{r}_i$ 为经过交换模块后查询图像和三维模型,$\{\hat{t}_i^m\}_{m=1}^{M}$ 为经过交换模块后三维模型渲染视图,输入到网络数据为

$$\begin{cases} S=\{s_i\}_{i=1}^{|B|} \\ s_i=(\hat{q}_i,\hat{r}_i,y_i,k_i)。 \\ \hat{r}_i=\{\hat{t}_i^m\}_{m=1}^{M} \end{cases} \tag{5}$$

2.3　图像与模型特征提取

如图 1 所示,$\hat{q}_i$ 通过图像编码器网络 $f_q(\cdot,\cdot)$ 提取图像高维特征向量 $\boldsymbol{v}_i^q=f_q(\hat{q}_i,\theta_q)$,其中,$\theta_q$ 是图像编码器网络参数。$\hat{r}_i$ 经过模型域编码器网络 $f_r(\cdot,\cdot)$ 得到三维模型高维特征向量 $\boldsymbol{v}_i^r=f_r(\hat{r}_i,\theta_r)$,$\theta_r$ 是模型域编码器网络参数。为进一步降低维度,减小两个域之间差距,使用共享参数神经网络 $f_{\text{mlp}}(\cdot,\cdot)$,将 $\boldsymbol{v}_i^q$、$\boldsymbol{v}_i^r$ 进一步处理。$\boldsymbol{v}_i^q$ 使用 $f_{\text{mlp}}(\cdot,\cdot)$

得到低维特征向量 $\bar{\boldsymbol{v}}_i^q = f_{\text{mlp}}(\boldsymbol{v}_i^q, \theta_{\text{mlp}})$，其中 θ_{mlp} 是网络参数。三维模型的多视图特征向量经过 Average 模块将模型 M 个视图特征向量进行平均得到低维特征向量 $\bar{\boldsymbol{v}}_i^r = \text{Average}(f_{\text{mlp}}(\boldsymbol{v}_i^r, \theta_{\text{mlp}}))$。$\bar{\boldsymbol{v}}_i^q$、$\bar{\boldsymbol{v}}_i^r$ 特征维度均为 128 维。

2.4 跨域特征对齐

特征对齐模块主要是减小两个域之间差距，包括实例级样本判别和图像模型匹配两个损失函数。

2.4.1 实例级样本判别

通过将每个实例视为一个独立类别，将三维模型和图像都划分到相同实例类，将每个样本实例投射到相同特征空间。$\bar{\boldsymbol{v}}_i^q$ 经过实例级分类头网络 $f_{\text{inst}}(\cdot,\cdot)$ 得到图像实例概率向量 $\boldsymbol{p}_i^q = f_{\text{inst}}(\bar{\boldsymbol{v}}_i^q, \theta_{\text{inst}})$，$\theta_{\text{inst}}$ 是分类头网络参数。$\bar{\boldsymbol{v}}_i^r$ 经过网络得到三维模型实例概率向量 $\boldsymbol{p}_i^r = f_{\text{inst}}(\bar{\boldsymbol{v}}_i^r, \theta_{\text{inst}})$，图像实例级样本判别损失函数为

$$L_{\text{img\_inst}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K_1} y_{i,j}\,\text{lb}(\boldsymbol{p}_{i,j}^q), \tag{6}$$

模型实例级样本判别损失函数：

$$L_{\text{model\_inst}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K_1} y_{i,j}\,\text{lb}(\boldsymbol{p}_{i,j}^r), \tag{7}$$

式中，$y_{i,j}$ 为实例样本个体类别，K_1 为个体类别数，N 为小批次中样本数量。

2.4.2 图像模型配对

该模块确定图像和模型是否来自相同实例进行特征对齐。当图像和模型来自相同实例时，希望二者相似概率大；当图像和模型来自不同实例时，希望二者相似概率小。这一模块旨在拉近相同实例图像和三维模型，推远不同实例图像和三维模型。批次中 $\bar{\boldsymbol{v}}_i^q$、$\bar{\boldsymbol{v}}_j^r$ 经过拼接单元 $\text{concat}(\cdot,\cdot)$ 变成一个向量 $\bar{\boldsymbol{v}}_{ij}^{qr} = \text{concat}(\bar{\boldsymbol{v}}_i^q, \bar{\boldsymbol{v}}_j^r)$。经过 $f_{\text{simi}}(\cdot,\cdot)$ 得到二者相似性 $\boldsymbol{p}^{i2m} = f_{\text{simi}}(\bar{\boldsymbol{v}}_{ij}^{qr}, \theta_{\text{simi}})$，$\theta_{\text{simi}}$ 是参数。图像和模型配对损失函数为

$$L_{\text{imm}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K_2} y_{i,j}^{i2m}\,\text{lb}(p_{i,j}^{i2m}), \tag{8}$$

式中：$y_{i,j}^{i2m}$ 为图像和模型是否为同一实例标签，当图像和模型是同一个实例时 $y_{i,j}^{i2m}=1$；当图像和模型不是同一个实例时 $y_{i,j}^{i2m}=0$；K_2 为类别数，取值为 2；N 为小批次中样本数量。

2.5 图像特征增强

根据文献[11]提出的损失函数，基于对比学习，以同实例图像为正样本、不同实例图像为负样本的方式定义特征增强损失函数。目的是通过拉近同一实例图像特征推远不同实例图像特征，增强特征向量区分性。定义域内特征增强损失函数为

$$L_{\text{fe}} = \sum_{i\in B}\frac{-1}{|P(i)|}\sum_{p\in P(i)}\text{lb}\left(\frac{\exp(\bar{\boldsymbol{v}}_i^q\cdot\bar{\boldsymbol{v}}_p^q)}{\sum_{j\in B}\exp(\bar{\boldsymbol{v}}_i^q\cdot\bar{\boldsymbol{v}}_j^q)}\right), \tag{9}$$

式中，$\bar{\boldsymbol{v}}^q$ 为训练批次中图像域网络输出图像特征向量，B 为小批量数据集，$P(i)=\{j\in B \text{ and } y_j=y_i\}$ 为批次中所有与查询图像相同实例图像样本。

3 试验与结果

3.1 数据集

为和以往工作进行比较，本研究和文献[7-9]保持一致，在 Pix3D[28]、CompCars[29]、StanfordCars[30]数据集上进行试验。Pix3D 数据集包含 4 个类、5 118 张图片和 322 个三维模型，其中 2 648 张图片用以训练，2 470 张图片用以测试。CompCars 数据集包含 5 696 张图像、94 个三维模型，训练集包含 3 798 张，测试集包含 1 898 张。StanfordCars 数据集包含 16,185 张图像、134 个三维模型，训练集包含 8 144 张，测试集包含 8 041 张。

3.2 评价指标

试验用 $A_{\text{Top-1}}$、$A_{\text{Top-10}}$、d_{HAU}、d_{IoU} 做为检索结果评价指标。$A_{\text{Top-1}}$ 为预测最相似三维模型和真实三维模型相同数量与所有查询数量比率；$A_{\text{Top-10}}$ 是前 10 个预测三维模型中有真实三维模型数量与所有查询数量比率；d_{HAU} 指的是 Hausdorff 距离；d_{IoU} 指的是交并比。

3.3 实现细节

算法用 ResNet50[31] 为特征提取骨干网络。输入图像域和模型域网络图像尺寸均为 224 像素×224 像素。每个三维模型被渲染成 12 张多视图。算法使用 Pytorch 实现。使用 Adam 做为优化器，学习率设置为 1×10^{-4}。训练批量大小为 16 个，总共训练 400 个 Epoch。

3.4 检索性能对比试验

试验在 3 个数据集上分别和经典方法[7-9,27,32-33]进行比较。如表 1~3 所示，整体上看，提出算法性能优于以往方法。$A_{\text{Top-1}}$ 最高提高4.5%。在 Pix3D 上 chair 类效果不如 HEG-TS[5]，这是因为 chair 类中实例图像间纹理十分相似，HEG-TS 提出一种可以生成纹理只关注几何信息，忽略纹理对算法影响。本研究方法在 Pix3D 其他类以及其他数据集上效果优于以往方法，其中在 StanfordCars 数据集上 $A_{\text{Top-1}}$ 指标提高到 88%，远超以往算法。

表 1　在 Pix3D 数据集上的检索性能
Table 1　Retrieval performance on Pix3D dataset

类别	方法	$A_{\text{Top-1}}$/%	$A_{\text{Top-10}}$/%	d_{HAU}	d_{IoU}
bed	UDF-CGI[32]	19.4	46.6	0.082 1	0.339 7
	Grabneret al.[27]	35.1	83.2	0.038 5	0.559 8
	LFD[8]	64.4	89.0	0.015 2	0.807 4
	HEG-TS[7]	65.3	95.4	0.012 2	0.821 3
	Linet al.[9]	73.3	96.1	0.009 3	0.892 7
	ULIP[33]	74.2	96.3	0.006 8	0.893 1
	本研究算法	**80.3**	**96.8**	**0.004 2**	**0.916 3**
chair	UDF-CGI[32]	17.3	49.1	0.055 9	0.302 7
	Grabneret al.[27]	41.3	73.9	0.030 5	0.546 9
	LFD[8]	58.1	81.8	0.017 0	0.716 9
	HEG-TS[7]	**87.9**	**97.9**	**0.004 1**	**0.906 3**
	Linet al.[9]	79.4	96.3	0.008 0	0.866 1
	ULIP[33]	83.7	97.4	0.005 2	0.883 2
	本研究算法	82.3	94.1	0.004 8	0.857 8
sofa	UDF-CGI[32]	21.7	52.2	0.050 3	0.382 4
	Grabneret al.[27]	44.1	89.9	0.019 7	0.776 2
	LFD[8]	67.0	94.4	0.007 5	0.902 8
	HEG-TS[7]	72.8	97.7	0.004 7	0.907 0
	Linet al.[9]	80.7	97.1	0.004 5	0.932 9
	ULIP[33]	81.5	97.5	0.004 8	0.934 7
	本研究算法	**82.2**	**97.8**	**0.002 5**	**0.943 8**
table	UDF-CGI[32]	12.0	34.2	0.100 3	0.171 5
	Grabneret al.[27]	33.9	66.1	0.060 7	0.450 0
	LFD[8]	53.3	80.1	0.028 8	0.638 3
	HEG-TS[7]	73.7	92.4	0.017 0	0.766 7
	Linet al.[9]	76.9	93.5	0.016 8	**0.808 8**
	ULIP[33]	78.0	94.0	0.015 2	0.791 3
	本研究算法	**78.4**	**94.0**	**0.009 3**	0.806 6
mean	UDF-CGI[32]	17.6	45.5	0.072 2	0.299 1
	Grabneret al.[27]	38.6	78.3	0.037 4	0.583 2
	LFD[8]	60.7	86.3	0.017 1	0.766 3
	HEG-TS[7]	74.9	95.8	0.009 5	0.850 3
	Linet al.[9]	78.9	96.1	0.008 6	0.874 6
	ULIP[33]	79.3	**96.3**	0.008 0	**0.875 1**
	本研究算法	**81.6**	95.1	**0.004 9**	0.873 8

注:黑体为最优结果。

表 2　在 CompCars 数据集上的检索性能
Table 2　Retrievalperformance on CompCars dataset

方法	$A_{\text{Top-1}}$/%	$A_{\text{Top-10}}$/%	d_{HAU}	d_{IoU}
UDF-CGI[32]	2.4	18.2	0.020 7	0.722 4
Grabneret al.[27]	10.2	36.9	0.015 8	0.780 5
LFD[8]	20.5	58.0	0.013 3	0.814 2
HEG-TS[7]	67.1	93.7	0.003 5	0.925 6
Linet al.[9]	77.8	**94.1**	0.002 3	**0.939 9**
ULIP[33]	78.8	94.3	0.002 3	0.937 2
本研究算法	**78.8**	92.1	**0.002 2**	0.938 5

注:黑体为最优结果。

表 3　在 StanfordCars 数据集上的检索性能
Table 3　Retrieval performance on StanfordCars dataset

方法	$A_{\text{Top-1}}$/%	$A_{\text{Top-10}}$/%	d_{HAU}	d_{IoU}
UDF-CGI[32]	3.7	20.1	0.019 8	0.716 9
Grabneret al.[27]	11.3	42.2	0.015 3	0.772 1
LFD[8]	29.5	69.4	0.011 0	0.835 2
HEG-TS[7]	68.4	92.1	0.003 4	0.921 0
Linet al.[9]	83.4	96.4	0.002 1	0.943 1
ULIP[33]	84.3	96.7	0.002 3	0.946 7
本研究算法	**88.8**	**96.7**	**0.001 9**	**0.959 1**

注:黑体为最优结果。

3.5　模块消融试验

本研究在 3 个数据集上进行消融试验,进一步分析算法性能和效果。如表 4~6 所示,试验结果充分验证提出不同模块有效性。

表 4　Pix3D 数据集上不同模块检索性能
Table 4　Retrieval performance of different modules on Pix3D dataset

方法	$A_{\text{Top-1}}$/%	$A_{\text{Top-10}}$/%	d_{HAU}	d_{IoU}
IMM	51.3	88.1	0.014 2	0.629 1
IMM+DE	55.1	88.9	0.014 3	0.653 5
IMM+ID	78.9	93.0	0.004 7	**0.891 0**
IMM+FE+ID	78.1	93.2	0.006 1	0.849 0
IMM+DE+ID	81.2	95.1	0.005 4	0.870 2
IMM+DE+ID+FE	**81.6**	**95.3**	**0.004 9**	0.873 8

注:黑体为最优结果。

表 5　在 CompCars 数据集上不同模块检索性能
Table 5　Retrieval performance of different modules on CompCars dataset

方法	$A_{\text{Top-1}}$/%	$A_{\text{Top-10}}$/%	d_{HAU}	d_{IoU}
IMM	63.9	90.3	0.003 5	0.900 4
IMM+DE	67.5	93.2	0.003 1	0.911 9
IMM+ID	70.1	89.8	0.003 0	0.906 6
IMM+FE+ID	75.1	89.7	0.002 4	0.930 5
IMM+DE+ID	73.8	90.6	0.002 6	0.925 2
IMM+DE+ID+FE	**78.8**	**92.1**	**0.002 2**	**0.938 5**

注:黑体为最优结果。

表 6　在 StanfordCars 数据集上不同模块检索性能
Table 6　Retrieval performance of different modules on StanfordCars dataset

方法	$A_{\text{Top-1}}$/%	$A_{\text{Top-10}}$/%	d_{HAU}	d_{IoU}
IMM	78.5	95.8	0.002 8	0.928 0
IMM+DE	80.4	96.3	0.002 7	0.935 4
IMM+ID	85.8	95.4	0.002 1	0.952 3
IMM+FE+ID	86.9	95.8	0.002 0	0.955 5
IMM+DE+ID	87.9	**97.0**	0.001 9	0.956 5
IMM+DE+ID+FE	**88.8**	96.7	**0.001 9**	**0.959 1**

注:黑体为最优结果。

从结果可以看出在仅仅利用图像配对损失(image model matching, IMM)情况下 3 个数据集

效果都不理想,在特征对齐是起到一定作用。数据交换模块(data exchange, DE)在CompCars数据集上提高最为明显,所有指标均大幅提高。实例级样本判别(instance discrimination, ID)在3个数据集上提高程度惊人。对比IMM+DE模块,算法在加上此模块后在CompCars、StanfordCars和Pix3D数据集上$A_{\text{Top-1}}$分别提高6.3%、7.5%、26.1%。分析原因是此模块通过将相同实例图像和模型分到一个类中能够大幅减小图像域和模型域距离提高检索精度。特征增强模块(feature enhancement, FE)在模型性能提高方面同样发挥着至关重要作用。尽管在CompCars和StanfordCars等数据集上,该模块能够带来显著性能提升,在处理Pix3D数据集时,其效果却相对有限。这主要是Pix3D数据集4个类别样本在特征表达上已经足够区分,特征增强模块在Pix3D上精度提高并不显著。

4 结论

本研究提出一个新的算法框架可以解决图像检索实例级别的三维模型。现有的算法大多使用对比损失和三元组损失函数。三元组损失函数本身存在局限,需解决最难负样本挖掘问题。对比损失函数对样本数量有要求。为解决以往算法不足,减小图像域和模型域间的域差距,算法提出实例级判别、图像模型配对、数据交换和特征增强等方法,可以检索出与查询图像最为相似三维模型。图像模型配对判断图像与模型是否为同一实例训练网络,初步减小域差距;数据交换在图像域与模型域间交换数据,让图像域网络具有模型域数据学习能力,模型域网络具有图像域数据学习能力,进一步减小域差距;实例判别将图像和模型分为个体实例,提取特征;特征增强在实例判别提取独立特征后增强特征间的辨别性,有利于减小域差距。试验结果表明,算法在常见数据集上效果非常显著,证实本研究方法可靠性和有效性。算法在三维模型数据量较少情况下虽取得良好效果,但若三维模型数量更多时,实例级损失会有影响,下一步将着手解决这一问题。

参考文献:

[1] WU Peng, LU Xiankai, SHEN Jianbing, et al. Clip fusion with bi-level optimization for human mesh reconstruction from monocular videos[C]//Proceeding of the 31st ACM International Conference on Multimedia. New York, USA: ACM, 2023:105-115.

[2] QIN Zheyun, HAN Cheng, WANG Qifan, et al. Unified 3D segmenter as prototypical classifiers[C]//Proceeding of the 37th International Conference on Neural Information Processing Systems. New York, USA: Curran Associates Inc., 2023:46419-46432.

[3] LIU Anan, ZHANG Chenyu, LI Wenhui, et al. Self-supervised auxiliary domain alignment for unsupervised 2D image-based 3D shape retrieval[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(12): 8809-8821.

[4] LI Tianbao, SU Yuting, SONG Dan, et al. Progressive fourier adversarial domain adaptation for object classification and retrieval[J]. IEEE Transactions on Multimedia, 2024, 26: 4540-4553.

[5] SONG Dan, YANG Yuanxiang, LI Wenhui, et al. Adaptive semantic transfer network for unsupervised 2D image-based 3D model retrieval[J]. Computer Vision and Image Understanding, 2024, 240(3): 1077-3142.

[6] DAI Yongxing, LIU Jun, SUN Yifan, et al. IDM: an intermediate domain module for domain adaptive person re-id[C]//Proceeding of the 20th International Conference on Computer Vision. Piscatawa, USA: IEEE, 2021: 11844-11854.

[7] FU Huan, LI Shunming, JIA Rongfei, et al. Hard example generation by texture synthesis for cross-domain shape similarity learning[C]//Proceeding of the 34th International Conference on Neural Information Processing Systems. New York, USA: Curran Associates Inc., 2020: 14675-14687.

[8] GRABNER A, ROTH P M, LEPETIT V. Location field descriptors: single image 3D model retrieval in the wild[C]//Proceeding of the 9th International Conference on 3D Vision. Quebec, Canada: IEEE, 2019: 583-593.

[9] LIN Mingxian, YANG Jie, WANG He, et al. Single image 3D shape retrieval via cross-modal instance and category contrastive learning[C]//Proceeding of the 20th International Conference on Computer Vision. Piscataway, USA: IEEE, 2021: 11385-11395.

[10] HE Kaiming, FAN Haoqi, WU Yuxing, et al. Momentum contrast for unsupervised visual representation learning[C]//Proceeding of the 33th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2020: 9726-9735.

[11] KHOSLA P, TETERWAK P, WANG C, et al. Supervised contrastive learning[C]//Proceeding of the 34th International Conference on Neural Information Processing Systems. New York, USA: Curran Associates Inc., 2020: 18661-18673.

[12] WU Z R, SONG S R, KHOSLA A, et al. 3D shapenets: a deep representation for volumetric shapes[C]//Proceeding of the 28th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA:

IEEE, 2015: 1912-1920.
[13] FURUYA T, OHBUCHI R. Deep aggregation of local 3D geometric features for 3D model retrieval[C]//Proceeding of the 2016 British Machine Vision Conference. York, UK: BMVC Press, 2016: 920--928.
[14] QI C R, SU H, NIEßNER M, et al. Volumetric and multi-view CNNs for object classification on 3D data [C]//Proceeding of the 29th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2016: 5648-5656.
[15] MATURANA D, SCHERER S. Voxnet: A 3d convolutional neural network for real-time object recognition [C]//Proceeding of the 2015 international conference on intelligent robots and systems. Piscataway, USA: IEEE, 2015: 922-928.
[16] QI C R, SU H, KAICHUN M, et al. Pointnet: deep learning on point sets for 3D classification and segmentation[C]//Proceeding of the 30th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2017: 77-85.
[17] QI C R, LI Y, SU H, et al. Pointnet++: deep hierarchical feature learning on point sets in a metric space[C]// Proceeding of the 31st International Conference on Neural Information Processing Systems. New York, USA: Curran Associates Inc., 2017: 5105-5114.
[18] MA Xu, QIN Can, YOU Haoxuan, et al. Rethinking network design and local geometry in point cloud: a simple residual MLP framework [C]//Proceeding of the 10th International Conference on Learning Representations. New York, USA: Curran Associates Inc., 2022: 661-673.
[19] WANG Yue, SUN Yongbin, LIU Ziwei, et al. Dynamic graph CNN for learning on point clouds [J]. ACM Transactions on Graphics, 2019, 38(5): 1-14.
[20] SU J C, GADELHA M, WANG R, et al. A deeper look at 3D shape classifiers[C]//Proceeding of the 15th European conference on computer vision. Heidelberg, Germany: Springer-Verlag, 2018: 645-661.
[21] FENG Yifan, ZHANG Zizhao, ZHAO Xibin, et al. GVCNN: group-view convolutional neural networks for 3D shape recognition[C]//Proceeding of the 31th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2018: 264-272.
[22] SU H, MAJI S, KALOGERAKIS E, et al. Multi-view convolutional neural networks for 3D shape recognition [C]//Proceeding of the 15th IEEE International Conference on Computer Vision. Piscataway, USA: IEEE, 2015: 945-953.
[23] ISAAC-MEDINA B, WILLCOCKS C, BRECKON T. Multi-view vision transformers for object detection [C]//Proceeding of the 26th International Conference on Pattern Recognition. Piscataway, USA: IEEE, 2022: 4678-4684.
[24] NIE Weizhi, ZHAO Yue, NIE Jie, et al. CLN: cross-domain learning network for 2D image-based 3D shape retrieval[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(3): 992-1005.
[25] LIU Anan, GUO Fubin, ZHOU Heyu, et al. Domain-adversarial-guided siamese network for unsupervised cross-domain 3-D object retrieval[J]. IEEE Transactions on Cybernetics, 2022, 52(12): 13862-13873.
[26] ZHOU Heyu, LIU Anan, NIE Weizhi. Dual-level embedding alignment network for 2D image-based 3D object retrieval[C]//Proceeding of the 27th ACM International Conference on Multimedia. New York, USA: ACM, 2019: 1667-1675.
[27] GRABNER A, ROTH P M, LEPETIT V. 3D pose estimation and 3D model retrieval for objects in the wild [C]//Proceeding of the 31th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2018: 3022-3031.
[28] SUN Xingyuan, WU Jiajun, ZHANG Xiuming, et al. Pix3D: dataset and methods for single-image 3D shape modeling[C]//Proceeding of the 31th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2018: 2974-2983.
[29] WANG Yaming, TAN Xiao, YANG Yi, et al. 3D pose estimation for fine-grained object categories[C]//Proceeding of of 16th European Conference on Computer Vision. Heidelberg, Germany: Springer-Verlag, 2019: 619-632.
[30] KRAUSE J, STARK M, DENG J, et al. 3D Object representations for fine-grained categorization [C]//Proceeding of the 14th IEEE International Conference on Computer Vision. Piscataway, USA: IEEE, 2013: 554-561.
[31] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//Proceeding of the 29th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2016: 770-778.
[32] AUBRY M, RUSSELL B C. Understanding deep features with computer-generated imagery [C]// Proceeding of the 15th IEEE International Conference on Computer Vision. Piscataway, USA: IEEE, 2015: 2875-2883.
[33] XUE Le, GAO M, MARTIN M R, et al. Ulip: learning a unified representation of language, images, and point clouds for 3d understanding[C]//Proceeding of the 36th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2023:1179-1189.

(编辑:陈燕)