中图法分类号:TN9;TP18;TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2025)12-3927-14

论文引用格式:Chen S L, Huang R Y, Huang S X, Chen Y and Li Q. 2025. Transformer attention-guided optimal view selection and classification for 3D models. Journal of Image and Graphics, 30(12):3927-3940(陈松乐, 黄茹玥, 黄思轩, 陈怡, 李骞. 2025. Transformer注意力引导的三维模型最优视图选择与分类方法. 中国图象图形学报, 30(12):3927-3940)[DOI:10.11834/jig.250037]

Transformer注意力引导的三维模型最优视图选择与分类方法

陈松乐[1,2],黄茹玥[1],黄思轩[1],陈怡[3],李骞[4*]

1. 南京邮电大学江苏省邮政大数据技术与应用工程研究中心,南京　210003; 2. 南京大学计算机软件新技术国家重点实验室,南京　210023; 3. 南京审计大学数字经济系,南京　211815; 4. 国防科技大学气象海洋学院,长沙　411107

摘　要:**目的**　现有的基于多视图的三维模型分类方法通常基于预设的多个视点渲染三维模型,然后将所有渲染的视图送入神经网络模型实现分类。显然由于冗余和无效视图的存在,每个视图对于分类目标的作用并不相同。选择对分类目标贡献大的视图,不仅有利于提高基于多视图的三维模型分类的性能,而且能够提供表征三维模型的代表性视图。**方法**　提出一种Transformer注意力引导的三维模型最优视图选择与分类方法。在从正十二面体20个视角对待预测的三维模型渲染后,首先采用卷积神经网络从多个视图提取特征信息,获得多视图局部特征Token序列,并对其进行位置编码,以保留其空间位置信息。随后,将可学习的全局分类Token与多视图特征Token序列合并,输入至Transformer编码器进行全局视图特征融合,获得初始全局分类特征。接下来,最优视图选择模块基于全局视图特征融合过程中的注意力得分矩阵计算各视图对初始全局分类Token的贡献,并选择得分高的视图作为最优视图。最后,将最优视图特征Token序列与初始全局分类Token拼接后输入到Transformer编码器进行最优视图融合,并获得最终的全局分类Token,将其输入分类预测模块获得最终分类概率,并输出选择的最优视图。本文在训练过程中采用了随机丢弃视图和对比学习策略,以进一步提高模型的泛化性能。**结果**　在ModelNet40基准数据集上,所提方法总体识别精度和平均识别精度分别为97.61%和96.36%,在达到当前先进分类水平的同时,基于Transformer注意力得分矩阵选择出的最优视图更具有表征性。**结论**　本文方法利用Transformer实现不同视图特征之间的融合,通过自注意力、残差连接以及多层堆叠机制,Transformer能够有效学习数据的复杂特征,并捕捉不同视图之间的全局上下文关系。同时,其注意力得分矩阵为最优视图选择提供了依据,在实现高效分类的同时,能够选择出最具有表征性的视图。

关键词:三维模型分类;Transformer;最优视图选择;对比学习;多视图学习

Transformer attention-guided optimal view selection and classification for 3D models

Chen Songle[1,2], Huang Ruyue[1], Huang Sixuan[1], Chen Yi[3], Li Qian[4*]

1. Jiangsu Provincial Postal Big Data Technology and Application Engineering Research Center, Nanjing University of Posts and

收稿日期:2025-02-12;**修回日期**:2025-04-28;**预印本日期**:2025-05-05

*通信作者:李骞　public_liqian@163.com

基金项目:软件新技术国家重点实验室创新基金项目(KFKT2022B19);江苏省高校自然科学基金项目(19KJB520047);南京邮电大学自然科学基金项目(NY220213,NY221105);江苏省社会科学基金项目(24EYB015);国家社会科学基金项目(25BJL058)

Supported by:Innovation Foundation of State Key Laboratory for Novel Software Technology of China(KFKT2022B19);Natural Science Foundation of University of Jiangsu Province, China(19KJB520047); Natural Science Foundation of Nanjing University of Posts and Telecommunications(NY220213,NY221105);Jiangsu Provincial Social Science Foundation Project(24EYB015);National Social Science Foundation of China(25BJL058)

Telecommunications, Nanjing 210003, China; 2. State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China; 3. School of Digital Economy, Nanjing Audit University, Nanjing 211815, China; 4. College of Meteorology and Oceanography, National University of Defense Technology, Changsha 411107, China

Abstract: Objective 3D model classification is a fundamental problem in the fields of computer graphics and computer vision, with wide-ranging applications in areas such as computer-aided design, mixed reality, autonomous driving, and robotic navigation. The challenges associated with 3D model classification primarily arise from three key aspects: the difficulty in representing 3D surface geometric features, the diversity of 3D transformations and deformations, and the incompleteness of geometric and topological structures. Existing multi-view-based 3D model classification methods typically render 3D models from multiple preset viewpoints and input all rendered views into a neural network for classification. However, due to the presence of redundant and ineffective views, not all views contribute equally to the classification task. Selecting views that substantially enhance classification performance can not only improve the overall accuracy of multi-view 3D model classification but also help identify representative views that effectively capture the essential characteristics of the 3D model. **Method** This paper proposes a Transformer attention-guided approach for optimal view selection and classification of 3D models. The 3D model is first rendered from 20 viewpoints arranged on a regular icosahedron. A convolutional neural network is then employed to extract feature information from these multiple views, producing a sequence of local multi-view feature tokens. Aiming to retain spatial location information, position encoding is applied to the token sequence. Next, a learnable global classification token is introduced and concatenated with the multi-view feature tokens, forming the input to a Transformer encoder that performs global view feature fusion and generates an initial global classification feature. Subsequently, the optimal view selection module calculates the contribution of each view to the initial global classification token using the attention score matrix from the feature fusion process. The highest-scoring views are selected as the optimal views. These optimal view feature tokens are then concatenated with the initial global classification token and input into the Transformer encoder for a second round of feature fusion, producing the final global classification token. This final token is passed through a classifier to generate the classification probabilities and simultaneously output the selected optimal views. Aiming to enhance generalization during training, the model incorporates random view dropping and contrastive learning strategies. **Result** This study experiments on the ModelNet40 dataset, which comprises 40 object categories. The dataset is suitable for research in 3D object recognition and is widely used for benchmarking algorithm performance. Evaluation metrics include overall accuracy (OA), average accuracy (AA), and speed. OA measures classification accuracy across the entire dataset, while AA calculates the mean accuracy across all categories, addressing issues related to class imbalance. The dataset, created by Stanford University, is widely used for performance evaluation of algorithms. First, the Transformer-based multi-view selection and 3D model classification method proposed in this paper are compared with other state-of-the-art deep learning-based 3D model classification methods to validate its effectiveness. Subsequently, ablation experiments are conducted to analyze the impact of different parameter settings on the performance of the proposed method, including multi-view representation, feature extraction backbone, Transformer hidden layer dimension, number of attention heads, contrastive learning strategy, and random view dropout module. On the ModelNet40 benchmark dataset, the proposed method achieves an overall recognition accuracy of 97.61% and an average recognition accuracy of 96.36%. In addition to reaching state-of-the-art classification performance, the optimal views selected based on the Transformer attention score matrix are shown to be highly representative. **Conclusion** The proposed method leverages the Transformer architecture to perform feature fusion across different views. By employing mechanisms such as self-attention, residual connections, and multi-layer stacking, the Transformer effectively learns complex features and captures global contextual relationships among different views. Furthermore, the attention score matrix generated by the Transformer serves as a basis for optimal view selection, enabling efficient classification while identifying the most representative views.

Key words: 3D model classification; Transformer; optimal view selection; contrastive learning; multi-view learning

0 引 言

三维模型分类是计算机图形学和计算机视觉领域中的一个基础问题，应用范围非常广泛，包括计算机辅助设计、混合现实、无人驾驶以及机器人导航等。三维模型分类的挑战主要来源于以下几个方面：三维表面几何特征难以表征、三维变换和形变的多样性、几何和拓扑结构的不完整等。

传统的三维模型分类方法使用手工设计的特征，如形状直方图（Ankerst 等，1999）、球谐描述子（Kazhdan 等，2003）和三维 SURF（speeded up robust features）（Knopp 等，2010）等，然后结合基于机器学习的分类器，如支持向量机（support vector machine，SVM）（Cortes 和 Vapnik，1995）、随机森林（Breiman，2001）等来实现对三维模型的分类。由于手工定义特征的表达能力和浅层算法泛化能力的不足，上述方法在适用的范围和鲁棒性上都有着很大的局限性。随着深度学习在自然语言与图像处理领域的快速发展，基于深度学习的三维模型分类方法已成为主流，依据三维数据输入方式的不同，可以将其分为3类：基于体素的方法、基于点云的方法和基于多视图的方法。

基于体素的方法（Maturana 和 Scherer，2015；Qi 等，2016；Zanuttigh 和 Minto，2017）使用体素化技术将三维对象转化为固定大小的体像素，通过深度体素卷积神经网络提取体素表示的深层特征以进行分类。尽管体素表示可以实现在三维模型上直接进行卷积与池化操作，但三维模型体素表示受限于较低的分辨率，导致会丢失三维模型的局部细节，同时三维卷积操作相较二维卷积操作需要更多的计算资源。

基于点云的方法（Qi 等，2017；Zhang 等，2020；Yang 等，2023）在三维模型采样后的点云上构建树、图等数据结构，保留了三维数据的原始特征。然而针对同一个物体，由于扫描设备和扫描方式的不同，获得的点云数据可能存在很大差异，点云数据往往存在稀疏性、非均匀性以及排列的无序性等问题，使得点云特征的提取面临着较多的挑战，基于点云特征的三维模型分类精度也有待于进一步提高。

基于多视图的三维模型分类方法通过从多个角度渲染三维模型生成二维图像，以利用二维图像特征进行三维模型的识别与分类。相比于三维体素表示和三维点云表示，基于多视图的分类方法的优势在于可以利用大量的图像数据库预训练深度神经网络，同时还能使用在图像识别任务中成功的先进网络架构，如 ResNet（residual neural network）（Huang 等，2013）、VGG（Visual Geometry Group）（Simonyan 和 Zisserman，2015）以及 DenseNet（densely connected convolutional network）（Huang 等，2017）等。由于上述原因，基于多视图的三维模型分类方法在精度上往往优于基于体素和点云表示的分类方法。

在基于多视图的三维模型分类研究中，Su 等人（2015）率先提出基于多视图卷积神经网络（multi-view convolutional neural network，MVCNN）的三维模型分类方法，首先通过卷积神经网络（convolutional neural network，CNN）获得每个局部视图特征，然后使用池化层聚合局部特征以得到全局分类特征，最后将其送入分类器中实现三维模型的分类。由于深度学习强大的表征能力，该方法的分类性能相比于基于手工特征的分类方法有了显著的提升，基于多视图的三维模型分类也获得了广泛的关注，研究人员在视图特征融合与视图特征聚合上提出多种改进方法。

Feng 等人（2018）提出组视图的概念，通过将多个视图分组，进行组内视图特征池化和组间特征融合，能够更有效地保留三维模型的全局信息。Han 等人（2019）将三维模型的多个视图按序列形式输入到 CNN，通过分层注意力聚合使得模型关注重要的视图特征，增强了模型对三维形状的理解。Liu 等人（2021）采用了分层多视图上下文建模的方法，自适应地计算特征权重并使用 KMeans 对视图特征进行聚合，然后通过双向长短期记忆网络（bidirectional long short-term memory，Bi-LSTM）进行组间视图特征融合。Wang 等人（2022b）提出一种基于多视点注意力卷积的三维点云分类方法，采用注意力卷积机制，在输入数据中寻找与当前输出有关的信息，以解决视图特征在池化过程中的信息丢失的问题。吴晗等人（2025）提出一种融合多视图一致性与互补信息的三维模型分类方法，通过引入可学习权重迭代网络和预分类模块，有效提升了多视图信息融合的准确性。上述方法通过单独设计权重（注意力）计算分支网络来反映视图和特征的区分性和互补性，提高视图特征的融合效果，进一步提升了分类的性能。

近年来，研究者成功地将自然语言处理模型

Transformer（Vaswani 等，2023）应用到视觉任务（周丽娟和毛嘉宁，2023），并取得了巨大的成功。在三维模型分类中，研究人员利用Transformer中的自注意力机制，实现不同层次的信息融合。Chen 等人（2021）为了解决不同视图块之间的信息无法交互的问题，设计了 MVT（multi-view Transformer）网络模型，首先对每个视图进行分块并将其输入局部Transformer中得到每个块的特征，然后将所有块的特征送入全局Transformer中以捕获跨视图的块与块之间的相关性。Li 等人（2023）在采用CNN提取投影视图的多尺度特征并对其进行聚合后，使用Transformer融合不同视图的特征，更好地捕获了不同视图之间的特征相关性，在ModelNet40基准数据集上，达到了95.4%的分类准确率。

尽管 Li 等人（2023）的方法基于 Transformer 实现了视图级的特征融合，然而其仍然存在一些局限。首先，该方法仍然使用池化的方式聚合经过 Transformer融合后的视图特征，不可避免地会引起局部特征信息的丢失；其次，Transformer在特征融合的每轮迭代中，都需要和全部的视图进行融合，不可避免地会受到冗余和无效视图的噪声干扰，从而影响分类性能。实际上，每个视图对于分类目标的作用并不相同，如果能够选择出对分类目标贡献度大的视图，不仅能够提高基于多视图的三维模型分类性能，而且能够提供表征三维模型的代表性视图。

针对以上问题，本文提出一种基于Transformer的三维模型最优视图选择与分类方法。首先，在Transformer进行视图特征融合的过程中，采用可学习的全局分类嵌入来表征三维模型的空间和几何信息，相比于池化的聚合方法，能够有效地避免特征的丢失。其次，在Transformer多层堆叠的过程中，以每一层的得分矩阵为依据，选择出最优的视图，并将最优的视图和可学习的全局分类嵌入进行融合，有效地避免了冗余和无效视图的噪声干扰，选择出的最优视图相比于基于信息熵视图选择的分类方法OVPT（optimal view prediction Transformer）（Wang等，2022a）更具有表征性。最后，采用随机丢弃视图策略与对比学习以增加模型的学习难度，进一步提高模型的泛化性能。

在ModelNet40基准数据集上选择4个视角的情况下，本文方法总体识别精度和平均识别精度分别为97.61%和96.36%，在达到先进分类水平的同时，基于Transformer注意力得分矩阵选择出的最优视图更具有表征性，为后续三维模型的展示和分析提供了支持。

1 本文方法

本文提出一种Transformer注意力引导的最优视图选择与分类方法。本节将首先对整体方法进行概述，介绍其模块构成与处理流程，然后分小节依次介绍各个组成部分的设计与实现。

1.1 方法概述

本文提出一种Transformer注意力引导的三维模型最优视图选择与分类方法，其神经网络结构如图1所示，主要包括多视图特征提取模块、全局多视图特征融合模块、最优视图选择模块、最优多视图特征融合模块以及分类模块。

本文设计的算法具体工作流程如下：

首先，将三维模型从正十二面体的20个预设视角进行渲染，利用CNN提取各视图的局部特征，构成一组视图特征Token，并加入位置编码以保留空间信息。随后，引入一个可学习的全局分类Token，并将其与上述视图Token拼接后输入Transformer编码器，实现全局特征融合并生成初步分类表示。

在此基础上，最优视图选择模块根据融合过程中的注意力得分，评估各视图对分类Token的贡献，从而筛选出关键视图。将这些高贡献视图的特征Token与初步分类Token一同再次输入Transformer，得到增强后的全局表示，最终用于三维模型的分类预测。

此外，为提升模型的泛化能力，本文在训练过程中引入了两项策略：1）随机丢弃部分视图以提升鲁棒性；2）采用对比学习以增强同类样本间的表示一致性。

1.2 视图特征提取模块

为了获取三维模型在多个视角下的语义特征，首先从N个预设视角渲染模型图像，并利用在ImageNet数据集上预训练的DenseNet网络（Huang 等，2017）对视图集$\boldsymbol{V}=\{\boldsymbol{v}_1,\cdots,\boldsymbol{v}_i,\cdots,\boldsymbol{v}_N\}$进行特征提取，得到特征序列$\boldsymbol{X}=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_i,\cdots,\boldsymbol{x}_N\}$，其中$\boldsymbol{x}_i\in\mathbf{R}^D$表示视图$\boldsymbol{v}_i$的特征向量，$D$为特征维度。

在此特征序列前加入一个可学习的全局分类Token，记做$\boldsymbol{x}_{\text{class}},\boldsymbol{x}_{\text{class}}\in\mathbf{R}^{1\times D}$，用于捕捉整个视图集

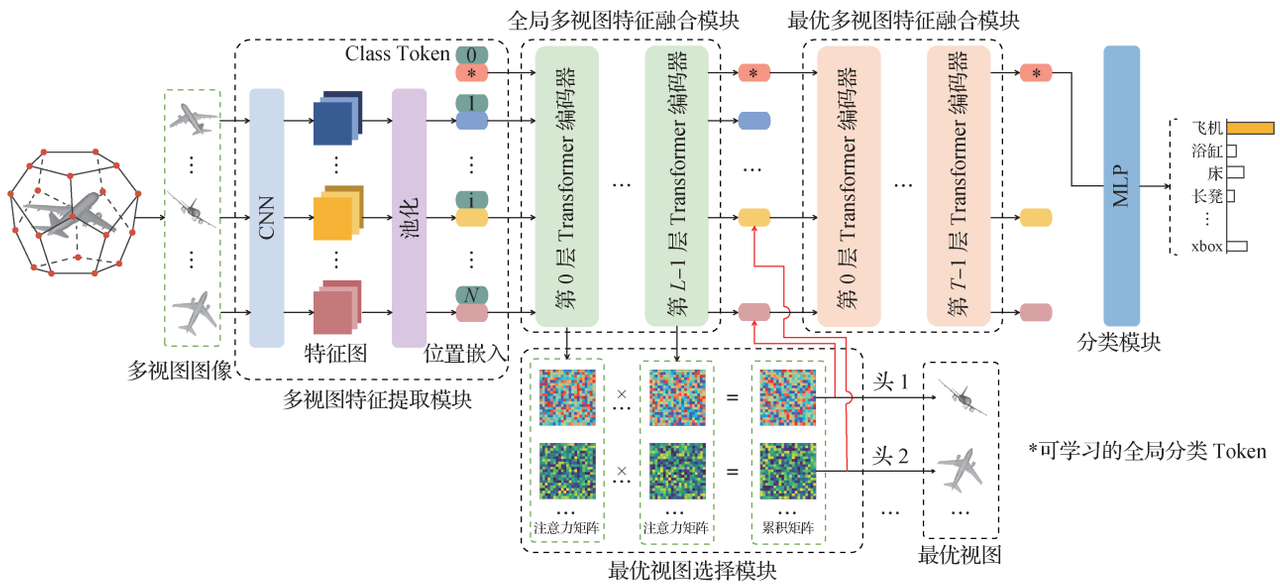

图 1　Transformer注意力引导的三维模型最优视图选择与分类方法网络架构

Fig. 1　Network architecture of Transformer attention-guided method for optimal view selection and classification of 3D models

的全局信息。同时，为保留各视图的空间顺序关系，引入可学习的位置编码嵌入 $\boldsymbol{E}_{\text{pos}}$。最终构建包含空间位置信息和分类表示的初始输入特征 $\boldsymbol{X}_0$，$\boldsymbol{X}_0 \in \mathbf{R}^{(N+1)\times D}$，具体为

$$\boldsymbol{X}_0 = [\boldsymbol{x}_{\text{class}}; \boldsymbol{x}_1, \cdots, \boldsymbol{x}_i, \cdots, \boldsymbol{x}_N] + \boldsymbol{E}_{\text{pos}} \tag{1}$$

1.3　全局视图特征融合模块

本文利用卷积神经网络对多个视角渲染图进行特征提取，得到包含不同角度空间结构信息的局部特征表示。由于单一视角所包含的信息具有局限性，难以反映三维模型的全貌，因此需要将多个视角的局部特征整合，以提升整体识别性能。

为了实现多视图信息的有效融合，本文采用Transformer作为特征融合模块，如图 2 所示。该模块以Transformer编码器结构为核心，采用多头自注意力机制（multi-head self-attention，MSA）增强特征间的关联，同时结合层归一化（layer normalization，LN）和多层感知机（multilayer perceptron，MLP）等操作实现深层语义建模。该结构允许模型在多个视图之间建立丰富的上下文联系，从而更全面地捕捉三维模型的几何特征。

在融合过程中，首先将每个视图的局部特征序列 $\boldsymbol{X}_0$ 输入Transformer编码器，进行归一化处理，具体为

$$\boldsymbol{X}_0' = LN(\boldsymbol{X}_0) \tag{2}$$

然后将归一化后的特征输入多头自注意力模块，计算注意力权重以捕捉不同视图之间的关系，具体为

$$\text{MSA}(\boldsymbol{X}_0') = f_{\text{attention}}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = f_{\text{softmax}}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{3}$$

式中，$\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 分别表示由输入特征 $\boldsymbol{X}_0'$ 映射得到的查询、键和值，d_k 为每个注意力头向量的维度。

接着，输出的注意力特征通过残差连接和 dropout 操作，与原始输入叠加，生成新的特征表示，具体为

$$\boldsymbol{X}_1' = f_{\text{dropout}}(\text{MSA}(\boldsymbol{X}_0')) + \boldsymbol{X}_0 \tag{4}$$

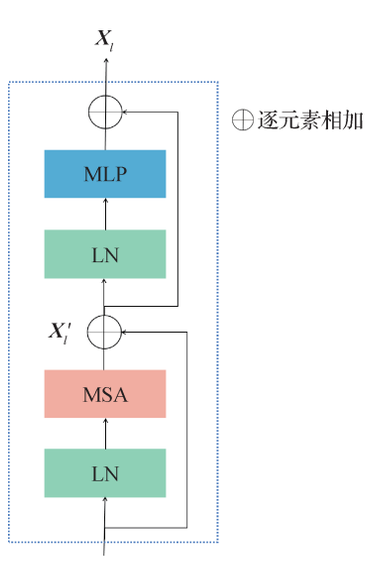

图 2　Transformer编码器层结构

Fig. 2　Structure of the Transformer encoder layer

为了增强非线性表达能力，接下来使用 MLP 模块对融合后的特征进一步映射，同时引入 dropout 和残差连接机制，生成最终的局部融合特征 $\boldsymbol{X}_1$，具体为

$$\boldsymbol{X}_1 = f_{\text{dropout}}(MLP(LN(\boldsymbol{X}_1'))) + \boldsymbol{X}_1' \tag{5}$$

整个 Transformer 编码器由多层堆叠组成，每一层都重复上述流程。对于第 l 层（$l = 2, \cdots, L$），其更新过程为

$$\boldsymbol{X}_l' = f_{\text{dropout}}(MSA(LN(\boldsymbol{X}_{l-1}))) + \boldsymbol{X}_{l-1} \tag{6}$$

$$\boldsymbol{X}_l = f_{\text{dropout}}(MLP(LN(\boldsymbol{X}_l'))) + \boldsymbol{X}_l' \tag{7}$$

通过多层堆叠的 Transformer 结构，模型能够有效整合不同视图之间的信息，并增强全局特征表示的建模能力，为后续最优视图选择和最终分类提供支持。

1.4 最优视图选择模块

本模块旨在基于注意力机制提取每个视图的显著性，进而选取具有判别能力的关键视图。具体而言，通过分析融合序列中的全局分类 Token 与各视图 Token 之间的注意力分布关系，从而度量每个视图对最终模型分类判断的贡献。

融合特征序列 $\boldsymbol{X}_L = [\boldsymbol{x}_L^0; \boldsymbol{x}_L^1, \boldsymbol{x}_L^2, \cdots, \boldsymbol{x}_L^N]$ 中，$\boldsymbol{x}_L^0$ 表示全局分类 Token，而 $\boldsymbol{x}_L^i (i \geqslant 1)$ 表示第 i 个视图对应的特征 Token。由于 Transformer 编码器由多层堆叠而成，且每层包含 M 个注意力头，因此在注意力矩阵中，可以获取每一层每个头的注意力向量，用于刻画 Token 之间的相对关注程度。

对于第 l 层，得到的注意力得分矩阵 $\boldsymbol{a}_l$ 具体为

$$\boldsymbol{a}_l = [\boldsymbol{a}_l^1, \boldsymbol{a}_l^2, \cdots, \boldsymbol{a}_l^M],\ l \in \{1, \cdots, L\} \tag{8}$$

式中，$\boldsymbol{a}_l^i$ 表示第 l 层第 i 个头的注意力得分矩阵，其具体结构为

$$\boldsymbol{a}_l^i = [\boldsymbol{a}_l^{i_0}, \boldsymbol{a}_l^{i_1}, \boldsymbol{a}_l^{i_2}, \cdots, \boldsymbol{a}_l^{i_N}], i \in \{1, \cdots, M\} \tag{9}$$

式中，$\boldsymbol{a}_l^{i_0}$ 为每个 Token 对全局分类 Token 的注意力分数向量，$\boldsymbol{a}_l^{i_1}$ 为每个 Token 对第 1 个视图 Token 的注意力分数向量，以此类推。图 3 给出了属于同一个头的层 2 和层 1 的得分矩阵示例。矩阵的第 i 行表示每个 Token 对第 i 个 Token 的注意力权重，同理，矩阵的第 j 列表示第 j 个 Token 对每个 Token 的注意力权重。

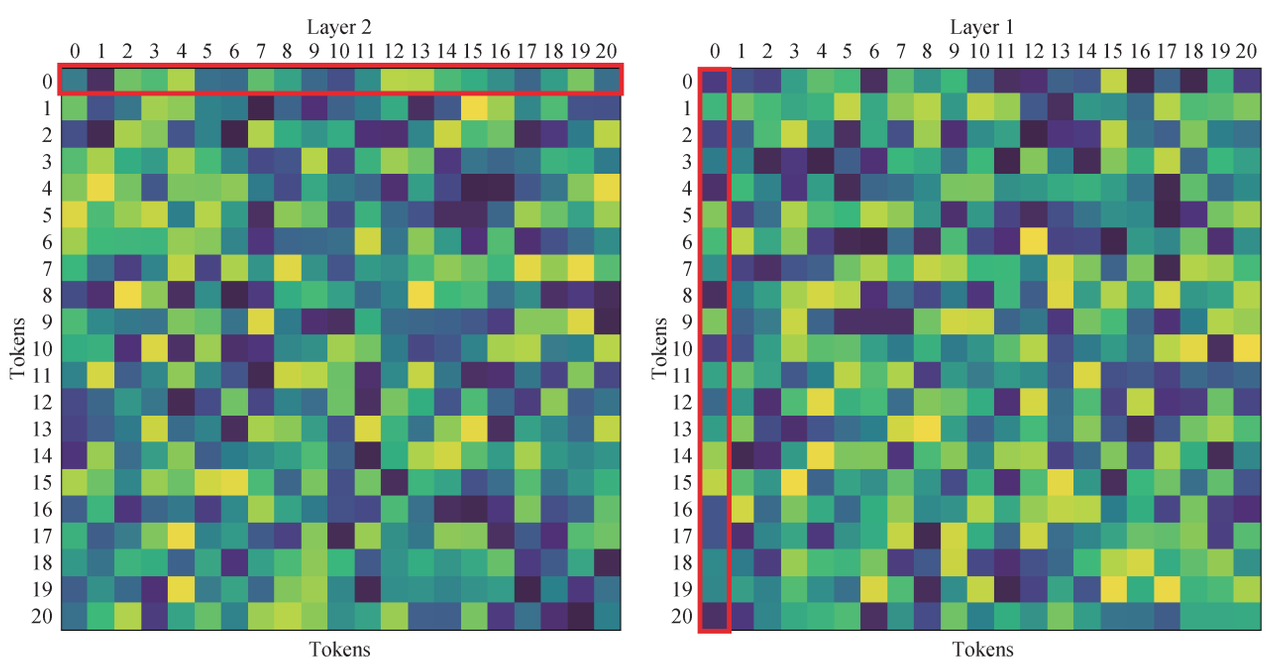

图 3 Transformer 得分矩阵示例（视图数为 20）

Fig. 3 Example of Transformer score matrix (views: 20)

在 Transformer 多层堆叠过程中，注意到对输入的 Token（即 $\boldsymbol{X}_0$）的注意力是个扩散的过程，对每一个头 i，递归地将所有层的注意力得分矩阵相乘，将得到头 i 中 $\boldsymbol{X}_L$ 对于输入 $\boldsymbol{X}_0$ 中每个 Token 的权重，计算为

$$\boldsymbol{a}_{\text{accu}}^i = \prod_{l=1}^{L} \boldsymbol{a}_l^i \tag{10}$$

以图 3 为例，右侧 Layer 1 中的红框对应的列为 Token 0 对所有 Token 的得分，左侧 Layer 2 中的红框对应的行是第 2 次迭代时，其他所有 Token 对 Token

0的得分。Layer 2和Layer 1相乘时，左侧红框（行）中的元素和右侧红框（列）中的元素逐元素相乘并相加，结果为Token 0经过2层变换后从输入$\boldsymbol{X}_0$中Token 0获得的累积注意力。

经过如式（10）所示的得分矩阵逐层相乘后，$\boldsymbol{a}_{\text{accu}}^{i}$矩阵捕获了头$i$中输入$\boldsymbol{X}_0$和输出$\boldsymbol{X}_L$之间的注意力累积信息。显然矩阵中的分值越大，则该列号对应的输入Token对行号对应的输出Token的贡献就越大。

输出的第0行对应可学习全局分类嵌入Token 0，Token 1～N为视图特征序列。如图1所示，Token 0将用于最后的分类，而$\boldsymbol{a}_{\text{accu}}^{i}$中第0行从第1列起，得分最高的值对应的列号为对头$i$全局分类特征贡献最大的视图索引，因此该视图为头$i$的最优视图。具体为

$$\boldsymbol{v}_{\text{best}}^{i}=\max\left(\boldsymbol{a}_{\text{accu}}^{i}[0,1:]\right) \tag{11}$$

对于M个头，根据式（11）选取每个头在$\boldsymbol{a}_{\text{accu}}$中最大值的列号，构成最优视图索引列表，即$[O_1;\ O_2;\ \cdots;\ O_M]$。

1.5 最优视图特征融合模块

在完成最优视图选择后，本模块依据最优视图的序列索引，从上一层特征序列$\boldsymbol{X}_L$中提取相应的视图特征Token，并与全局分类Token拼接，构成新的输入序列，表示为

$$\boldsymbol{X}_L'=[\boldsymbol{x}_L^0;\boldsymbol{x}_L^{O_1},\boldsymbol{x}_L^{O_2},\cdots,\boldsymbol{x}_L^{O_M}] \tag{12}$$

式中，$\boldsymbol{x}_L^0$为全局分类Token，$\boldsymbol{x}_L^i$为第i个被选中视图的特征表示，共选择M个最优视图。

该序列随后被送入另一个Transformer编码器中进行进一步的特征融合。该编码器共包含T层，其计算流程与式（6）和式（7）一致。经过多层编码后，最终获得融合后的特征序列$\boldsymbol{X}_T$，即

$$\boldsymbol{X}_T=[\boldsymbol{x}_T^0;\boldsymbol{x}_T^{O_1},\boldsymbol{x}_T^{O_2},\cdots,\boldsymbol{x}_T^{O_M}] \tag{13}$$

在该模块中，仅使用筛选出的最优视图参与融合，通过多层Transformer的表达能力，有效抑制了冗余或低质量视图对分类任务的干扰。这种筛选与融合相结合的策略进一步增强了特征的判别性，有助于提升整体模型的分类性能。

1.6 特征分类模块

最终融合得到的序列$\boldsymbol{X}_T$中包含了整合视图信息后的语义表示，尤其是其中的全局分类Token $\boldsymbol{x}_T^0$，其携带了从多个最优视图中聚合而来的判别特征。

在分类阶段，模型利用该Token作为输入，通过全连接层映射为预测向量$\boldsymbol{Y}$，具体计算过程为

$$\boldsymbol{Y}=f_{\text{linear}}(\boldsymbol{x}_T^0) \tag{14}$$

接着通过softmax函数对$\boldsymbol{Y}$进行归一化处理，输出各类别的预测概率，具体为

$$\boldsymbol{P}(\boldsymbol{y}\,|\,\boldsymbol{x})=f_{\text{softmax}}(\boldsymbol{Y})=\frac{\boldsymbol{e}^{\boldsymbol{Y}}}{\sum_{i=1}^{C}\boldsymbol{e}^{\boldsymbol{Y}_i}} \tag{15}$$

式中，C表示类别总数，$\boldsymbol{Y}_i$为第i类的得分。最终，模型输出的概率最大值对应的索引即为该三维模型所属的预测类别。

1.7 随机丢弃视图学习与对比学习

本文采用随机丢弃视图策略与对比学习以增加模型的学习难度，进一步提高模型的泛化性能。

随机丢弃策略是一种正则化技术，通常使用dropout在训练过程中丢弃一部分神经元，以减小模型对特定神经元的依赖。受到该思想的启发，本文在训练过程中除了随机丢弃部分神经元外，还采用随机丢弃视图的策略，以进一步防止模型过拟合。具体而言，在通过CNN获得局部视图特征序列$\boldsymbol{X}=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_i,\cdots,\boldsymbol{x}_N\}$后，采用动态掩码策略，即对$N$个视图随机生成Transformer的掩码，使得其中K个视图对应的掩码位为True，从而增大模型的学习难度。

对比学习（contrastive learning）旨在通过对比样本之间的相似性和差异性来学习数据的特征表示。本文采用对比学习进一步提高模型的特征能力。通过设计对比损失函数，使得不同类别全局分类Token相似性最小化，同时使得具有相同类别的全局分类Token相似性最大化。为了防止损失被负样本（相似度小的不同类别的样本）所主导，引入常数α，只有相似度大于α的负样本才会贡献损失，具体的对比损失函数定义为

$$\mathcal{L}_{\text{con}}=\frac{1}{B^2}\left[\sum_{j:\boldsymbol{y}_i=\boldsymbol{y}_j}^{B}1-Sim(\boldsymbol{x}_i,\boldsymbol{x}_j)+\sum_{j:\boldsymbol{y}_i\neq\boldsymbol{y}_j}^{B}\max\left((Sim(\boldsymbol{x}_i,\boldsymbol{x}_j)-\alpha),0\right)\right] \tag{16}$$

式中，B表示batch size，$\boldsymbol{x}_i$和$\boldsymbol{x}_j$表示经过L2归一化的特征向量，$Sim(\boldsymbol{x}_i,\boldsymbol{x}_j)$是$\boldsymbol{x}_i$和$\boldsymbol{x}_j$的点积。

本文使用交叉熵损失$\mathcal{L}_{\text{cross}}$和对比学习损失$\mathcal{L}_{\text{con}}$共同训练模型。模型总损失函数定义为

$$\mathcal{L}=\mathcal{L}_{\text{cross}}(\boldsymbol{y},\boldsymbol{y}')+\mathcal{L}_{\text{con}}(\boldsymbol{x}) \tag{17}$$

式中，$\mathcal{L}_{\text{cross}}(\boldsymbol{y},\boldsymbol{y}')$是预测标签$\boldsymbol{y}'$与真实标签$\boldsymbol{y}$之间的交叉熵损失。

2 实验结果与分析

为验证所提出方法的有效性，本节从多个角度进行了实验分析。首先介绍所使用的数据集和评价准则，然后说明实验设置与环境参数。随后，将本文方法与多种先进的三维模型分类方法进行对比，并通过一系列消融实验，进一步分析各模块对整体性能的影响。

2.1 数据集介绍

本文在ModelNet40数据集上进行实验。ModelNet40是一个广泛用于3D物体识别领域的数据集，包含40个物体类别。这些物体涵盖各类日常物品，如椅子、桌子、键盘、汽车等。该数据集的训练集有9 843个物体，测试集包含2 468个物体。该数据集由美国斯坦福大学创建，广泛应用于三维模型领域的研究，提供了一个用于评估和比较不同算法性能的标准基准。

2.2 评价准则

总体识别精度（overall accuracy，OA）、平均识别精度（average accuracy，AA）和分类速度是三维模型分类研究中常用的性能评价准则。其中，OA是指分类模型在整个数据集上的准确率，表示模型正确分类的样本数量在总样本数量中的占比。OA计算为

$$OA = \frac{P}{Z} \times 100\% \tag{18}$$

式中，P表示准确分类的模型数量，Z表示总体三维模型数量。

AA常用于多类别分类任务中，表示模型对每个类别分类准确率的平均值，可以解决不同类别之间样本数量不平衡问题，AA计算为

$$AA = \frac{AA_1 + AA_2 + \cdots + AA_C}{C} \tag{19}$$

式中，AA_i表示模型中每个类别的准确率，C为类别的数量。

2.3 实验设置及环境参数

实验在Ubuntu 20.04.4服务器上进行，CPU型号为Intel（R）Core（TM）i9-10900X，GPU型号为NVIDIA GeForce RTX 3090（24 GB），使用Python 3.9.18，PyTorch 2.1.1+cuda12.1版本进行开发。批次大小（batchsize）设为8，该设置在GPU显存允许的条件下可最大化批处理规模，从而提升训练效率。在训练过程中使用Adam优化器，初始学习率设置为0.000 1，并采用指数型衰减策略。该学习率设置参考了相关工作OVPT中的推荐参数。如图4所示，该设置在本任务中能够较好地平衡收敛速度与训练稳定性。迭代周期（epoch）设为40，模型在该轮次内已完成收敛。

2.4 与先进方法的对比

首先将本文提出的基于Transformer的三维模型多视图选择与分类方法与其他先进的基于深度学习的三维模型分类方法进行对比，以验证本文方法的有效性。这些方法包括MVCNN（Su等，2015）、MHBN（multi-view harmonized bilinear network）（Yu等，2018）、MVTN（multi-view Transformation network）（Hamdi等，2021）、RotationNet（Kanezaki等，2018）、MVT（Chen等，2021）和OVPT。

不同三维模型分类方法的总体识别精度（OA）和平均识别精度（AA）如表1所示。由于OVPT与本文方法在性能上最为接近，本文通过实验进一步给出其在4个视角下的性能，记做OVPT*，以便对比分析。可以看出，本文方法的总体识别精度（OA）为97.61%，相较于MVCNN、MHBN、MVTN、RotationNet、MVT、OVPT*和OVPT分别提高5.51%、2.91%、4.11%、0.24%、0.11%、0.28%和0.13%。本文方法的平均精度（AA）为96.74%，相较于MVCNN、MHBN、MVTN和RotationNet分别提高6.36%、3.35%、4.16%和0.52%，相比于OVPT（6个

表1 不同三维模型分类方法精度对比

Table 1 Accuracy comparison of different classification methods for 3D model

模型	视图数	OA/%	AA/%
MVCNN	12	92.10	90.00
MHBN	6	94.70	93.01
MVTN	12	93.50	92.20
RotationNet	20	97.37	95.84
MVT	20	97.50	–
OVPT*	4	97.33	96.07
OVPT	6	97.48	**96.74**
本文	4	**97.61**	96.36

注：加粗字体表示各列最优结果。所有数据由论文作者提供，“–”表示作者未提供该项数据。

视图)低0.38%,但均采用4个视图时,较OVPT*方法提高0.29%。总体来看,本文方法在总体识别精度和平均识别精度上达到了先进的水平。

图4展示了本文方法随着训练epoch的增加,总体识别精度和平均识别精度逐渐收敛的曲线图,图4(a)为epoch-OA曲线图,图4(b)为epoch-AA曲线图。

为了进一步评估各方法的执行效率,本文对比了不同方法在实验Ubuntu服务器上的分类时长,其由渲染视图时长和基于渲染的视图进行推理的时长构成。实验统一使用PyTorch3D renderer库进行高效的批处理并行渲染。每个指标采用5次运行的平均结果,结果如表2所示。

在表2中,前4种方法均基于CNN架构实现分类,而后4种方法则引入Transformer结构,以增强分类特征表示能力。除RotationNet外,未使用Transformer的方法在推理阶段的耗时明显低于使用Transformer的方法。RotationNet由于需要通过对多个视图顺序进行排列组合来估计姿态,因此推理时间在所有方法中最长。在基于Transformer的方法中,本文方法需将全部20个视图输入到Transformer中以实现最优视图选择,从而导致推理时间略高于OVPT。然而从表2可见,即使采用并行渲染,每种方法的渲染时长仍远高于推理时长。因此,本文方法运行效率处于可接受的范围内。

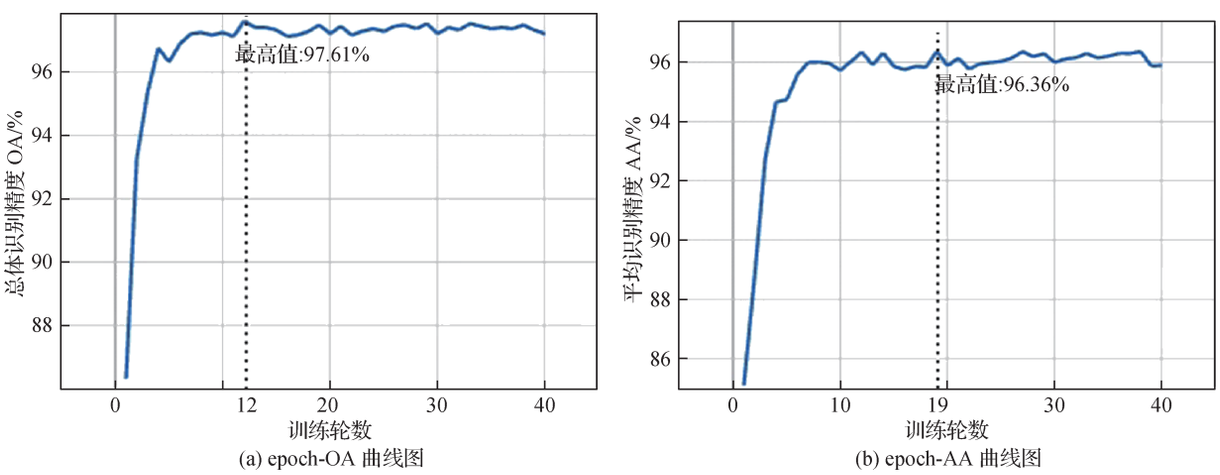

图4　OA和AA训练收敛过程

Fig. 4　Convergence process of OA and AA during training ((a) epoch-OA curve; (b) epoch-AA curve)

表2　不同三维模型分类方法速度对比

Table 2　Speed comparison of different classification methods for 3D model

模型	渲染/选择视图	渲染时长/ms	推理时长/ms	总时长/ms
MVCNN	12/12	110.2	**4.3**	114.5
MHBN	6/6	**64.8**	4.6	**69.4**
MVTN	12/12	110.2	10.7	120.9
RotationNet	20/20	172.6	110.5	283.1
MVT	20/20	172.6	56.3	228.9
OVPT*	20/4	172.6	16.1	188.7
OVPT	20/6	172.6	18.3	190.9
本文	20/4	172.6	43.2	215.8

注:加粗字体表示各列最优结果。

OVPT同样为基于视图选择的三维模型分类方法,同时与本文方法的分类性能最为接近,接下来,本文将对比分析这两种方法选择的最优视图。随机选取5种三维模型,本文方法和OVPT*选择的4个最优视图如图5所示,其中蓝色框标出的视图为最优视图,每种类别的第1行为本文方法,第2行是OVPT方法。

OVPT根据视点熵的信息量选择最优视图,可以看出其选择出的二维视图都是固定序列[4,9,13,14]。OVPT认为每个视图都是由固定摄像机的固定角度渲染而成,因此20个视图的信息量都是固定排序的,无法随着三维模型的不同选择不同的最优视角。本文方法依据对全局分类Token的贡献得分选择出最优视图,选择出的视图为分类提供了最多的信息。选择出的最优视图并不局限于固定的视点,而是为了满足分类的需要,具有更好的区分性。因此,本文方法选择出的最优视图更具有表征性。

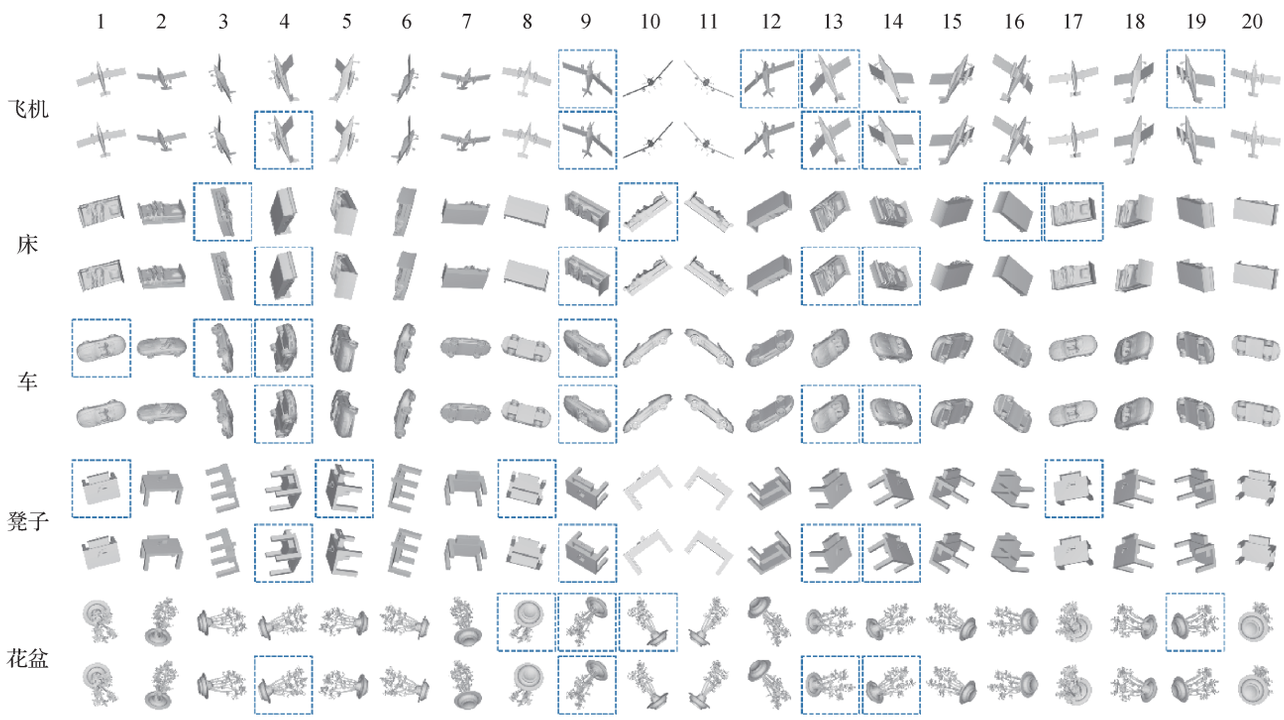

图5　视图选择结果对比(各类别上层为本文方法,下层为OVPT方法)

Fig. 5　Comparison of view selection results (top row: our method, bottom row: OVPT method)

2.5　消融实验

为分析不同的参数设置与选择对所提出的三维模型最优视图选择与分类方法性能的影响,本文进行了多项消融实验。

2.5.1　多视图表示方式对分类性能的影响

相机的数量、位置和角度决定了多视图的数量和质量。因此,选取数量适中、信息量丰富的多视图对于三维模型分类至关重要。本文采用两种多视图表示方式,第1种是以Z轴作为旋转轴虚拟摄像机环绕一周摆放,第2种采用正十二面体视点设置的方式。两种多视图表示方式的分类性能如表3所示,注意力头数为2,即选择2个最优视图。从表3的实验结果可以看出,在相同模型设置下,采用正十二面体视点设置方法的模型总体识别精度(OA)达到97.53%,平均识别精度(AA)达到了96.59%。与第1种多视图表示方式相比,正十二面体视点设置方法下的总体识别精度提高4.30%,平均识别精度提高4.82%。采用正十二面体视点设置方法在模型分类任务中取得了更好的表现。这表明正十二面体视点设置方法能够更均匀地覆盖模型的表面,提供了更丰富和多样化的视角信息,从而使得模型能够更准确地进行分类。

表3　不同多视图表示方式下模型识别精度对比

Table 3　Comparison of the recognition accuracy under different multi-view representations

/%

多视图表示方式	注意力头数	对比学习	OA	AA
环绕一周	2	√	93.23	91.77
正十二面体	2	√	**97.53**	**96.59**

注:加粗字体表示各列最优结果。

2.5.2　不同backbone对分类性能的影响

采用正十二面视点设置,渲染生成三维模型的多视图表示后,需要对每个视图进行局部特征提取。特征提取的好坏,很大程度上将影响后续的视图特征融合以及最终的三维模型分类。因此,本文选取ResNet18、ResNet34、ResNet50以及DenseNet这几种主流的CNN网络框架提取视图的特征,以探索不同backbone对多视图三维模型分类性能的影响。这其中,ResNet18和ResNet34均属于轻量级残差网络,层数较浅,计算代价较低,适合快速提取中低层次特征;ResNet50引入了更深的网络结构和Bottleneck模块,具备更强的抽象能力;而DenseNet通过密集连接机制增强了特征复用与梯度传播能力,在图像识别等任务中表现优异。本文分别使用2个和4个

注意力头进行实验，结果如表4所示。

根据表4可知，在相同注意力头数下，使用DenseNet网络框架的识别精度最高。注意力头数为2时，DenseNet的总体识别精度（OA）达到97.53%，相较于ResNet18、ResNet34和ResNet50分别提高0.65%、0.65%和0.13%。平均识别精度（AA）达到96.59%，相较于ResNet18、ResNet34和ResNet50分别提高0.77%、1.37%和0.44%。注意力头数为4时，DenseNet的总体识别精度达到97.61%，相较于ResNet18、ResNet34和ResNet50分别提高0.53%、0.37%和0.97%。平均识别精度达到96.36%，相较于ResNet18、ResNet34和ResNet50分别提高1.14%、0.37%和1.19%。因此，本文采用DenseNet作为特征提取模块的网络框架。

表4 不同backbone对分类性能的影响

Table 4 Impact of different backbones on the classification performance

基线	注意力头数	OA/%	AA/%
ResNet18	2	96.88	95.82
ResNet34	2	96.88	95.22
ResNet50	2	97.40	96.15
DenseNet	2	97.53	**96.59**
ResNet18	4	97.08	95.22
ResNet34	4	97.24	95.99
ResNet50	4	96.64	95.17
DenseNet	4	**97.61**	96.36

注：加粗字体表示各列最优结果。

2.5.3 Transformer隐层维度对分类性能的影响

本文尝试了3种不同的Transformer模型设置，旨在评估Transformer隐层维度（hidden size）对模型识别精度的影响，实验结果如表5所示。可以看出，当隐层维度为192维时，模型总体识别精度为97.61%，相较于隐层维度为384维和768维的总体识别精度高出0.32%和0.41%。模型的平均识别精度为96.36%，相较于隐层维度为384维和768维的总体识别精度高出0.17%和0.53%。通过分析可知，采用较轻量级的隐层维度（192维）相比于更大的维度设置（384维和768维），模型具有更好的分类性能。

表5 Transformer隐层维度对分类性能的影响

Table 5 Impact of transformer hidden layer dimensions on the classification performance

模型	隐层维度	OA/%	AA/%
tiny	192	**97.61**	**96.36**
small	384	97.29	96.19
base	768	97.20	95.83

注：加粗字体表示各列最优结果。

2.5.4 注意力头数对分类性能的影响

Transformer注意力头数决定了选择的最优视图数量，对模型的分类精度有着重要的影响，本文设置了4种不同数量的注意力头数，实验结果如表6所示。可以看出，当注意力头数为2时，模型总体识别精度和平均识别精度分别为97.53%和96.59%，当注意力头数为4时，模型总体识别精度和平均识别精度分别为97.61%和96.36%，两者性能相当，但当注意力头数大于4后，总体识别精度不增反降，这说明并不是注意力头数越多越好，较少的注意力头数往往能取得更好的性能。

表6 注意力头数对分类性能的影响

Table 6 Impact of the number of attention heads on the classification performance

注意力头数	OA/%	AA/%
2	97.53	96.59
4	**97.61**	**96.36**
6	97.53	96.23
8	97.41	96.12

注：加粗字体表示各列最优结果。

2.5.5 对比学习、随机丢弃视图和最优视图选择对分类性能的影响

在训练过程中，本文采用随机丢弃视图和对比学习以提高模型的泛化性能。本节将考察这两者与最优选择模块对分类性能的影响，结果如表7所示。

从表7可以看出：1）没有添加对比学习、随机丢弃视图和最优视图选择模块的基线模型中，总体识别精度为96.47%，平均识别精度为95.14%。仅添加对比学习模块时，总体识别精度为96.96%，平均识别精度为95.53%，表明对比学习模块的引入对模

型的识别精度能力有了明显提升。2)仅添加随机丢弃视图模块,总体识别精度提升0.45%,平均识别精度提升0.93%,显示了随机丢弃视图对模型性能的正面影响。3)仅添加最优视图选择模块,总体识别精度和平均识别精度分别下降0.20%和0.04%,初步判定单独使用最优视图选择模块不利于模型识别精度的提升。4)添加对比学习和随机丢弃视图模块,总体识别精度为97.33%,平均识别精度为96.31%。在此基础上,进一步引入最优视图选择模块,总体识别精度和平均识别精度分别为97.61%和96.36%,相比在引入最优视图选择模块之前分别提升0.28%和0.05%。由此可见,在模型添加对比学习和随机丢弃视图模块时,最优视图选择模块在选择出最优视图的同时,同样提升了模型的识别精度。

表7　对比学习、随机丢弃视图和最优视图选择模块对模型识别精度的影响

Table 7　Impact of contrastive learning, random view dropout, and optimal view selection module on the recognition accuracy

对比学习	随机丢弃视图	最优视图选择	OA/%	AA/%
×	×	×	96.47	95.14
√	×	×	96.96	95.53
×	√	×	96.92	96.07
×	×	√	96.27	95.10
√	√	×	97.33	96.31
√	×	√	97.24	95.73
×	√	√	96.92	95.60
√	√	√	**97.61**	**96.36**

注:加粗字体表示各列最优结果。“√”和“×”分别表示使用和未使用对应模块。

3　结　论

本文提出一种Transformer注意力引导的三维模型最优视图选择与分类方法,该方法通过在Transformer的特征融合过程中引入可学习的全局分类嵌入,能够有效保留三维模型的空间和几何信息,避免了传统池化方法中可能出现的特征丢失问题。此外,本文采用基于注意力得分矩阵的视图选择策略,有效减少了冗余和无效视图的干扰。该方法不仅提高了分类精度,选择出的最优视图也为后续三维模型的展示与分析提供了有效支持。该方法还结合了随机丢弃视图策略和对比学习,进一步增强了模型的泛化能力。在ModelNet40基准数据集上,本文提出的方法取得了97.61%的总体识别精度和96.36%的平均识别精度,在达到当前最先进的分类水平的同时,基于Transformer注意力得分矩阵选择出的最优视图更具有表征性。

本文提出的方法存在如下局限:首先,该方法基于已有的渲染视图选择最优的视图,所有的视图都需要进行特征提取和信息融合,计算量较大,不利于需要实时处理的场景;其次,该方法基于Transformer得分矩阵选择出最优视图,没有对最优视图之间的关系进行进一步的解释与描述。上述两个方面的不足为本文的后续研究提供了方向。

参考文献(References)

Ankerst M, Kastenmüller G, Kriegel H P and Seidl T. 1999. 3D shape histograms for similarity search and classification in spatial databases//Advances in Spatial Databases: 6th International Symposium. Hong Kong, China: Springer: 207-226 [DOI: 10.1007/3-540-48482-5_14]

Breiman L. 2001. Random forests. Machine Learning, 45(1): 5-32 [DOI: 10.1023/A:1010933404324]

Chen S, Yu T and Li P. 2021. MVT: multi-view vision transformer for 3D object recognition [EB/OL]. [2025-02-12]. https://arxiv.org/pdf/2110.13083.pdf

Cortes C and Vapnik V. 1995. Support-vector networks. Machine Learning, 20(3): 273-297 [DOI: 10.1007/BF00994018]

Feng Y F, Zhang Z Z, Zhao X B, Ji R R and Gao Y. 2018. GVCNN: group-view convolutional neural networks for 3D shape recognition//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 264-272 [DOI: 10.1109/CVPR.2018.00035]

Hamdi A, Giancola S and Ghanem B. 2021. MVTN: multi-view transformation network for 3D shape recognition//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 1-11 [DOI: 10.1109/ICCV48922.2021.00007]

Han Z Z, Lu H L, Liu Z B, Vong C M, Liu Y S, Zwicker M, Han J W

and Chen C L P. 2019. 3D2SeqViews: aggregating sequential views for 3D global feature learning by CNN with hierarchical attention aggregation. IEEE Transactions on Image Processing, 28(8): 3986-3999 [DOI: 10.1109/TIP.2019.2904460]

Huang G, Liu Z, Van Der Maaten L and Weinberger K Q. 2017. Densely connected convolutional networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 2261-2269 [DOI: 10.1109/CVPR.2017.243]

Huang Q X, Su H and Guibas L. 2013. Fine-grained semi-supervised labeling of large shape collections. ACM Transactions on Graphics (TOG), 32(6): #190 [DOI: 10.1145/2508363.2508364]

Kanezaki A, Matsushita Y and Nishida Y. 2018. RotationNet: joint object categorization and pose estimation using multiviews from unsupervised viewpoints//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 5010-5019 [DOI: 10.1109/CVPR.2018.00526]

Kazhdan M, Funkhouser T and Rusinkiewicz S. 2003. Rotation invariant spherical harmonic representation of 3D shape descriptors//Proceedings of 2003 Eurographics/ACM SIGGRAPH Symposium on Geometry Processing. Aachen, Germany: Eurographics Association: 156-164

Knopp J, Prasad M, Willems G, Timofte R and van Gool L. 2010. Hough transform and 3D SURF for robust three dimensional classification//Proceedings of the 11th European Conference on Computer Vision. Heraklion, Greece: Springer: 589-602 [DOI: 10.1007/978-3-642-15567-3_43]

Li J, Liu Z, Li L, Lin J Q, Yao J and Tu J M. 2023. Multi-view convolutional vision Transformer for 3D object recognition. Journal of Visual Communication and Image Representation, 95: #103906 [DOI: 10.1016/J.JVCIR.2023.103906]

Liu A A, Zhou H Y, Nie W Z, Liu Z G, Liu W, Xie H T, Mao Z D, Li X Y and Song D. 2021. Hierarchical multi-view context modelling for 3D object classification and retrieval. Information Sciences, 547: 984-995 [DOI: 10.1016/J.INS.2020.09.057]

Maturana D and Scherer S. 2015. VoxNet: a 3D convolutional neural network for real-time object recognition//Proceedings of 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems. Hamburg, Germany: IEEE: 922-928 [DOI: 10.1109/IROS.2015.7353481]

Qi C R, Su H, Mo K C and Guibas L J. 2017. PointNet: deep learning on point sets for 3D classification and segmentation//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 77-85 [DOI: 10.1109/CVPR.2017.16]

Qi C R, Su H, Nießner M, Dai A, Yan M Y and Guibas L J. 2016. Volumetric and multi-view CNNs for object classification on 3D data//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 5648-5656 [DOI: 10.1109/CVPR.2016.609]

Simonyan K and Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2025-02-12]. https://arxiv.org/pdf/1409.1556.pdf

Su H, Maji S, Kalogerakis E and Learned-Miller E. 2015. Multi-view convolutional neural networks for 3D shape recognition//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 945-953 [DOI: 10.1109/ICCV.2015.114]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L and Polosukhin I. 2023. Attention is all you need [EB/OL]. [2025-02-12]. https://arxiv.org/pdf/1706.03762.pdf

Wang W J, Chen G, Zhou H R and Wang X L. 2022b. OVPT: optimal viewset pooling transformer for 3D object recognition//Proceedings of the 16th Asian Conference on Computer Vision. Macao, China: Springer: 486-503 [DOI: 10.1007/978-3-031-26319-4_29]

Wang W J, Wang T and Cai Y. 2022a. Multi-view attention-convolution pooling network for 3D point cloud classification. Applied Intelligence, 52(13): 14787-14798 [DOI: 10.1007/s10489-021-02840-2]

Wu H, Hu L C, Yang Y, Jie B and Luo Y L. 2025. Multiview consistent and complementary information fusion method for 3D model classification. Journal of Image and Graphics, 30(3): 811-823 (吴晗, 胡良臣, 杨影, 接标, 罗永龙. 2025. 融合多视图一致和互补信息的深度3D模型分类. 中国图象图形学报, 30(3): 811-823) [DOI: 10.11834/jig.240060]

Yang M M, Chen J J and Velipasalar S. 2023. Cross-modality feature fusion network for few-shot 3D point cloud classification//Proceedings of 2023 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 653-662 [DOI: 10.1109/WACV56688.2023.00072]

Yu T, Meng J J and Yuan J S. 2018. Multi-view harmonized bilinear network for 3D object recognition//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 186-194 [DOI: 10.1109/CVPR.2018.00027]

Zanuttigh P and Minto L. 2017. Deep learning for 3D shape classification from multiple depth maps//Proceedings of 2017 IEEE International Conference on Image Processing. Beijing, China: IEEE: 3615-3619 [DOI: 10.1109/ICIP.2017.8296956]

Zhang M, Wang Y F, Kadam P, Liu S and Jay Kuo C C. 2020. Pointhop++: a lightweight learning model on point sets for 3D classification//Proceedings of 2020 IEEE International Conference on Image Processing (ICIP). Abu Dhabi, United Arab Emirates: IEEE: 3319-3323 [DOI: 10.1109/ICIP40778.2020.9190740]

Zhou L J and Mao J N. 2023. Vision Transformer-based recognition tasks: a critical review. Journal of Image and Graphics, 28(10): 2969-3003 (周丽娟, 毛嘉宁. 2023. 视觉Transformer识别任务

研究综述. 中国图象图形学报, 28(10): 2969-3003)[DOI: 10.11834/jig.220895]

作者简介

陈松乐,男,副教授,硕士生导师,主要研究方向为计算机视觉、计算机图形学和深度学习。
E-mail: chensongle@njupt.edu.cn
李骞,通信作者,男,副教授,硕士生导师,主要研究方向为计算机视觉、计算机图形学和深度学习。
E-mail: public_liqian@163.com
黄茹玥,女,硕士研究生,主要研究方向为计算机视觉与深度学习。E-mail: hry1024nike@gmail.com
黄思轩,女,硕士研究生,主要研究方向为计算机图形学与深度学习。E-mail: hsxuan0608@163.com
陈怡,女,教授,博士生导师,主要研究方向为电子商务和大数据。E-mail: dongtaichen@nau.edu.cn