基于深度监督隐空间构建的语义分割改进方法

王柏涵，姜晓燕，范柳伊
（上海工程技术大学电子电气工程学院，上海 201600）

摘　要：现有卷积操作在语义分割任务中难以有效捕捉长距离区域间的关系，导致分割结果不符合人类常识。为此，提出一种基于深度监督隐空间构建的语义分割改进方法。采用“特征图-隐空间-特征图”流程，将图像空间的像素特征转换为隐空间中的节点特征，将区域之间的位置和语义关系转换为节点之间的连接权重，实现了从特征图到隐空间的特征转换。在隐空间构建过程中，使用Kullback-Leibler散度损失函数监督投影矩阵，以避免从特征图到隐空间节点的转换过程中丢失特征；使用InfoNCE损失函数监督节点特征表征与真实标签表征，使得图像特征与标签保持一致。该方法在构建的隐空间上使用图神经网络进行语义推理，学习节点之间的关系，赋予模型学习区域间语义关系的能力，从而改善分割结果中的反常识现象。在公开数据集CityScapes上的实验结果表明，相比基线分割网络，该方法的平均交并比（mIoU）为81.1%，相较于基线分割网络mIoU提升2.6个百分点，能有效提升分割结果。

关键词：语义分割；卷积神经网络；深度监督；图神经网络；反常识现象
源代码链接：https://github.com/atiuo/ProjectFuse-ImprovedGlore
中图分类号：TP391　　**文献标志码：**A　　**DOI:**10.19678/j.issn.1000-3428.0067369

Semantic Segmentation Improvement Method Based on Deep Supervision for the Construction of Latent Space

WANG Bohan，JIANG Xiaoyan，FAN Liuyi
（School of Electronic and Electrical Engineering，Shanghai University of Engineering Science，Shanghai 201600，China）

【Abstract】The existing convolution operations cannot effectively capture the relationships between long-distance regions in semantic segmentation tasks，resulting in segmentation results that do not conform to human common sense. Accordingly，a semantic segmentation improvement method based on deep supervised latent space construction is proposed. This article adopts the"feature map-hidden space-feature map"process to convert pixel features in an image space into node features in a hidden space，and convert the position and semantic relationships between regions into connection weights between nodes，thereby achieving feature conversion from the feature map to the hidden space. In the process of constructing the hidden space，the Kullback-Leibler divergence loss function is used to supervise the projection matrix，to avoid losing features during the transformation process from feature maps to hidden space nodes. It uses Information Noise Contrastive Estimation（InfoNCE）loss function to supervise node feature and real label representations，ensuring consistency between image features and labels. The proposed method uses Graph Neural Network（GNN）for semantic inference on the constructed latent space，learning the relationships between nodes and endowing the model with the ability to learn semantic relationships between regions，thereby improving the anti-common sense phenomenon in segmentation results. The experimental results on the publicly available dataset CityScapes demonstrate that compared to the baseline segmentation network，the mean Intersection over Union（mIoU）of the proposed method is 81.1%，which is 2.6 percentage points higher than that of the baseline segmentation network and can effectively improve the segmentation results.

【Key words】semantic segmentation；Convolutional Neural Network（CNN）；deep supervision；Graph Neural Network（GNN）；anti-common sense phenomenon

0　引言

语义分割是旨在图像中每个像素分配其所属类别的标签并进行密集预测。在计算机视觉领域，图像的语义分割是一项基础且重要的任务，是当前的重要研究方向和研究热点，且在场景理解[1]、医学图像分析[2]、机器人系统感知[3]等方面具有重要应用。

近年来，卷积神经网络（CNN）以其强大的纹理特征学习能力，在多个计算机视觉任务中成为研究热点。文献[4]提出一种全卷积网络（FCN），该方法将传统的全连接层替换为卷积层，使得网络可以接受任意大小的输入图像，同时实现基于深度学习、端到端的语义分割流程。RONNEBERGER 等[5]提出一种结构呈U字形的语义分割网络U-Net，该方法采

收稿日期：2023-04-06　　**修回日期：**2023-05-31
基金项目：国家自然科学基金联合项目（U2033218）；国家自然科学基金重点项目（61831018）。
通信作者 E-mail：adeclaire18@foxmail.com

用编码器-解码器结构，在编码阶段使用卷积神经网络，在解码阶段通过上采样和反卷积操作逐渐将特征映射的分辨率恢复到输入图像尺寸，同时利用跳层连接，将编码器中的特征映射与解码器中的特征映射进行融合，以提高分割结果的准确性。为扩大卷积核感受野且捕捉多尺度的上下文信息，CHEN等[6]提出空洞空间金字塔池化模块，通过融合多尺度特征提升分割性能。上述经典的语义分割方法，其主干网络均采用了传统的编解码结构及其变体，通过设计不同的特征融合策略、上采样方式，有效提升语义分割的精度。

随着深度学习的发展，针对性地对现有语义分割方法进行改进。例如，针对规则物体的边缘区域分割结果不平滑问题，DHINGRA等[7]提出使用2个图卷积网络，分别用于全局特征提取和边界细化，实现对边界区域的像素进行建模；同时还使用边界监督损失函数，通过迭代地调整边界区域的权重来优化网络，从而更好地利用边界区域的信息。针对图像分割任务中上下文特征聚合效率低的问题，YUAN等[8]认为1个像素所属的类别与其所属物体强相关，因此提出利用像素所属类别表征像素，通过加权聚合上下文信息，增强像素的特征表征，实现分割结果的提升。

尽管上述针对性改进的方法均提升了语义分割效果，解决了特定问题，但语义分割领域仍然存在2个方面的问题：1）从实现过程上分析，现有分割方法大多数是基于卷积神经网络，而传统的卷积操作本身具有较强的局限性，其感受野大小与参数量成正比，因此，卷积操作尽管可以较好地学习局部的纹理特征，但是难以高效捕捉远距离区域间的关系，而在语义分割任务中，1个像素的所属类别不仅和其相邻像素有关，还可能与图像中的任意像素有关；2）从分割结果分析，现有方法存在反常识现象，该现象体现为分割结果中2个紧密相邻的区域被分割为没有任何语义相关性的标签[9]，究其原因在于现有方法的监督学习对象仅为像素级的语义标签，难以学习到类别间的相互关系和物体内的一致性。

为增强网络针对区域间关系的学习能力，改善分割结果中的反常识现象，本文提出一种基于深度监督隐空间构建的语义分割改进方法。本文工作如下：1）针对分割结果出现反常识现象问题，提出使用“特征图-隐空间-特征图”流程，将图像特征与自然语言标签共同投影至高维空间，使用对比损失函数辅助监督训练过程，通过保持图像特征与语言特征的一致性，实现特征转换过程的准确性；2）针对“特征图-隐空间-特征图”流程中出现的分配失衡和位置信息丢失问题，使用深度监督技术约束投影矩阵分布，避免部分像素特征丢失，在隐空间构建过程中加入节点位置编码，保留特征的位置信息。

1 相关研究

1.1 语义分割模型FCN

FCN网络是一种用于语义分割的深度神经网络模型，其实现过程分为3个步骤：

1）使用预训练的卷积神经网络（如VGG[10]、ResNet[11]）提取图像特征，此过程包含多次卷积与下采样操作，得到分辨率较低的多通道特征图，该过程称为编码阶段。为了获得密集的像素级预测，FCN需要将特征图上采样到原始图像分辨率，该阶段称为解码阶段。FCN中使用的上采样方式为转置卷积，也称为反卷积，可以将输入的低分辨率特征图进行扩展，并且在扩展过程中进行卷积操作，从而得到高分辨率的特征图。转置卷积的实现过程为给定输入特征图、卷积核、步长、填充大小和目标的输出特征图尺寸，先对输入特征图进行零填充，零填充的目的是为了保证输出张量与输入张量的尺寸相同。将卷积核与特征图进行卷积操作，得到输出特征图。

2）将得到的特征图与来自主干网络相应层的特征图通过跳层连接相结合，使得网络能够同时利用高层和低层的特征、保留空间信息，提高分割的准确性。

3）合并后的特征图通过1个Softmax层，以获得分割类别的概率分布。具有最高概率的类别被选为每个像素的预测标签，结果是1个分割掩码，表明输入图像中每个像素的类别。

本文提出的基于深度监督隐空间构建的语义分割改进方法作用于FCN上采样之后、Softmax之前的特征图。由于FCN的主干网络已在ImageNet上进行预训练，因此输出特征图具有初始语义信息。

1.2 CNN与图神经网络的结合

与传统的CNN不同，图神经网络（GNN）在处理非欧几里得数据时具有明显的优势[12]。近年来，随着相关研究的发展，图神经网络被广泛应用于物体检测、语义分割、行为识别[13-14]等任务中。其中，一些方法试图通过结合CNN和GNN来捕捉抽象的语义关系，此类方法的实现流程可以被总结为“特征图-隐空间-特征图”流程。该流程主要包括3个步骤：1）将特征从坐标空间映射到隐藏的交互空间，构建1个语义感知图；2）对图进行推理，更新节点特征；3）将图映射回坐标空间，获得更新的特征图。此类方法大多数在前2个步骤上有所不同：如何从特征图中构建图，如何执行消息传递以更新节点特征。LI等[15]提出一种符号图推理方法，该方法使用图卷积网络[16]，对1组符号节点进行推理，旨在明确表示先验知识图中的不同语义，捕捉不同区域之间的长程依赖关系。CHEN等[17]提出1个轻量但高效的全局推理模块GloRe，该模块通过全局池化和加权特征传播，实现坐标空间与隐空间的映射，并通过图卷积在交互空间的小图上进行关系推理。

上述方法都指出了区域间语义关联的重要性，并

利用GNN来提取语义特征，最终达到更好的特征提取性能，有利于下游任务的开展。但此类方法存在分配失衡和位置模糊2个主要问题。分配失衡是指随着训练的进行，网络倾向于将一部分像素分配多个节点，而另一部分像素则不分配任何节点。此现象导致后者特征在GNN中缺失，降低网络性能。位置模糊是指上述方法仅根据语义相似性来衡量区域间的连通性，并未考虑位置信息。但是在语义分割任务中，决定1个像素的所属类别，不仅考虑与语义特征相近的像素特征，还要考虑位置邻近的像素特征。

本文提出的基于深度监督隐空间构建的语义分割改进方法将针对上述问题进行改进，通过加入位置编码和深度监督技术，实现从特征图到隐空间转换过程中保留位置信息，避免部分像素语义特征丢失。

1.3 对比学习

对比学习[18]通过学习如何将相似样本在特征空间中彼此聚集，并将不相似样本分散开，以达到学习特征表示的目的。对比学习中的目标是最大化同类样本之间的相似度，最小化异类样本之间的相似度。InfoNCE[19]是对比学习中常用的一种损失函数，通过将对比学习中的目标转化为最大化正确匹配样本对的相似度、最小化不匹配样本对的相似度，实现特征表示的学习。

为保证构建的隐空间表征能力足够、特征转换有效，本文提出使用基于InfoNCE的损失函数。通过最大化正样本对在特征空间中的相似度、最小化负样本对在特征空间中的相似度，使用语义标签辅助监督训练过程，达到图像特征与语义特征一致的目的。

1.4 深度监督技术

深度监督技术[20]是一种用于深度神经网络训练的方法，旨在解决传统的端到端神经网络训练中的梯度消失和梯度爆炸问题。该方法在网络结构中引入额外的辅助损失函数，并在网络的不同层次处进行监督，以增加训练期间的梯度流动，加快收敛速度并提高模型的泛化能力。深度监督技术的核心思想是通过将多个损失函数分布在网络的不同层次中，增加网络的训练深度，从而减轻梯度消失和梯度爆炸问题[21]。在深度监督技术中，网络的每个辅助损失函数都可以看作是1个短路径，将梯度反向传播到较早的层次。通过这些短路径，辅助损失函数可以提供额外的监督信号，以加速网络的收敛和优化。

为保证特征转换的质量与效率，本文基于深度监督思想，提出使用基于Kullback-Leibler(KL)散度的损失函数监督投影矩阵分布，使用基于InfoNCE的损失函数监督节点特征矩阵。

2 本文方法

2.1 整体流程

本文提出基于深度监督隐空间构建的语义分割改进方法，位于语义分割网络的解码器之后，作用于Softmax之前的特征图。该方法的整体结构如图1所示。其中，实线空心单向箭头表明该步骤在推理过程和训练过程中均有发生，虚线单向箭头表明该步骤仅存在于训练阶段，虚线双向箭头表示箭头两侧用于计算损失函数的预测值与真实值。图1中的主分支描述了本文所提方法的推理过程，该过程分为3个阶段：基于特征转换的隐空间构建、基于图神经网络的特征更新、反投影。上下2个分支仅在训练过程中发生，分别对应投影矩阵和节点特征矩阵的监督学习，这2个矩阵均在隐空间的构建阶段生成。

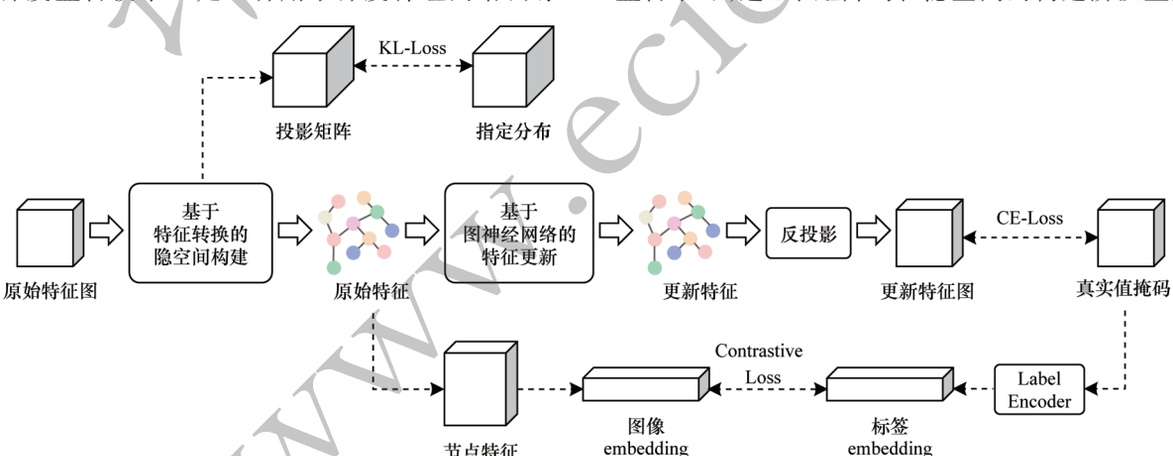

图1 基于深度监督隐空间构建的语义分割改进网络的整体流程

Fig.1 Overall procedure of semantic segmentation improvement network based on deep supervised latent space construction

2.2 基于特征转换的隐空间构建

在本文提出的语义分割网络模型中，基于特征转换的隐空间构建模块位于语义分割网络模型输出的特征图之后，其作用是将具有相似语义特征的区域投影到隐空间内的同1个节点，并将区域之间的语义相似度和位置相近度表示为节点之间的边缘权重。基于特征转换的隐空间构建流程如图2所示，隐空间的表现形式由若干个节点和带权边缘组成的图数据。

在隐空间构建的实现过程中，投影矩阵是关键模块，决定了从特征图到隐空间的映射关系。隐空间由2个矩阵表征：节点特征矩阵和邻接矩阵。节点特征矩阵表示每个节点的语义特征向量，而邻接矩阵则表示节点之间的边缘权重。

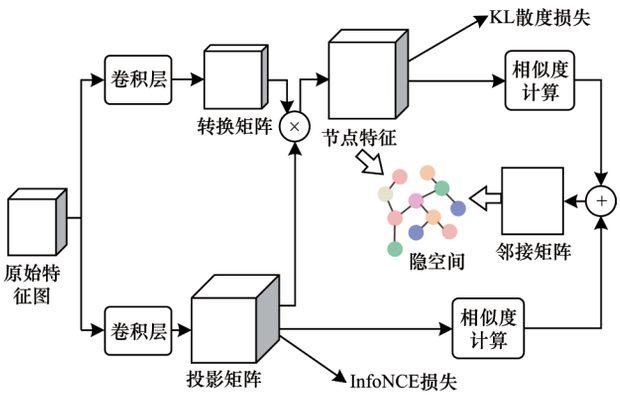

图2 基于特征转换的隐空间构建流程

Fig.2 Procedure of latent space construction based on feature transformation

2.2.1 投影矩阵

从特征图到投影矩阵的变换，最直接的方法是区域生长法。该方法通过将具有相似像素值的像素聚合到一起，形成1个连通的区域，实现对图像的分割。该方法的优点在于它的简单性和可扩展性，由于区域生长法只需要一些简单的图像处理操作和像素之间的相似度比较，因此可以快速地实现像素级的分类。但在本任务的实际操作中，使用区域生长法不仅需要手动选择种子点、相似度阈值等参数，而且由于特征图是高维度的且数值为浮点数，因此算法的实现也相应更加复杂，需要花费更多的计算时间和存储空间。为解决该问题，受到GloRe方法的启发，本文从特征图到投影矩阵的转换由投影矩阵实现。投影矩阵作用在于：将坐标空间中的像素特征投影为语义空间中的节点特征，在此过程中，坐标空间中具有相似特征的像素，将会被分配给语义空间中的同一节点。生成投影矩阵的主要依据是编码器-解码器输出的特征图。具体地，对于特征图 $\boldsymbol{X}\in\mathbb{R}^{C\times H\times W}$，其对应的投影矩阵为 $\boldsymbol{P}\in\mathbb{R}^{C\times H\times W}$，其中 C 为特征图的通道数，在本文中等于类别数，N 为隐空间中的节点个数（该数值需人为预先设定），H、W 分别为原始图像的高和宽。

投影矩阵中处于 (n,h,w) 位置的元素，其数值取值范围为(0,1)，该位置元素数值大小代表将原特征图 (h,w) 处的像素投影至第 n 个节点的概率大小。具体地，根据特征图 $\boldsymbol{X}$，可通过式(1)得到投影矩阵 $\boldsymbol{P}$：

$$\boldsymbol{P}=\mathrm{Conv}_p(\boldsymbol{X}) \tag{1}$$

其中：卷积操作的卷积核大小为1。

2.2.2 节点特征矩阵

节点特征矩阵描述了特征图投射在隐空间后生成的节点特征。生成节点特征矩阵的过程也是从像素级别的特征图转换到节点级别的隐空间、建模节点特征的过程。该过程的输入是投影矩阵和特征图，输出为节点特征 $\boldsymbol{V}\in\mathbb{R}^{N\times D}$，其中 D 为节点的特征维度。

生成节点特征矩阵的直接方法是将投影矩阵与特征图进行矩阵相乘，但存在2个问题：1)从计算量的角度看，此处构建的节点特征将用于后续的图神经网络，图神经网络中存在特征聚合与表征更新操作，若原始特征图通道数较多，则用此方法生成的节点特征维度也相应较高，导致图神经网络阶段需要较高的计算量，增加推理时间，降低运行效率；2)从特征分布看，像素级别的特征图转换到节点级别的隐空间，两者存在一定的特征分布差异，而直接使用矩阵乘法难以消除这种分布差异。为解决该问题，本文方法将先对特征图进行特征转换，再生成节点特征矩阵。具体地，首先将像素特征 $\boldsymbol{X}$ 转换为语义特征 $\boldsymbol{S}\in\mathbb{R}^{D\times H\times W}$，其目的在于升维或降维。当特征图维度较高时，为降低在后续操作中的参数量，将高维度特征转换为低维度特征；当特征图维度较低时，为增加网络的描述能力，将低维度特征转换为高维度特征。该转化过程由卷积操作实现：

$$\boldsymbol{S}=\mathrm{Conv}_t(\boldsymbol{X}) \tag{2}$$

其中：卷积操作的卷积核大小为1。将投影矩阵与语义特征进行矩阵相乘，得到节点特征矩阵，即可实现从像素级别到节点级别的特征转换：

$$\boldsymbol{V}=\boldsymbol{P}\cdot\boldsymbol{S}^{\mathrm{T}} \tag{3}$$

2.2.3 邻接矩阵

本文提出使用邻接矩阵 $\boldsymbol{A}$ 描述节点特征的相似度 $\boldsymbol{A}_{\mathrm{sem}}$ 及其位置相近度 $\boldsymbol{A}_{\mathrm{pos}}$，其目的在于以更加高效的方式，将具有相似语义且距离较近的2个节点使用更高的权重边缘连接。具体地，将投影矩阵 $\boldsymbol{P}$ 展平即可得到 N 个长度为 $H\times W$ 的一维向量 $\boldsymbol{p}_i$，该向量即为隐空间中第 i 个节点的位置编码。本文将进行形状变换后的投影矩阵记为 $\boldsymbol{P}'\in\mathbb{R}^{C\times H\times W}$。因此，将 $\boldsymbol{P}'$ 与其自身计算Cosine相似度，即可得到描述节点位置相近度的矩阵：

$$\boldsymbol{A}_{\mathrm{pos}}=\frac{\boldsymbol{p}_i\cdot\boldsymbol{p}_j}{|\boldsymbol{p}_i|\cdot|\boldsymbol{p}_j|} \tag{4}$$

其中：分子表示2个向量的点积；分母表示2个向量的模长。类似地，已有节点特征矩阵，展平其每个通道即可得到该节点的语义编码 $\boldsymbol{v}_i$，将进行形状变换后的节点特征矩阵 $\boldsymbol{V}'$ 与其自身计算Cosine相似度，即可得到描述节点位置相近度 $\boldsymbol{A}_{\mathrm{sem}}$。最后将2个矩阵相加，即可得到目标邻接矩阵 $\boldsymbol{A}$。至此，从特征图到隐空间的转换完成，隐空间由若干个节点及边缘特征组成，节点特征与边缘特征分别由 $\boldsymbol{V}$ 和 $\boldsymbol{A}$ 表示。

2.3 基于图神经网络的特征更新

基于图神经网络的特征学习模块将节点特征和邻接矩阵作为输入，输出最新的节点特征 $\tilde{\boldsymbol{V}}$。该模块的目的是对图进行特征学习，捕捉语义空间中的关系以及坐标空间中的位置属性。该过程表达式如下：

$$\tilde{\boldsymbol{V}}=\boldsymbol{F}_g(\boldsymbol{V},\boldsymbol{A},W) \tag{5}$$

其中：W 代表图神经网络的可学习参数。

本文使用图卷积网络(GCN)对得到的节点特征进行学习。图卷积网络是一种用于处理图数据的卷积神经网络,它在节点上进行卷积操作,类似于在图像上进行像素卷积操作。图 3 所示为给定 1 组图数据,图卷积网络将节点特征与其邻居特征进行聚合,从而生成新的节点特征表示。

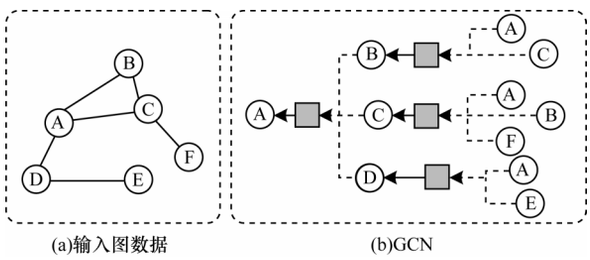

图 3 图卷积网络的图传递示意图

Fig.3 Schematic diagram of graph transfer in graph convolutional network

对于节点特征矩阵和邻接矩阵,图卷积层执行消息计算和特征聚合操作,并输出节点的更新特征 $\tilde{\boldsymbol{V}}$。在本文方法中,为保留节点本身的特征,将对节点本身和其邻居进行不同的特征计算:

$$\boldsymbol{m}_v = W_s(\boldsymbol{h}_v) \tag{6}$$

$$\boldsymbol{m}_u = \mathrm{AGG}\{W_n(\boldsymbol{h}_u)\} \tag{7}$$

其中:W_s 和 W_n 表示针对节点表征进行的线性变换;$\boldsymbol{m}_v$ 表示节点本身特征;$\boldsymbol{m}_u$ 代表节点邻域聚合后的特征;AGG 操作可以是 1 个简单的平均化、最大化或求和,以汇总邻居的特征。通过式(8)获得更新的节点特征:

$$\tilde{\boldsymbol{V}} = \sigma\{\mathrm{SUM}(\boldsymbol{m}_u, \boldsymbol{m}_v)\} \tag{8}$$

其中:σ 为激活层,用于增加网络的非线性。

2.4 反投影模块

反投影模块的作用在于:将更新后的隐空间语义特征 $\tilde{\boldsymbol{V}}$ 转换为特征图 $\tilde{\boldsymbol{X}} \in \mathbb{R}^{C\times H\times W}$,该过程可视为将像素特征投影为节点特征的逆过程。具体地,首先将隐空间特征转换至语义特征,该过程是式(3)的逆向操作,接着将得到的语义特征还原至原始尺寸的特征图,如式(9)所示:

$$\tilde{\boldsymbol{X}} = \mathrm{Conv}_t'(\boldsymbol{P}^{\mathrm{T}} \cdot \tilde{\boldsymbol{V}}) \tag{9}$$

至此,特征图经过图神经网络,完成了特征更新,且整个流程是端到端、可训练、可进行反向传播的。

2.5 深度监督与损失函数

深度监督技术可通过在网络结构中引入辅助损失函数,并在不同层次处提供额外的监督信号,达到加速网络收敛的目的。为保证特征转换的质量与效率,本文基于深度监督思想,提出使用 KL 散度损失函数监督投影矩阵分布,使用 InfoNCE 损失函数监督节点特征矩阵。

2.5.1 使用 KL 散度损失函数监督投影矩阵的方法

投影矩阵的生成与应用是端到端的,投影矩阵的数值可通过训练学习得到。但是在实践过程中,本文发现原始的“特征图-隐空间-特征图”流程中存在分配失衡问题:随着训练的进行,网络倾向于将一部分像素分配给多个节点,而另一部分像素不分配任何 1 个节点。投影矩阵作为特征图与隐空间相互转换的关键模块,其数据分布直接影响着特征转换的效果。因此,本文基于深度监督思想,提出使用基于 KL 散度的损失函数监督投影矩阵分布。KL 散度损失函数是一种用于衡量 2 个概率分布之间的差异性指标,其值越小表示 2 个分布越相似,反之则越不相似。对于 2 个形状同为 $N\times(H\times W)$ 的特征分布 $\boldsymbol{P}'$、$\boldsymbol{Q}$,其 KL 散度计算式如下:

$$\mathrm{Loss}_{\mathrm{KL}}(\boldsymbol{P}', \boldsymbol{Q}) = \sum_{i=1}^{N} \boldsymbol{p}'_i \log_a \frac{\boldsymbol{P}'_i}{\boldsymbol{Q}_i} \tag{10}$$

其中:$\boldsymbol{P}'_i$ 和 $\boldsymbol{Q}_i$ 分别表示 $\boldsymbol{P}'$、$\boldsymbol{Q}$ 第 i 层的概率分布。

本文提出 2 种使用 KL 散度损失函数监督投影矩阵的方法。第 1 种方法是首先对真实值掩码进行如式(1)所示的转换操作,将得到的矩阵称为投影参考矩阵,记作 $\hat{\boldsymbol{P}}$,然后计算 $\hat{\boldsymbol{P}}$ 与 $\boldsymbol{P}'$ 的 KL 散度损失。此方法通过显式建模从特征图到隐空间过程中投影矩阵的生成方式,使得网络学习到的投影矩阵能够将原特征图中具有相似语义特征的像素分配到同 1 个节点。第 2 种方法则是使用与投影矩阵形状相同的均匀分布矩阵 $\boldsymbol{Q}$,通过最小化 $\boldsymbol{P}'$ 与 $\boldsymbol{Q}$ 的 KL 散度,使得投影矩阵的数值分布接近于均匀分布。此方法可视为第 1 种方法的反面,可以减少低置信度的出现频率,进而避免转换过程中的特征丢失。

2.5.2 使用 InfoNCE 损失函数监督节点特征矩阵的方法

为确保本文方法中从特征图到隐空间的转换过程有效,并且能够构建具有足够表征能力的隐空间,本文提出采用 InfoNCE 损失函数,使用分割真实掩码监督得到的节点特征。具体地,首先,针对样本图像 $\boldsymbol{I}_x$ 得到的节点特征,使用多层感知机(MLP),将节点特征转换为特征向量 $\boldsymbol{q} \in \mathbb{R}^{1\times D}$。接着,$\boldsymbol{I}_x$ 对应的真实值编码为 $\boldsymbol{l}^+ \in \mathbb{R}^{1\times C}$ 作为对比损失中的正样本,其元素数值为 1 或 0,代表掩膜中是否出现某个类别。同时随机生成 K 个负样本特征向量 $\boldsymbol{l}^- \in \mathbb{R}^{K\times C}$,每个特征向量的元素数值同样为 1 或 0,生成的 $\boldsymbol{l}^-$ 应满足:$\sum_{i=1}^{K} \boldsymbol{l}_i^- \cdot \boldsymbol{l}^+ = 0$。使用多层感知机将 $\boldsymbol{l}^+$ 转换为 $\boldsymbol{z}^+ \in \mathbb{R}^{1\times D}$,$\boldsymbol{l}^-$ 转换为 $\boldsymbol{z}^- \in \mathbb{R}^{K\times D}$。至此,对比损失的计算式如下:

$$\mathrm{Loss}_{\mathrm{NCE}}(\boldsymbol{q}, \boldsymbol{z}^+, \boldsymbol{z}^-) = -\frac{1}{K}\log_a \frac{\exp(\boldsymbol{q}\cdot\boldsymbol{z}^+/\tau)}{\sum_{i=1}^{K}\exp(\boldsymbol{q}\cdot\boldsymbol{z}_i^-/\tau)} \tag{11}$$

2.5.3 损失函数

本文采用的损失函数由 3 个部分组成:预测结果与真实标签之间的像素级损失、投影矩阵与目标矩阵的 KL 散度损失、节点特征与语义标签的

InfoNCE对比损失。损失函数计算式如下：

$$\text{Loss}=\text{Loss}_{\text{CE}}(\boldsymbol{Y},\hat{\boldsymbol{Y}})+\alpha\cdot\text{Loss}_{\text{KL}}(\boldsymbol{P}',\boldsymbol{Q})+\beta\cdot\text{Loss}_{\text{NCE}}(\boldsymbol{q},\boldsymbol{z}^{+},\boldsymbol{z}^{-}) \tag{12}$$

其中：Loss_{CE}表示计算输出的分割掩码$\boldsymbol{Y}$与真实值$\hat{\boldsymbol{Y}}$的交叉熵损失；Loss_{KL}表示计算投影矩阵$\boldsymbol{P}'$与目标同尺寸矩阵$\boldsymbol{Q}$的KL散度损失；Loss_{NCE}表示计算图像表征与标签正负样本的对比损失；α、β为可调节参数，用于减小2个损失值的差异。交叉熵损失的计算式如下：

$$\text{Loss}_{\text{CE}}(\boldsymbol{Y},\hat{\boldsymbol{Y}})=-\sum_{i=1}^{H\times W}\sum_{j=1}^{C}\hat{\boldsymbol{Y}}_{ij}\log_a\boldsymbol{Y}_{ij} \tag{13}$$

其中：$\hat{\boldsymbol{Y}}_{ij}$表示第$i$个样本是否为第$j$类，取值为0或1；$\boldsymbol{Y}_{ij}$表示模型将第$i$个像素预测为第$j$个类别的概率，取值范围[0,1]。

3 实验与结果分析

3.1 数据集与评价指标

本文在公开数据集CityScapes[22]上训练和评估该方法。CityScapes由德国斯图加特大学的计算机视觉小组于2016年创建，是1个大规模的城市街景数据集。该数据集拍摄了不同天气状况下，德国和瑞士50个城市的街景图像，包括城市中心、住宅区、高速公路和乡村道路等场景。该数据集提供5 000张精细标注数据，为避免模型过拟合、增强模型泛化能力，该数据仅公开了精细标注数据中4 500个样本的真实掩码，用于研究人员的训练与验证，另有500个样本未提供真实掩码，需在线提交验证。在实验中，本文采取2 975个样本用于训练，500个样本用于测试。每个样本都针对19个类别分别进行像素级标注。

本文从模型运行效果和运行效率2个方面评价提出的网络模型。在模型运行效果方面，使用平均交并比（mIoU，计算中用m_{mIoU}）评价模型分割结果的精度，其计算式如下：

$$m_{\text{mIoU}}=\frac{T_{\text{TP}}+F_{\text{FN}}}{T_{\text{TP}}+F_{\text{FN}}+T_{\text{TN}}+F_{\text{FP}}} \tag{14}$$

其中：T_{TP}表示针对真实值为真的像素，模型预测也为真；F_{FN}表示针对真实值为假的像素，模型预测也为假；T_{TN}表示针对结果为假的像素，模型预测为真；F_{FP}表示针对真实值为真的像素，模型预测为假。mIoU取值范围为[0,1]，该数值越大，分割结果越准确，网络效果越好。

在模型运行效率方面，使用每秒浮点运算数（FLOPS）和模型参数量衡量所提方法的计算复杂度。

3.2 实验环境与设置

本文实验环境基于64位的Ubuntu18.04操作系统，使用PyTorch 1.7.0框架进行网络训练，计算机配置显卡为NVIDA GeForce GTX 3090，CUDA版本为11.0。在训练过程中，本文采用Adam优化器，初始学习率为0.01，按照如下poly策略，动态设置学习率：

$$l_{\text{lr}}=l_{\text{lr init}}\times\left(1-\frac{i_{\text{iter}}}{i_{\text{iter total}}}\right)^{\text{power}} \tag{15}$$

其中：l_{lr}表示学习率；$l_{\text{lr init}}$表示初始学习率；i_{iter}表示当前迭代次数；$i_{\text{iter total}}$表示最大迭代次数；power设置为0.9。在训练过程中采用的数据增强手段包括颜色变换、随机翻转、裁剪缩放，输入数据尺寸统一为320×320像素，Batch size设置为16。本文使用像素级交叉熵损失函数、KL散度损失函数和InfoNCE损失函数联合优化网络，忽略未标记的背景像素点。在评估过程中，使用相应数据集的原始图片，不对输入图片进行任何形式的数据变换。

3.3 消融实验

本文的消融实验包括2个部分：1）通过删改不同模块配置，验证不同深度监督损失函数的有效性、寻找最佳的生成边缘权重方式；2）通过改变建图过程中的节点数目、节点维度，寻找最佳隐空间构建的参数取值。表1所示为消融实验结果。

表1 消融实验结果

Table 1 Results of ablation experiment

方法	投影参考矩阵	均匀分布	语义相似性	位置相近度	InfoNCE	FLOPS/10^9	参数量/10^6	mIoU/%
Baseline						43.49	54.68	78.5
方法①	√		√			42.78	53.16	78.8
方法②		√	√			42.71	53.16	79.8
方法③		√		√		42.94	53.16	79.7
方法④		√	√	√		43.19	53.16	80.3
方法⑤		√	√	√	√	43.19	54.31	81.1

本文使用FCN+GloRe作为基线方法，其中FCN的编码器为ResNet101，GloRe模块位于FCN编码器之后，该模块的节点数为2 048个，节点特征维度为1 024。从方法①和方法②可观察到：本文方法基于深度监督思想，使用KL散度损失监督投影矩阵过程中，当投影矩阵分布的学习对象为均匀分布时，相比投影参考矩阵分布，mIoU提高1.0个百分点；而Baseline方法没有对投影矩阵进行任何形式的监督或约束，其生成过程是完全隐式的。为寻找生成边缘权重的最佳方法，实验比较3种不同的方法：仅考虑语义相似性、仅考虑位置相似性、同时考虑语义和位置相似性。3种方式分别对应表1中的方法②、方法③、方法④。同时考虑节点间的语义相似性和位置相近度，相比原始的仅考虑语义相似性，mIoU可提升0.5个

百分点；而 Baseline 方法中的邻接矩阵是可训练的，随着网络的迭代在不断变化。对比方法④和方法⑤，使用 InfoNCE 损失函数监督对比节点特征与标签特征，mIoU 提升 0.8 个百分点。综上，与 Baseline 方法相比，本文方法在构建隐空间过程中基于深度监督思想，一方面使用 KL 散度损失函数监督投影参考矩阵分布，以解决原方法中的分配失衡问题；另一方面根据区域之间的语义相似性与位置相似性，显性建模邻接矩阵，以解决原方法中的位置模糊问题。对比 Baseline 方法，方法⑤的 mIoU 提升 2.6 个百分点，降低了模型参数量与 FLOPS，验证本文方法的优越性。

隐空间作为本文提出方法的关键组成部分，其节点个数与单个节点特征维度需提前设定。表 2 所示为针对分辨率为 1 024×2 048 像素、类别数为 19 的 CityScapes 数据集，隐空间参数对模型性能的影响，加粗表示最优数据。从表 2 可以看出，当节点个数设置为 32 个、单个节点特征的维度为 16 时，模型性能达到最佳效果。

表 2　隐空间参数对模型性能的影响

Table 2　The influence of hidden space parameters on model performance

节点数/个	节点特征维度	mIoU/%
32	8	80.7
32	16	**81.1**
32	32	80.9
32	64	80.8
8	16	80.6
16	16	80.8
64	16	81.0

3.4　实验结果与 State-of-Art（SOTA）方法对比

表 3 所示为本文方法与其他 SOTA 方法在 CityScapes 数据集上的实验结果对比，其中，“*”表示该方法使用多尺度训练。本文方法采用 ResNet101 作为主干网络、FCN 作为基础预分割模型、单尺度监督训练，隐空间的节点个数与节点维度选择表 2 中表现最佳的配置，即节点个数为 32 个、节点特征维度为 16；GloRe[17] 表示 Baseline 方法。文献[23]提出的 SegFormer 是一种编码器基于多层 Transformer 结构、解码器基于多层感知机的语义分割方法。CSWin-T[28]、HRViT-b2[29] 则是 2 种基于 Transformer 结构、针对语义分割任务设计的编码器。因此，CSWin-T 和 HRViT-b2 采用 2 种更新的编码器连接 SegFormer 解码器的语义分割网络。

从表 3 可以看出，实时语义分割方法 BiSeNet 系列和 ESANet 方法在 FLOPS 和参数量方面具有明显优势，但是在精度上有所不足。本文提出的方法、CSWin-T 以及 HRViT-b2 3 种方法在分割精度上具有一定优势。其中，CSWin 和 HRViT 均为基于 Transformer 的大型语义分割网络。尽管 HRViT 在参数量及计算量方面进行重点优化与改进，但相比其他方法，此类方法的计算量与参数量仍然较大。本文提出的方法采用 GloRe 流程，与 Transformer 类方法实现原理完全不同，本文方法的优势在于以图隐式表征图像，通过图神经网络学习区域间的关系，且因隐空间的存在，本文方法未来在多模态数据融合方向具有一定潜力。

表 3　不同方法在 CityScapes 数据集上的实验结果对比

Table 3　Experimental results comparison among different methods on the CityScapes dataset

方法	主干网络	FLOPS/10^9	参数量/10^6	mIoU/%
BiSeNetV1*[24]	ResNet18	8.05	23.37	78.86
BiSeNetV2*[25]	ResNet18	7.78	22.23	77.08
ESANet[26]	ResNet50	12.51	25.14	79.20
PSPNet*[27]	ResNet101	14.62	49.01	79.70
OCR*[8]	ResNet101	70.50	72.40	80.60
SegFormer[23]	MiT-B2	67.97	90.81	81.00
CSWin-T+SFH[28]	CSWin-T	60.86	75.18	81.56
HRViT-b2+SFH[29]	HRViT-b2	50.87	59.37	82.81
GloRe[17]	ResNet101	43.63	54.68	78.50
本文方法	ResNet101	43.51	54.31	81.10

相比基线方法 GloRe，本文提出的方法在参数量与计算量有轻微减少的同时，精度获得了明显的提升。其原因为本文方法提出将 GloRe 模块应用于分割网络之后而非编码器之后，可有效避免 GNN 中节点维度过高，减少参数量，本文方法提出深度监督隐空间构建过程，避免了原始方法中特征转换步骤的特征丢失与特征冗余，有利于后续特征更新过程，同时，由于深度监督过程仅存在于训练阶段，因此推理过程的计算量并没有显著增加。

3.5　实验结果可视化

3.5.1　投影矩阵可视化结果与分析

作为本方法的重要组成部分，投影矩阵在将特征从像素转换至隐空间方面起着关键作用。为深入探究 KL 散度约束的有效性，从数据集中随机选择 1 个样本，可视化在不同训练阶段、KL 散度监督不同对象情况下，投影矩阵的分配情况。考虑到原始高维投影矩阵不易可视化，本文采用在通道维度上取其最大值的方法，在投影矩阵的每个像素位置取通道最大值，进行可视化分析。

投影矩阵在不同训练阶段的可视化结果如图 4 所示。投影矩阵某个位置在通道维度的最大值越大，表明网络将该像素分配给某个隐空间节点的置信度越高，体现到图像上即为该位置的亮度越亮。对比图 4 中 2 行的差异，可以看到第 1 行，在使用了 KL 散度损失函数监督投影矩阵与投影参考矩阵的分布差异时，随着迭代的进行，图片亮度未产生明显变化。而在使用 KL 散度损失函数监督投影矩阵与同形状的均匀分布矩阵的分布差异时，随着迭代的

进行，图片亮度越来越低，即网络将该位置像素分配给某个节点的概率差别变小。此现象表明KL散度损失函数的使用在一定程度上起到了困难像素挖掘的作用，促使更多像素参与到了GNN的学习中，从而实现了对分割结果的改进。

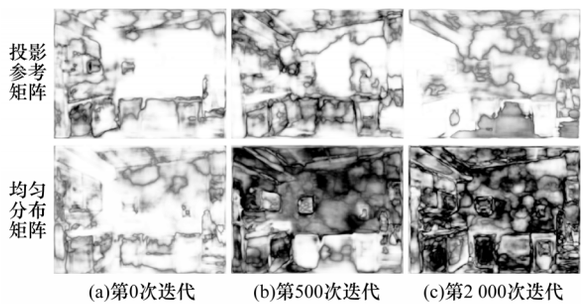

图4 投影矩阵在不同训练阶段的可视化结果

Fig.4 Visualization results of the projection matrix at different training stages

3.5.2 分割结果可视化结果与分析

本节从CityScapes数据集的验证集样本中，挑取2个具有一定代表性的样本，分别对其分割结果进行可视化。图5所示为不同方法语义分割的可视化结果。第1列为样本A，第2列为样本B。

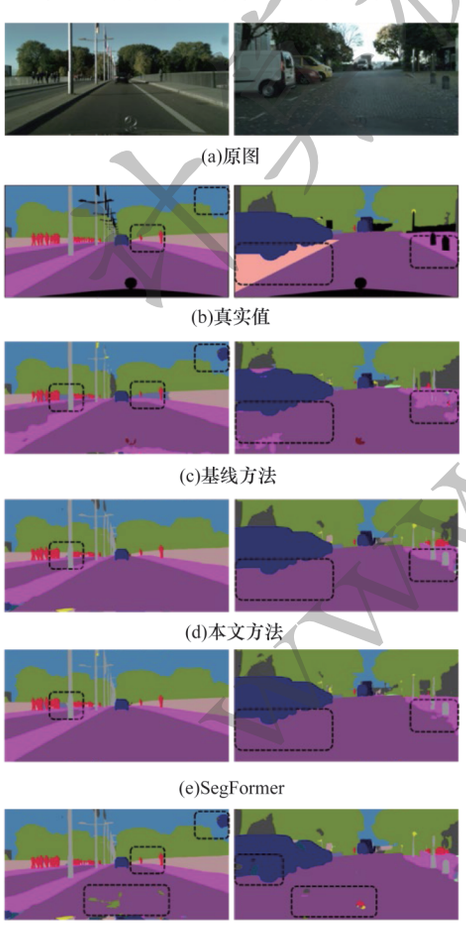

图5 不同方法的语义分割可视化结果

Fig.5 Visualization results of semantic segmentation among different methods

从样本A中可以看出：基线方法在图片右上角的“天空”中分割出一小块“汽车”，此反常识现象在本文方法中得到完全修正；基线方法在图片水平方向中间区域的“围墙”远端分割出“墙”，而在本文方法中此现象得到了极大改善。分析其原因可以发现“围墙”和“墙”两者的纹理特征有一定差别，但语义相近。基线方法主要根据纹理特征进行分割，而本文方法考虑不同类别语义特征之间的关联性，从而改进分割效果。在样本B中，基线方法在图片左侧的“汽车”上方分割出“人行道”、在图片下方的“道路”中间分割出“摩托车”，此类反常识现象均在本文方法的分割结果中得到了极大改善。样本B的真实值显示，画面左侧及下方存在“地形”类别，但观察其输入图片，发现此区域的纹理特征与地面基本一致。此类标注是从“汽车可行驶”条件为标注依据，即使停车位与地面纹理特征一致，但由于上方存在停放车辆，因此此区域未被分割为“道路”。此类场景需要更多先验知识进行理解，对比样本B中的图5(c)~图5(f)可知，现有分割方法在此类场景下仍存在改进空间。

4 结束语

本文提出一种基于深度监督隐空间构建的语义分割改进方法。该方法采用“特征图-隐空间-特征图”范式，在隐空间构建过程中，加入位置编码保留特征的位置信息，以缓解原流程中的“位置模糊”问题；使用KL散度损失函数监督投影矩阵，以缓解原流程中的“分配失衡”问题。为确保从特征图到隐空间的转换过程有效，并且能够构建具有足够表征能力的隐空间，本文采用InfoNCE损失函数，使用分割真实掩码标签，监督转换得到节点特征。在CityScapes数据集上的实验结果表明，本文网络能够有效改善语义分割中出现的不规则、反常识区域，并在精度方面相较于基线有显著提升。后续将对本文提出的特征转换及隐空间构建方法进行改进，使其适用于图像、自然语言、激光点云等多模态任务，通过将多模态特征统一投影至同一隐空间，增强模型的表征能力，进而提升其在下游任务中的表现。

参考文献

[1] NADEEM U, SHAH S A, SOHEL F, et al. Deep learning for scene understanding[J]. Handbook of Deep Learning Applications, 2019, 5: 21-51.

[2] 褚张晴晴，钟志强，颜子夜，等. 基于特征融合与注意力机制的脑肿瘤分割算法[J]. 计算机工程，2023，49(10)：154-161.
CHU Z Q Q, ZHONG Z Q, YAN Z Y, et al. Brain tumor segmentation algorithm based on feature fusion and attention mechanism[J]. Computer Engineering, 2023, 49(10): 154-161. (in Chinese)

[3] MIYAMOTO R, NAKAMURA Y, ADACHI M, et al. Vision-based road-following using results of semantic segmentation for autonomous navigation[C]//Proceedings of the 9th International Conference on Consumer Electronics. Washington D. C., USA: IEEE Press, 2019: 174-179.
[4] SHELHAMER E, LONG J, DARRELL T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 3431-3440.
[5] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation[C]// Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention. Berlin, Germany: Springer, 2015: 234-241.
[6] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. DeepLab: semantic image segmentation with deep convolutional Nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(4): 834-848.
[7] DHINGRA N, CHOGOVADZE G, KUNZ A. Border-SegGCN: improving semantic segmentation by refining the border outline using graph convolutional network[C]// Proceedings of International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2021: 865-875.
[8] YUAN Y H, CHEN X L, WANG J D, et al. Object-contextual representations for semantic segmentation[EB/OL]. [2023-03-01]. https://arxiv.org/abs/1909.11065v2.
[9] AKULA A R, WANG K Z, LIU C S, et al. CX-ToM: counterfactual explanations with theory-of-mind for enhancing human trust in image recognition models[J]. iScience, 2022, 25(1): 103581.
[10] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. [2023-03-01]. https://www.arxiv.org/pdf/1409.1556.pdf.
[11] SRIVASTAVA R K, GREFF K, SCHMIDHUBER J. Training very deep networks[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems. New York, USA: ACM Press, 2015: 28.
[12] 袁立宁,胡皓,刘钊. 基于多通道图卷积自编码器的图表示学习[J]. 计算机工程,2023,49(2):150-160,174.
YUAN L N, HU H, LIU Z. Graph representation learning based on multi-channel graph convolutional autoencoders[J]. Computer Engineering, 2023, 49(2): 150-160, 174. (in Chinese)
[13] 刘宽,奚小冰,周明东. 基于自适应多尺度图卷积网络的骨架动作识别[J]. 计算机工程,2023,49(10):264-271.
LIU K, XI X B, ZHOU M D. Skeleton action recognition based on adaptive multi-scale graph convolution network[J]. Computer Engineering, 2023, 49(10): 264-271. (in Chinese)
[14] 李威庭. 基于改进图卷积神经网络的面部动作单元识别算法研究[D]. 北京:北京工业大学,2021.
LI W T. Research on improved graph convolutional neural network computing model for facial action unit recognition[D]. Beijing: Beijing University of Technology, 2021. (in Chinese)
[15] LI Y, GUPTA A. Beyond grids: learning graph representations for visual recognition[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. New York, USA: ACM Press, 2018: 9225-9235.
[16] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[EB/OL]. [2023-03-01]. http://arxiv.org/abs/arXiv:1609.02907.
[17] CHEN Y, ROHRBACH M, YAN Z, et al. Graph-based global reasoning networks[C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2019: 433-442.
[18] YOU Y N, CHEN T L, SUI Y D, et al. Graph contrastive learning with augmentations[EB/OL]. [2023-03-01]. http://www.arXiv:2010.13902.
[19] YU Q Y, LOU J M, ZHAN X Y, et al. Adversarial contrastive learning via asymmetric InfoNCE[C]//Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2022: 53-69.
[20] LEE C Y, XIE S, GALLAGHER P, et al. Deeply-supervised Nets[EB/OL]. [2023-03-01]. http://de.arxiv.org/pdf/1409.5185.
[21] ZHANG L, CHEN X, ZHANG J, et al. Contrastive deep supervision[C]//Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2022: 1-19.
[22] CORDTS M, OMRAN M, RAMOS S, et al. The CityScapes dataset[C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2015: 2.
[23] XIE E Z, WANG W H, YU Z D, et al. SegFormer: simple and efficient design for semantic segmentation with transformers[EB/OL]. [2023-03-01]. https://arxiv.org/abs/2105.15203.
[24] YU C Q, WANG J B, PENG C, et al. BiSeNet: bilateral segmentation network for real-time semantic segmentation[C]// Proceedings of the European Conference on Computer Vision. Berlin, Germany: Springer, 2018: 325-341.
[25] YU C Q, GAO C X, WANG J B, et al. BiSeNetV2: bilateral network with guided aggregation for real-time semantic segmentation[J]. International Journal of Computer Vision, 2021, 129: 3051-3068.
[26] SEICHTER D, KÖHLER M, LEWANDOWSKI B, et al. Efficient RGB-D semantic segmentation for indoor scene analysis[C]//Proceedings of International Conference on Robotics and Automation. Washington D. C., USA: IEEE Press, 2021: 13525-13531.
[27] ZHAO H S, SHI J P, QI X J, et al. Pyramid scene parsing network[C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2017: 2881-2890.
[28] DONG X Y, BAO J M, CHEN D D, et al. CSWin Transformer: a general vision Transformer backbone with cross-shaped windows[C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2022: 12124-12134.
[29] GU J Q, KWON H, WANG D L, et al. Multi-scale high-resolution vision transformer for semantic segmentation[C]// Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2022: 12094-12103.

编辑 薛晋栋