文章编号:1672-3961(2025)04-0048-08 DOI:10.6040/j.issn.1672-3961.0.2024.165

基于注意力和视图信息的单图三维模型检索

韩小凡[1,2],刁振宇[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:为提取有效特征描述符,减小图像和三维模型巨大差异,提出一种基于注意力和视图信息的方法。该方法在模型特征提取模块引入空间注意力机制,提高模型特征描述符的有效性;将三维模型二维视图引入到查询图像特征学习过程中,缩小图像域与模型域域间差异。在 Pix3D、Comp Cars、Stanford Cars 3 个代表性基准数据集进行试验,结果表明检索精度较现有经典方法提高约 5%。所提出方法能够使单幅图像有效检索相似三维模型,提高检索准确率。

关键词:三维模型;单幅图像;空间注意力;基于单图的三维模型检索;域间差异
中图分类号:TP183 **文献标志码**:A
引用格式:韩小凡,刁振宇,张承宇,等.基于注意力和视图信息的单图三维模型检索[J].山东大学学报(工学版),2025,55(4):48-55.
HAN Xiaofan, DIAO Zhenyu, ZHANG Chengyu, et al. Single image 3D model retrieval based on attention and view information[J]. Journal of Shandong University (Engineering Science), 2025, 55(4):48-55.

Single image 3D model retrieval based on attention and view information

HAN Xiaofan[1,2], DIAO Zhenyu[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 extract effective feature descriptors and reduce the significant differences between 2D images and 3D models, a method based on attention and view information was proposed. The method introduced a spatial attention mechanism into the model's feature extraction module to enhance the effectiveness of the model's feature descriptors. The 2D views of 3D models were incorporated into the process of learning query image features to reduce the domain gap between the image domain and the model domain. Experiments were conducted on three representative benchmark datasets: Pix3D, Comp Cars, and Stanford Cars. The results showed that the best retrieval accuracy improved by 5%. The proposed method effectively retrieved similar 3D models from a single image and improved the retrieval accuracy.

Keywords: 3D model; single image; spatial atention; 3D model retrieval based on single image; domain gap

0 引言

近年来,许多学者对基于单图的三维模型检索进行深入研究,取得显著成果[1-3]。为更好表示三维模型,目前大量研究工作将三维模型处理成多张视图形式表示[4-6]。这种处理方式可以最大程度呈现三维模型,有效减小域差距。为进一步解决域差距问题,一些研究采用对比学习方法进行度量学习[7]。相对于传统三元损失,对比学习可以解决挖掘最难正负样本问题。

基于单图的三维模型检索涉及两个关键性问题:(1)如何从二维图像和三维模型中提取出有效且高质量的特征描述符。(2)二维图像和三维模型之间存在巨大域差距。为更好解决这两个关键性问题,继续采用三维模型不同角度视图表示三维模

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(2023TSGC0244)
第一作者简介:韩小凡(2000—),女,山东滨州人,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:202221200983@ stu.ujn.edu.cn
*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

型,通过对比学习进行度量学习。提出在模型特征提取模块引入空间注意力机制,提高模型特征描述符有效性,为后续检索工作提供坚实基础。提出将三维模型二维视图引入到查询图像特征学习过程中,缩小图像域和模型域的域间差异。

1　相关工作

1.1　三维模型检索

目前三维模型检索的研究主要分为两个方向,基于模型的三维模型检索和基于图像的三维模型检索。

基于模型的三维模型检索旨在根据给定三维模型,检索出一系列相似的三维模型。大多数基于模型的三维模型检索方法是提取出色的三维模型特征,通过衡量三维模型之间的相似度完成检索。为提取出色的三维模型特征,文献[8]提出由连续主成分分析、傅里叶描述子和 Zernike 矩构成的方法联合表示三维模型。为更好衡量三维模型之间相似度,文献[9]提出联合成本函数方法,通过平衡不同模块获取的两个距离,解决局部和全局特征之间不相关性问题。这些方法通常都需要一个已知三维模型进行查询,三维模型在实际应用中获取困难。

另一个主要方向是基于图像的三维模型检索,旨在根据给定真实图像,检索出一系列相似的三维模型。这一方向主要分为有监督和无监督两种方法。无监督方法不需要二维图像或三维模型标签,检索结果往往聚焦类别级别且检索精度有限。为缩小域间差距,文献[10]提出无监督双层嵌入对齐网络,将三维模型每张视图通过跨域视图注意机制自适应计算权重,聚合成一个紧凑描述符,缩小源域和目标域之间的差距。文献[11]提出深度相关联合网络,基于判别损失和相关损失联合学习两个不同的深度神经网络。为实现更好聚类效果,文献[12]提出改进多聚类语义表示学习方法,为三维模型产生更可靠伪标签。

相较于无监督检索方法,有监督检索方法需要给定图像和三维模型标签,检索精度往往更高且通常聚焦实例级别检索。很多数据集已经提供了相应标签信息,为有监督检索方法奠定了良好基础。在实际应用中,实例级别检索更具有应用价值。本研究将关注实例级别的基于图像的三维模型检索。为缩小域间差距,文献[13]提出通过三元损失进行相似性度量学习。文献[14]提出位置场描述符(location field descriptors, LFD),利用位置场信息将三维模型和二维图像嵌入到共同的低级表示空间。文献[15]提出通过三维模型纹理合成模块生成负样本(hard example generation by texture synthesis, HEG-TS),改进跨域相似性度量学习。这些方法都是基于三元损失进行度量学习。三元损失需要挖掘最难正负样本,文献[7]采用对比学习进行度量学习,取得了良好效果。文献[16]提出预训练框架,该框架可以学习图像、文本和点云的统一表示(learning a unified representation of language, images, and point clouds, ULIP),在相同的特征空间对齐域间差异。为更好表示三维模型,文献[13]提出通过基于八叉树和多视图图像的方法表示三维模型。目前很多方法将三维模型处理成多个角度视图表示,但大部分都没有关注提升模型特征描述符的有效性,提出在模型特征提取模块引入空间注意力机制,提高模型特征描述符的有效性。

1.2　对比学习

对比学习已经在提高表示学习的性能方面取得显著进展[17-22],引起广泛关注。对比学习的核心思想是将样本映射到一个共同嵌入空间中,推动相似的样本更近,不同的样本更远。

对比学习广泛应用在许多领域。在聚类任务中,文献[23]提出将属于同一类点簇拉到一起,将来自不同类别样本分开。文献[24]提出对比多视图学习方法网络,对多个未标记的三维模型进行聚类。在预测任务中,文献[25]提出通过预测两个增强图像是否来自同一原始图像,学习图像编码器。在图像分类任务中,文献[26]提出通过使用强大的自回归模型,在潜在空间中预测未来学习表示。文献[27]提出在预训练过程借助有监督对比损失,有效提高模型泛化性能。在三维模型检索问题中,文献[28]提出采用对比学习将特征对齐到同一子空间,获得更高判别性融合特征。在基于图像的三维模型检索任务中,采用对比学习缩小图像和三维模型域间差异取得了良好效果。本研究继续采用对比学习进行度量学习,提出将三维模型二维视图引入到查询图像特征学习过程中,缩小图像域和模型域域间差异。

2　试验方法

本研究提出了一种基于注意力和视图信息方法。该方法在模型特征提取模块引入空间注意力机制,提高模型特征描述符有效性;将三维模型二维视图引入到查询图像特征学习过程中,缩小图像

域和模型域域间差异。图 1 为本研究的整体框架图。三维模型通过平面着色转换为多视角的灰度图像[7]。通过视图补充模块将三维模型二维视图引入到查询图像特征学习过程中；提取图像和三维模型的特征描述符；在类别和实例两个级别分别缩小图像和三维模型之间距离。

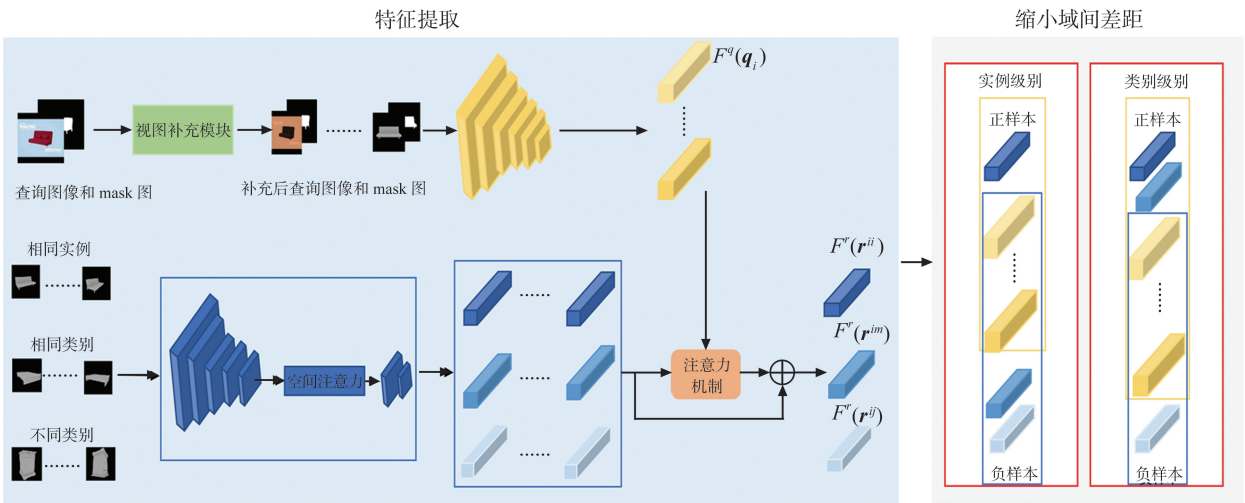

图 1 整体框架图

Fig.1 Overall framework diagram

2.1 视图补充模块

在该节中具体介绍视图补充模块。如图 2 所示，给定一个查询图像，随机选择 n 张和该查询图像具有相同实例标签的三维模型二维视图，设定随机选择视图和给定查询图像具有相同的实例和类别标签。随机选取视图颜色空间较为单一，本研究只对给定查询图像进行颜色转换。在训练样本中随机选取一张图像，通过颜色转换机制将随机选取图像颜色转换到给定查询图像上[29]，达到扩大查询图像颜色空间和增加训练样本多样性目的。本研究在随机选取视图的同时，视图通过预处理提取相应 mask 图。在后续训练过程中，随机选择的视图被用作查询图像样本进行训练。

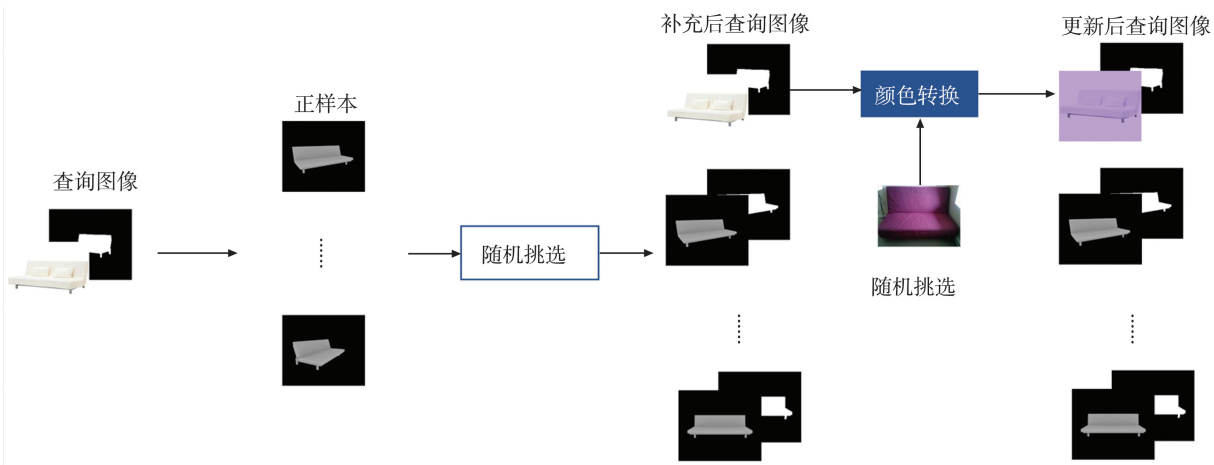

图 2 视图补充模块的具体细节

Fig.2 Specific details of the view supplement module

2.2 特征提取

假设数据集有 N 个实例，其中第 i 个实例 N_i 由查询图像 $\boldsymbol{q}_i$，查询图像对应的 mask 图 $\boldsymbol{m}_i$，三维模型 $\boldsymbol{s}_i$，以及语义标签 $\boldsymbol{y}_i$ 组成。其中每个三维模型由 K 个灰度图像表示 $\{\boldsymbol{r}_k^i\}_{k=1}^K$。

经过视图补充模块后，对 $\boldsymbol{q}_i$ 和 $\{\boldsymbol{r}_k^i\}_{k=1}^K$ 进行特征提取。将给定的 $\boldsymbol{q}_i$ 和相应的 $\boldsymbol{m}_i$ 合并通过图像特征提取器 F^q 提取查询图像特征表示 $F^q(\boldsymbol{q}_i)$。将三维模型多张视图分别通过模型特征提取器 F^r 提取每张视图特征表示 $F^r(\boldsymbol{r}_k^i)$。其次采用注意力机制，通过 Softmax 函数为每个三维模型的每张视图分配一个关于查询图像特征 $F^q(\boldsymbol{q}_i)$ 的权重 $\{\boldsymbol{\alpha}_k^{ii}\}_{k=1}^K$，将每张视图进行加权求和获得关于第 i 个查询图像的三维模型特征表示 $F^r(\boldsymbol{r}^{ii})$。

在提取模型特征时,希望提高模型特征描述符的有效性。本研究在提取模型特征时引入空间注意力机制。模型特征提取器以 ResNet18 为基础网络[30]。在第 4 个卷积组后引入空间注意力机制。

空间注意力机制旨在通过引入注意力模块[31],使模型能够自适应学习不同区域注意力权重。如图 3 所示,获得第 4 个卷积组特征 $\boldsymbol{F}$,根据式(1)(2),特征 $\boldsymbol{F}$ 在同一个通道维度分别进行最大池化和平均池化,得到 $\boldsymbol{F}_1 \in \mathbf{R}^{1\times H\times W}$, $\boldsymbol{F}_2 \in \mathbf{R}^{1\times H\times W}$。根据式(3),将 $\boldsymbol{F}_1$ 和 $\boldsymbol{F}_2$ 拼接经过卷积核大小为 7×7 的卷积运算,通过激活函数 Sigmoid 得到权重系数 $\boldsymbol{M}_s$。根据式(4),输入的 $\boldsymbol{F}$ 与 $\boldsymbol{M}_s$ 相乘得到缩放后新特征:

$$\boldsymbol{F}_1 = \mathrm{AvgPool}(\boldsymbol{F}), \tag{1}$$

$$\boldsymbol{F}_2 = \mathrm{MaxPool}(\boldsymbol{F}), \tag{2}$$

$$\boldsymbol{M}_s = \sigma(f^{7\times 7}([\boldsymbol{F}_1;\boldsymbol{F}_2])), \tag{3}$$

$$\boldsymbol{F}_{sa} = \boldsymbol{M}_s\boldsymbol{F}, \tag{4}$$

式中,σ 为 Sigmoid 激活函数,$f^{7\times 7}$ 为卷积核大小为 7×7 的卷积运算,AvgPool 表示平均池化,MaxPool 表示最大池化。

在该模块引入空间注意力机制,提高了模型特征描述符有效性。为后续拉近图像和三维模型之间距离提供了良好前提。

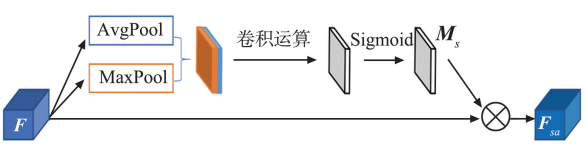

图 3　空间注意力

Fig.3　Spatial attention

2.3　跨域对比学习

完成特征提取,分别在类别级别和实例级别对齐图像和三维模型特征。在实例级别,利用对比学习拉近查询图像和具有相同实例标签的三维模型之间的距离,推远查询图像和具有不同实例标签的三维模型之间的距离,每个查询图像都只有一个正样本:

$$L_{\mathrm{inst}} = -\sum_{i\in 1}^{B}\left(\ln\frac{\exp(F^q(\boldsymbol{q}_i)\cdot F^r(\boldsymbol{r}^{ii})/\tau)}{\sum_{j\in 1}^{B}\exp(F^q(\boldsymbol{q}_i)\cdot F^r(\boldsymbol{r}^{ij})/\tau)}\right), \tag{5}$$

式中,B 为训练批次大小,τ 为超参数,$F^q(\boldsymbol{q}_i)$ 为查询图像特征向量,$F^r(\boldsymbol{r}^{ii})$ 为 $F^q(\boldsymbol{q}_i)$ 的正样本,$F^r(\boldsymbol{r}^{ij})$ 为 $F^q(\boldsymbol{q}_i)$ 的负样本。

在类别级别,利用对比学习不断拉近查询图像和同一类别三维模型之间的距离,推远查询图像和不同类别三维模型之间的距离,每个查询图像有多个正样本和多个负样本:

$$L_{\mathrm{cats}} = -\sum_{i\in 1}^{B}\frac{1}{N_m}\sum_{\substack{m\in 1,\\ y_m=y_i}}^{B}\left(\ln\frac{\exp(F^q(\boldsymbol{q}_i)\cdot F^r(\boldsymbol{r}^{im})/\tau)}{\sum_{j\in 1}^{B}\exp(F^q(\boldsymbol{q}_i)\cdot F^r(\boldsymbol{r}^{ij})/\tau)}\right), \tag{6}$$

式中,N_m 表示和查询图像属于相同类别的模型的个数,$F^q(\boldsymbol{q}_i)$ 为查询图像特征向量,$F^r(\boldsymbol{r}^{im})$ 表示与 $F^q(\boldsymbol{q}_i)$ 属于相同类别的的正样本,$F^r(\boldsymbol{r}^{ij})$ 表示与 $F^q(\boldsymbol{q}_i)$ 属于不同类别的的负样本。

总体损失函数如下:

$$L = L_{\mathrm{inst}} + \beta L_{\mathrm{cats}}, \tag{7}$$

式中 β 表示超参数。

3　试验细节与结果展示

3.1　试验设置

本研究分别对图像特征提取器和模型特征提取器应用 ResNet50 和 ResNet18[30],其中第一个卷积层的通道数量分别修改为 4 个和 1 个,特征输出维度为 128。生成查询图像对应的 mask 采用预先训练好的 MaskR-CNN[32] 和 OCRNet[33]。

查询图像、mask 图像和渲染图像都调整到 224 像素×224 像素的大小。每个三维模型包含 12 张渲染图像。设置视图补充模块随机选取 2 张视图。采用 Adam 优化器,学习率为 5×10^{-5},betas 为(0.5,0.999)。设置 Epoch 和 BatchSize 为 400 和 10,超参数 β 和 τ 为 0.2 和 0.1。

3.2　数据集

为了评估本研究的有效性,在 Pix3D[34]、CompCars 和 StanfordCars[35] 3 个数据集上进行试验并遵循文献[7,15,16]中的试验设置和评估指标。

Pix3D 数据集由 9 个类别组成。根据文献[7]的方法,本研究仅在 4 个类别(即床、椅子、沙发和桌子)上进行试验。这 4 个类别共包含 5 118 张查询图像和 322 个三维模型,其中 2 648 张用于训练,2 470 张用于评价。Comp Cars 和 StanfordCars 专注更具挑战性的细粒度检索任务。在 Comp Cars 中,共有 94 个汽车模型和 5 696 张查询图像,其中3 798 张用于训练,1 898 张用于评价。在 Stanford Car 中,共有 134 个汽车模型和 16 185 张查询图像,其中 8 144 张用于训练,8 041 张用于测试。

3.3　比较结果

为了更好展示比较结果,采用 $A_{\mathrm{ccTop\text{-}1}}$、$A_{\mathrm{ccTop\text{-}10}}$、$d_{\mathrm{HAU}}$、$d_{\mathrm{IoU}}$ 4 个指标进行评估。$A_{\mathrm{ccTop\text{-}1}}$ 和 $A_{\mathrm{ccTop\text{-}10}}$ 表示

第1个和前10个预测形状的准确性。d_{HAU}和d_{IoU}用来测量2个三维模型之间的距离。评估指标的详细解释请参见文献[7,15,16]。与现有经典方法的比较结果显示在表1~3中。本研究在Pix3D数据集上效果有大幅度提升,针对单独类别,除椅子类别外所有评估指标均优于现有经典方法。椅子类别,本研究的d_{HAU}、d_{IoU}均优于现有经典方法,$A_{ccTop-1}$、$A_{ccTop-10}$均优于文献[7],与现有经典方法中最好结果也十分相近。在CompCars和Stanford Cars数据集上效果也有着明显提升。

表1 Pix3D数据集的试验结果

Table 1 Experimental results on the Pix3D dataset

类别	方法	A_{ccTop1}/%	$A_{ccTop10}$/%	d_{HAU}	d_{IoU}
bed	UDF-CGI[36]	19.40	46.60	0.082 1	0.339 7
	Grabneret al[37]	35.10	83.20	0.038 5	0.559 8
	LFD	64.40	89.00	0.015 2	0.807 4
	HEG-TS	65.30	95.40	0.012 2	0.821 3
	Linet al	73.30	96.10	0.009 3	0.892 7
	ULIP	74.20	96.30	0.006 8	0.893 1
	本研究方法	**77.70**	**98.90**	**0.002 7**	**0.922 6**
chair	UDF-CGI	17.30	49.10	0.055 9	0.302 7
	Grabneret al	41.30	73.90	0.030 5	0.546 9
	LFD	58.10	81.80	0.017 0	0.716 9
	HEG-TS	**87.90**	**97.90**	0.004 1	0.906 3
	Linet al	79.40	96.30	0.008 0	0.866 1
	ULIP	83.70	97.40	0.005 2	0.883 2
	本研究方法	87.80	97.80	**0.003 2**	**0.906 9**
sofa	UDF-CGI	21.70	52.20	0.050 3	0.382 4
	Grabneret al	44.10	89.90	0.019 7	0.776 2
	LFD	67.00	94.40	0.007 5	0.902 8
	HEG-TS	72.80	97.70	0.004 7	0.907 0
	Linet al	80.70	97.10	0.004 5	0.932 9
	ULIP	81.50	97.50	0.004 8	0.934 7
	本研究方法	**84.80**	**98.70**	**0.002 0**	**0.952 0**
table	UDF-CGI	12.00	34.20	0.100 3	0.171 5
	Grabneret al	33.90	66.10	0.060 7	0.450 0
	LFD	53.30	80.10	0.028 8	0.638 3
	HEG-TS	73.70	92.40	0.017 0	0.766 7
	Linet al	76.90	93.50	0.016 8	0.808 8
	ULIP	78.00	94.00	0.015 2	0.791 3
	本研究方法	**78.40**	**95.70**	**0.008 9**	**0.809 2**
mean	UDF-CGI	17.60	45.50	0.072 2	0.299 1
	Grabneret al	38.60	78.30	0.037 4	0.583 2
	LFD	60.70	86.30	0.017 1	0.766 3
	HEG-TS	74.90	95.80	0.009 5	0.850 3
	Linet al	78.90	96.10	0.008 6	0.874 6
	ULIP	79.30	96.30	0.008 0	0.875 1
	本研究方法	**85.10**	**97.80**	**0.003 7**	**0.904 0**

注:黑体字为该列最优效果。

表2 Comp Cars数据集的试验结果

Table 2 Experimental results on the Comp Cars dataset

类别	方法	A_{ccTop1}/%	$A_{ccTop10}$/%	d_{HAU}	d_{IoU}
car	UDF-CGI	2.40	18.20	0.020 7	0.722 4
	Grabneret al	10.20	36.90	0.015 8	0.780 5
	LFD	20.50	58.00	0.013 3	0.814 2
	HEG-TS	67.10	93.70	0.003 5	0.925 6
	Linet al	77.80	94.10	0.002 3	**0.939 9**
	ULIP	**78.90**	94.30	0.002 3	0.937 2
	本研究方法	78.60	**95.20**	**0.002 1**	0.938 1

注:黑体字为该列最优效果。

表3 Stanford Cars数据集的试验结果

Table 3 Experimental results on the Stanford Cars dataset

类别	方法	A_{ccTop1}/%	$A_{ccTop10}$/%	d_{HAU}	d_{IoU}
car	UDF-CGI	3.70	20.10	0.019 8	0.716 9
	Grabneret al	11.30	42.20	0.015 3	0.772 1
	LFD	29.50	69.40	0.011 0	0.835 2
	HEG-TS	68.40	92.10	0.003 4	0.921 0
	Linet al	83.40	96.40	**0.002 1**	0.943 1
	ULIP	84.30	96.70	0.002 3	0.946 7
	本研究方法	**85.10**	**96.80**	0.002 3	**0.949 3**

注:黑体字为该列最优效果。

3.4 消融试验

3.4.1 关于空间注意力位置的消融试验

为了验证本研究选择空间注意力位置的合理性,在Pix3D和Comp Cars数据集上分别将空间注意力应用于ResNet18的第1个卷积层后、卷积组之间、第4个卷积组后进行消融试验.

表4的结果表明,对于两个数据集,空间注意力应用于ResNet18第4个卷积组后的检索效果会优于应用于其他位置。该消融试验说明获取ResNet18的最后一个卷积组的信息之后应用空间注意力会使得检索效果更好。进一步验证了研究空间注意力应用位置的合理性。

表4 空间注意力位置的消融试验结果

Table 4 Ablation results for spatial attention

数据集	类别	位置	A_{ccTop1}/%
Pix3D	mean	第1个卷积层后	82.00
		卷积组之间	82.00
		第4个卷积组后	**82.90**
Comp	car	第1个卷积层后	77.70
		卷积组之间	77.10
		第4个卷积组后	**78.00**

注:黑体字为该列最优效果。

3.4.2 关于视图数量的消融试验

为了验证本研究在视图补充模块所设置的选择视图张数的合理性,本研究在Pix3D和Comp

Cars 数据集上分别应用视图补充模块时随机选择 2、3、4 张视图进行消融试验。

对于 Pix3D 据集,表 5 的结果表明,随机选取 4 张视图检索效果优于随机选取 2 张和 3 张视图。对于 Comp Cars 数据集,表 5 的试验结果表明,随机选取 2 张视图的检索效果优于随机选取 3 张和 4 张视图的检索效果。在 Pix3D 数据集上,随机选取 2 张视图的检索效果和随机选取 4 张视图的检索效果相差不大。考虑到多类数据集和单类数据集的综合检索效果,随机选取 2 张视图展示出了一定的优越性。进一步验证了本研究在视图补充模块设置随机选取 2 张视图进行训练的合理性。

表 5　随机选取的视图张数的消融试验结果

Table 5　Ablation results for the number of attempts chosenat random

数据集	类别	视图数量/张	A_{ccTop1}/%
Pix3D	mean	2	85.10
		3	84.60
		4	**85.30**
Comp	car	2	**78.60**
		3	76.10
		4	77.30

注:黑体字为该列最优效果。

3.4.3　关于空间注意力和视图补充模块有效性的消融试验

为了验证空间注意力和视图补充模块的有效性,在 Pix3D 和 Comp Cars 数据集上分别进行了消融试验。采用未加入空间注意力和视图补充模块的检索方法作为消融试验的基准[7]。

在两个数据集上应用空间注意力的检索效果见表 6。表 6 的结果表明应用空间注意力使检索精度相较于现有经典方法有一定程度的提高。通过该消融试验可以验证空间注意力机制应用的有效性。

表 6　空间注意力和视图补充模块有效性的消融试验结果

Table 6　Ablation results for effectiveness of spatial attention and view complement modules

数据集	类别	方法	A_{ccTop1}/%
Pix3D	mean	基准	78.90
		基准+空间注意力	82.90
		基准+空间注意力+视图补充	**85.10**
Comp	car	基准	77.80
		基准+空间注意力	78.00
		基准+空间注意力+视图补充	**78.60**

注:黑体字为该列最优效果。

在两个数据集上应用视图补充模块的检索效果见表 6。表 6 的结果表明应用视图补充模块可以使检索精度进一步提高。无论是多类别还是单类别数据集,视图补充模块的应用都可以使检索效果变得更优越。

空间注意力模块和视图补充模块的应用都能够在一定程度上提高检索精度,这也证实了本研究的优越性。

4　结论

本研究提出一种基于注意力和视图信息的单幅图像三维模型检索方法。该方法能够显著提高特征描述符有效性,有效缩小图像和三维模型域间差距。在模型特征提取模块引入空间注意力机制,关注特征重要信息,提高模型特征描述符的有效性,为后续缩小域间差距和检索工作提供良好前提;将三维模型二维视图引入到查询图像特征学习过程中,进一步提高拉近图像域与模型域之间的域间差距的效果。试验结果表明,该方法效果优于现有经典方法。提取特征描述符借助空间注意力,拉近图像域和模型域的差距借助三维模型视图信息都显著提高了检索精度。本研究对单幅图像可以有效检索到相似三维模型,检索效果体现了一定的优越性。

参考文献:

[1] MU P P, ZHANG S Y, ZHANG Y, et al. Image-based 3D model retrieval using manifold learning[J]. Frontiers of Information Technology & Electronic Engineering, 2018, 19(11): 1397-1408.

[2] HU N, ZHOU H Y, LIU A A, et al. Collaborative distribution alignment for 2D image-based 3D shape retrieval [J]. Journal of Visual Communication and Image Representation, 2022, 83: 103426.

[3] ZHOU H Y, NIE W Z, SONG D, et al. Semantic consistency guided instance feature alignment for 2D image-based 3D shape retrieval[C]//Proceedings of the 28th ACM International Conference on Multimedia. New York, USA: ACM, 2020: 925-933.

[4] HE X W, HUANG T T, BAI S, et al. View n-gram network for 3D object retrieval[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 7514-7523.

[5] LIN D Y, LI Y Q, CHENG Y, et al. Multi-view 3D object retrieval leveraging the aggregation of view and instance attentive features[J]. Knowledge-Based Systems, 2022, 247: 108754.

[6] ALZU′BI A, ABUARQOUB A, AL-HOMOUZ A. Aggregated deep convolutional neural networks for multi-view 3D object retrieval[C]//Proceedings of 2019 11th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops. Dublin, Ireland: IEEE, 2019: 1-5.

[7] LIN M X, YANG J, WANG H, et al. Single image 3D shape retrieval via cross-modal instance and category contrastive learning[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE, 2021: 11385-11395.

[8] LIN S F, WU C, HSU C, et al. An efficient 3D model retrieval based on principal axes analysis and feature integration[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2011, 25(4): 583-604.

[9] PAN X Q, CHEN Y R, KUO C C. 3D shape retrieval via irrelevance filtering and similarity ranking (IF/SR)[C]//Proceedings of Computer Vision-ACCV 2016 Workshops. Taipei, China: Springer, 2017: 630-646.

[10] ZHOU H Y, LIU A A, NIE W Z. Dual-level embedding alignment network for 2D image-based 3D object retrieval[C]//Proceedings of the 27th ACM International Conference on Multimedia. New York, USA: ACM, 2019: 1667-1675.

[11] NIE W Z, LIU A A, ZHAO S C, et al. Deep correlated joint network for 2D image-based 3D model retrieval[J]. IEEE Transactions on Cybernetics, 2020, 52(3): 1862-1871.

[12] CHU J H, ZHAO X Q, SONG D, et al. Improved semantic representation learning by multiple clustering for image-based 3D model retrieval[J]. International Journal on Semantic Web and Information Systems, 2022, 18(1): 1-20.

[13] ZOU Q F, LIU L G, LIU Y. Instance-level 3D shape retrieval from a single image by hybrid-representation-assisted joint embedding[J]. The Visual Computer, 2021, 37(7): 1743-1756.

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

[15] FU H, LI S M, JIA R F, et al. Hard example generation by texture synthesis for cross-domain shape similarity learning[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. New York, USA: Curran Associates Inc, 2020: 14675-14687.

[16] XUE L, GAO M F, XING C, et al. Ulip: learning unified representation of language, image and point cloud for 3D understanding[C]//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE, 2023: 1179-1189.

[17] HE K M, FAN H Q, WU Y X, et al. Momentum contrast for unsupervised visual representation learning[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020: 9729-9738.

[18] HJELM R D, FEDOROV A, LAVOIE-MARCHILDON S, et al. Learning deep representations by mutual information estimation and maximization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(1): 722-737.

[19] LÖWE S, O′CONNOR P, VEELING B. Putting an end to end-to-end: gradient-isolated learning of representations[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. New York, USA: Curran Associates Inc, 2019: 3039-3051.

[20] MISRA I, MAATEN L V. Self-supervised learning of pretext-invariant representations[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020: 6707-6717.

[21] TIAN Y, KRISHNAN D, ISOLA P. Contrastive multiview coding[C]//Proceedings of Computer Vision-ECCV 2020: 16th European Conference. Glasgow, UK: Springer-Verlag, 2020: 776-794.

[22] WU Z F, WANG S N, GU J T, et al. Clear: contrastive learning for sentence representation[J]. ACM Transactions on Intelligent Systems and Technology, 2020, 14(4): 1-34.

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

[24] PENG B, LIN G, LEI, J, et al. Contrastivemulti-view learning for 3D shape clustering[J]. IEEE Transactions on Multimedia, 2024, 26: 6262-6272.

[25] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations[DB/OL]. (2020-02-13)[2020-03-30]. https://doi.org/10.48550/arXiv.2002.05709

[26] OORD A V, LI Y, VINYALS O. Representation learning with contrastive predictive coding[DB/OL]. (2018-07-10)[2019-01-22]. https://doi.org/10.48550/arXiv.1807.03748

[27] SUN J P, LEI S. A study of few-shot image classification model based on contrastive learning and self-attention[C]//Proceedings of 2023 IEEE

International Conference on Electrical, Automation and Computer Engineering. Changchun, China: IEEE, 2023: 1142-1148.

[28] CHEN Q, CHEN Y N. Multi-view 3D model retrieval based on enhanced detail features with contrastive center loss[J]. Multimedia Tools and Applications, 2022, 81(8): 10407-10426.

[29] REINHARD E, ADHIKHMIN M, GOOCH B, et al. Color transfer between images[J]. IEEE Computer Graphics and Applications, 2001, 21(5): 34-41.

[30] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016: 770-778.

[31] WOO S, PARK J, LEE J Y, et al. Cbam: convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision. Munich, Germany: Springer-Verlag, 2018: 3-19.

[32] HE K M, GKIOXARI G, DOLLÁR P, et al. Mask r-cnn[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 2961-2969.

[33] YUAN Y H, CHEN X L, WANG J D. Object-contextual representations for semantic segmentation[C]//Proceedings of the Computer Vision-ECCV 2020: 16th European Conference. Glasgow, UK: Springer-Verlag, 2020: 173-190.

[34] SUN X Y, WU J J, ZHANG X M, et al. Pix3d: dataset and methods for single-image 3D shape modeling[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake, USA: IEEE, 2018: 2974-2983.

[35] WANG Y M, TAN X, YANG Y, et al. 3D pose estimation for fine-grained object categories[C]//Proceedings of Computer Vision-ECCV 2018 Workshops. Cham, Switzerland: Springer-Verlag, 2019: 619-632.

[36] AUBRY M, RUSSELL B C. Understanding deep features with computer-generated imagery[C]//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015: 2875-2883.

[37] GRABNER A, ROTH P M, LEPETIT V. 3D pose estimation and 3D model retrieval for objects in the wild[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake, USA: IEEE, 2018: 3022-3031.

(编辑:陈燕)

(上接第 47 页)

[23] SCHLEGL T, SEEBÖCK P, WALDSTEIN S M, et al. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery[C]// Information Processing in Medical Imaging. Orlando, USA: Springer International Publishing, 2017: 146-157.

[24] BERGMANN P, LÖWE S, FAUSER M, et al. Improving unsupervised defect segmentation by applying structural similarity to autoencoders[EB/OL]. (2018-07-05)[2024-03-12]. https://arxiv.org/abs/1807.02011v3

[25] RUFF L, VANDERMEULEN R, GOERNITZ N, et al. Deep one-class classification[C]// Proceedings of the international conference on machine learning. Stockholm, Sweden: ICML, 2018: 4393-4402.

[26] DEHAENE D, FRIGO O, COMBREXELLE S, et al. Iterative energy-based projection on a normal data manifold for anomaly localization[EB/OL]. (2020-02-10)[2024-03-12]. https://arxiv.org/abs/2002.03734v1

[27] SALEHI M, SADJADI N, BASELIZADEH S, et al. Multiresolution knowledge distillation for anomaly detection[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021: 14902-14912.

[28] NAPOLETANO P, PICCOLI F, SCHETTINI R. Anomaly detection in nanofibrous materials by CNN-based self-similarity[J]. Sensors, 2018, 18(1): 209-224.

[29] DAI Y M, GIESEKE F, OEHMCKE S, et al. Attentional feature fusion[C]//2021 IEEE Winter Conference on Applications of Computer Vision (WACV). Waikoloa, Hawaii, USA:IEEE, 2021: 3560-3569.

(编辑:熊小原)