软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software [doi: 10.13328/j.cnki.jos.007155]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

GT-4S: 基于图 Transformer 的场景草图语义分割*

张拯明[1,2], 郭 燕[1,2], 马翠霞[1,2], 邓小明[1,2], 王宏安[1,2]

[1](人机交互北京市重点实验室 (中国科学院 软件研究所), 北京 100190)

[2](中国科学院大学 计算机科学与技术学院, 北京 100049)

通信作者: 马翠霞, E-mail: cuixia@iscas.ac.cn

摘 要: 场景草图由多个前、背景物体组成, 能够直观、概括地表达复杂的语义信息, 在现实生活中有着广泛的实际应用, 逐渐成为计算机视觉和人机交互领域的研究热点之一. 作为场景草图语义理解的基础任务, 场景草图语义分割的相关研究相对较少, 现有的方法多是对自然图像语义分割的方法进行改进, 不能克服草图自身的稀疏性和抽象性等特点. 针对以上问题, 直接从草图笔画入手, 提出一种图 Transformer 模型结合草图笔画的时空信息来解决自由手绘场景草图语义分割任务. 首先将矢量场景草图构建成图结构, 笔画表示为图的节点, 笔画在时序和空间上的关联表示为图的边. 然后通过边增强的 Transformer 模块捕获笔画的时空全局上下文信息. 最后将编码后的时空特征进行多分类优化学习. 在 SFSD 场景草图数据集上的实验结果表明, 所提方法可以利用笔画时空信息对场景草图进行有效的语义分割, 实现优秀的性能.

关键词: 场景草图; 语义分割; 图; Transformer

中图法分类号: TP391

中文引用格式: 张拯明, 郭燕, 马翠霞, 邓小明, 王宏安. GT-4S: 基于图Transformer的场景草图语义分割. 软件学报. http://www.jos.org.cn/1000-9825/7155.htm

英文引用格式: Zhang ZM, Guo Y, Ma CX, Deng XM, Wang HA. GT-4S: Graph Transformer for Scene Sketch Semantic Segmentation. Ruan Jian Xue Bao/Journal of Software (in Chinese). http://www.jos.org.cn/1000-9825/7155.htm

GT-4S: Graph Transformer for Scene Sketch Semantic Segmentation

ZHANG Zheng-Ming[1,2], GUO Yan[1,2], MA Cui-Xia[1,2], DENG Xiao-Ming[1,2], WANG Hong-An[1,2]

[1](Beijing Key Laboratory of Human-computer Interaction (Institute of Software, Chinese Academy of Sciences), Beijing 100190, China)

[2](School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: The scene sketch is made up of multiple foreground and background objects, which can directly and generally express complex semantic information. It has a wide range of practical applications in real life and has gradually become one of the research hotspots in the field of computer vision and human-computer interaction. As the basic task of the semantic understanding of scene sketch, scene sketch semantic segmentation is rarely studied. Most of the existing methods are improved from the semantic segmentation of natural images, which cannot overcome the sparsity and abstraction of sketches. To solve the above problems, this study proposes a graph Transformer model directly from sketch strokes. The model combines the temporal-spatial information of sketch strokes to solve the semantic segmentation task of free-hand scene sketches. First, the vector scene sketch is constructed into a graph with strokes as the nodes of the graph and temporal and spatial correlations between strokes as the edges of the graph. The temporal-spatial global context information of the strokes is then captured by the edge-enhanced Transformer module. Finally, the encoded temporal-spatial features are optimized for multi-classification learning. The experimental results on the SFSD scene sketch dataset show that the proposed method can effectively segment scene sketches using stroke temporal-spatial information and achieve excellent performance.

Key words: scene sketch; semantic segmentation; graph; Transformer

---

* 基金项目: 国家自然科学基金 (62272447); 北京市自然科学基金-海淀原始创新联合基金 (L222008)
收稿时间: 2023-08-11; 修改时间: 2023-10-21; 采用时间: 2023-12-30; jos 在线出版时间: 2024-05-08

随着各种触屏设备的不断发展, 用户能够直接在屏幕上进行草图手绘, 模仿传统纸笔的交互方式, 使得人机交互更加自然直观. 草图是一种简单、有效的交流工具, 有利于用户抽象思维的直观表达. 草图不仅提升了用户的参与度和表达效率, 还促进了创造性思维的发展, 为人们的交流和合作带来了更大的便利和创新空间. 因此草图逐渐成为一种重要的人机交互方式, 得到了广泛的应用. 草图具有文本的时序语义特性, 又具有图像一图胜千言的效果, 人们看到一幅草图后往往能马上联想到其所表达的语义信息. 但是由于草图自身的抽象性、稀疏性、多样性, 对于计算机来说仍然难以实现草图智能理解. 当前对于草图理解和分析的研究主要包括草图分割[1,2]、草图生成[3,4]、草图识别[5,6]、基于草图的图像检索[7,8]等. 单物体草图的研究已较为成熟, 目前越来越多的工作将重点转向场景草图[9−12]. 事实上, 场景草图在实际应用中更加常见, 人们在绘制草图时更倾向于绘制一个完整的场景[13], 场景中包含语义类别多样、空间位置和遮挡情况复杂的多个对象[14]. 因此, 对于场景草图的研究是草图理解任务中不可或缺的一部分.

与图像语义分割不同, 草图语义分割是从笔画层次对草图中的不同类别进行划分, 是草图理解的一个基本任务. 根据分割粒度和语义标签类型, 草图语义分割可以分为单目标草图语义分割和场景级草图语义分割[15]. 当前大多数研究关注于单目标草图语义分割[1,16−19] (例如将手绘的飞机草图按照机头、机身、机翼、机尾对笔画进行分割), 从场景层面对草图语义分割进行的研究工作还比较少[20−22] (将草图场景中包含的多个前背景物体按照类别进行分割). Zou 等人[20]直接将图像语义分割模型迁移到场景草图语义分割任务中, 结果显示性能较差. 这是因为草图与图像不同, 大部分区域为空白区域, 具有稀疏性. LDP[21]通过改进图像语义分割方法, 提升了模型对笔画的细节感知能力, 能够在一定程度上缓解草图稀疏性所带来的影响, 但是基于图像格式的草图语义分割方法丢失了矢量化笔画所特有的一些重要信息. S^3NN[22]以草图笔画为输入, 将视觉、时序和空间特征进行融合, 以提取草图的多样化特征, 最终实现了较好的场景草图语义分割效果. 但是该模型通过直接拼接长短时记忆网络所提取的时序特征和图卷积神经网络所提取的空间特征, 不能很好地平衡笔画的时空依赖.

图 Transformer 在图表示学习领域得到了广泛的应用, 它将图结构嵌入到 Transformer 架构中, 通过全局注意力来缓解稀疏消息传递机制的基本限制, 比如过度平滑[23]、过度压缩[24]和表现力界限[25]. 已有研究[26,27]证明了将草图笔画表示为图结构的实用性, 其在草图识别和分割任务上都取得了较好的效果, 因此本文使用图结构来对草图笔画进行建模. Transformer 作为一类强大的模型, 在自然语言处理和计算机视觉领域都发挥着重要作用, 当前已有研究[28,29]将 Transformer 引入到草图任务中并取得了较好的结果.

在草图绘制过程中, 笔画在时序上的上下文信息表达了绘画者的思维过程, 而空间上邻近笔画之间组成的图形传递了绘画者表达的意图, 这都包含了用于草图理解的重要信息. 本文充分考虑到草图笔画间所具有的时序关系和空间关系, 将场景草图中的笔画按照时序和空间邻近关系构建成图结构. 然后通过图 Transformer 多头自注意力机制和边增强机制传递笔画之间的信息, 捕获笔画的时空特征进行多分类, 实现场景草图的语义分割. 在场景草图数据集上进行实验验证, 本文所提出的图 Transformer 模型 GT-4S (graph Transformer for scene sketch semantic segmentation) 取得了最优的分割结果.

综上所述, 本文所提出的方法具有以下贡献.

(1) 首次提出了基于图 Transformer 模型来解决场景草图语义分割的问题, 该方法以矢量笔画为输入, 从几何、纹理、时空等多个角度对笔画信息进行编码, 有效地解决了草图稀疏性所带来的影响.

(2) 设计了一种边增强的方法, 通过融合包含时空上下文信息的笔画邻接矩阵与包含全局信息的注意力矩阵, 将图模型嵌入到 Transformer 架构中, 提高了 Transformer 对笔画的编码能力.

(3) 本文所提出的 GT-4S 方法在场景草图数据集 SFSD 上进行了大量的实验验证, 实验结果表明本文的方法取得了最先进的性能.

1 相关工作

1.1 草图语义分割

草图语义分割任务旨在为草图的每个分割部分预测正确的语义标签, 在草图理解和基于部件的草图分析中具

有重要作用[27]. 早期的草图语义分割通常使用手工提取特征[30−34]对草图笔画进行分类. 随着深度学习技术的快速发展, 各种神经网络被广泛应用于草图语义分割领域. 基于卷积神经网络 (convolutional neural network, CNN) 的模型将草图语义分割视为图像分割任务, 更加关注草图的局部边缘特征. MCPNet[35]通过在采样点集上学习和聚合多尺度深度表示来捕获草图的稀疏空间结构信息. SFSegNet[36]是一个针对草图的全卷积分割网络, 该网络包含改进的全卷积网络和仿射变换编码器, 使用重加权策略忽略背景像素来避免前、背景样本不均匀的问题, 并通过仿射变换规范笔画抖动. 基于循环神经网络 (recurrent neural network, RNN) 的模型通常提取笔画的时序特征, 将草图分割视为序列预测问题. SketchSegNet[16]首次将 RNN 应用于草图笔画分割问题, 借鉴了人类对于笔画绘制顺序的习惯和笔画之间的上下文信息, 利用解码器依次确定输入笔画的语义标签. 基于图神经网络 (graph neural network, GNN) 的模型可以有效地学习结构关系, 充分利用笔画之间的空间关系. SketchGNN[19]将输入草图视为二维点集, 通过构建多层次的图结构, 使用静态-动态分支图卷积网络架构来提取笔画内和笔画间的特征, 进而预测每个点的标签来实现对草图笔画的分割.

1.2 图模型

在分割任务中, 图已经得到了广泛的应用[37]. CRF[38]是基于图模型提出的一种有效图像分割方法. ZS3Net[39]采用图上下文编码来应对复杂场景中的语义类表示问题. 目前, 一些基于 GNN 的模型被用于语义分割任务. Graph-FCN[40]首次将图卷积神经网络 (graph convolution network, GCN) 应用于图像语义分割, 通过将图像网格数据扩展为图结构, 将语义分割问题转换为图节点分类问题. Chen 等人[41]提出一种全局推理方法, 首先将特征从坐标空间投影到交互空间, 然后利用图卷积网络推理节点之间的信息, 随后将此信息与节点特征融合, 再反投影到坐标空间. DGCNet[42]使用 GCN 来高效建模上下文信息进行语义分割. SiGCN[43]包含了一个支持诱导的图推理模块, 通过 GCN 来捕获不同语义级别的显著性查询目标. 当前一些工作使用稀疏连通图来表示草图, 然后使用 GNN 来处理各种草图上的任务. MGT[26]是首个将草图表示为图, 并将 GNN 应用于草图识别的工作, 使用 GNN 同时捕获全局和局部几何笔画结构和时间信息. ENDE-GNN[27]使用的 GNN 主干网络不仅关注草图笔画间和笔画内特征, 还关注笔画的绘制顺序, 以用于草图语义分割任务.

1.3 基于 Transformer 的架构

Transformer[44]作为一种结合多头注意力机制和前馈神经网络的深度学习模型, 最早应用于自然语言处理中的机器翻译任务. 随着基于 Transformer 的模型在自然语言处理领域取得巨大成功, 研究者们尝试将 Transformer 应用在计算机视觉领域. DETR[45]首先将 Transformer 应用在目标检测任务上. ViT[46]将输入图像分割成连续的图像块, 使用原始的 Transformer 编码器获取图像分类的视觉表示. SETR[47]首次在语义分割任务中使用 Transformer, 重新定义语义分割任务, 将其视为序列到序列的预测任务, 取得了突出效果. SegFormer[48]将无位置编码的分层 Transformer 编码器、轻量级多层感知机解码器结合起来, 避免了复杂设计, 实现高效的语义分割框架. Swin Transformer[49]在 ViT 的基础上引入滑动窗口机制和层次结构, 在语义分割任务中取得了比大多数 CNN 模型更好的结果, 且模型的参数量远小于 ViT. 已有研究将 Transformer 引入到基于草图的任务中, Sketchformer[28]将 Transformer 用于学习手绘草图的深度表示, 能够完成草图分类、基于草图的图像检索以及草图的重建和插值等多个任务. TVT[29]在基于草图的零样本图像检索任务中, 利用 ViT 对全局结构信息进行建模, 以更好地实现草图和图像之间的对齐. Tripathi 等人[50]提出一种草图引导的 ViT 编码器来学习查询条件下的图像特征, 从而实现与所查询草图更强的对齐.

1.4 图 Transformer 模型

图 Transformer 将图结构嵌入到 Transformer 架构中, 在避免严格结构归纳偏差的同时, 克服了局部邻域聚合的局限性[51]. GraphTrans[52]在标准的 GNN 模块之上添加一个 Transformer 模块, 其中 GNN 模块作为专门框架来学习节点近邻结构的局部表示, 而 Transformer 模块则以位置无关的方式计算所有成对节点的相互作用, 增强了模型的全局推理能力. 通过精心设计节点、边、图级的自监督任务, GROVER[53]可以从大量无标签数据中学习到丰富的隐含信息. 为了编码这些复杂信息, GROVER 将消息传递网络 (message passing network, MPS) 集成到

Transformer 架构中, 提供了一类更具表达力的编码器. SAN[54]提出了一种可学习的位置编码 (learned positional encoding, LPE), 利用拉普拉斯矩阵来学习给定图中每个节点的位置, 然后把 LPE 添加到图的节点特征中, 并传递给全连通 Transformer. Graphormer[55]使用度中心性作为神经网络的附加信号, 中心性编码根据每个节点的入度和出度为其分配两个实值嵌入向量. 当前已有工作将图 Transformer 应用在草图任务上, MGT[26]首先将草图建模为稀疏连通图, 然后将其输入到 Transformer 中以提供先验知识, 用于捕获草图的几何结构和时间信息. 本文将草图笔画构建成相应的图结构, 使用边增强的图 Transformer 网络作为笔画的特征提取器, 充分挖掘笔画之间的时序上下文信息和空间关联, 提高分割网络的性能.

2 面向场景草图的语义分割方法 GT-4S

针对场景草图所存在的抽象性、稀疏性、多样性、语义复杂性等特点, 本文提出了一种新颖的图 Transformer 语义分割网络 (GT-4S). 该网络能够有效地融合笔画几何与纹理多尺度特征, 并通过笔画间的时空上下文关联来增强对时空特征的学习. 总体框架图如图 1 所示, 主要由 3 个模块组成: (1) 图结构构建与多尺度笔画特征提取模块; (2) 图 Transformer 特征编码器; (3) 语义分割模块. 具体而言, 给定一张场景草图, 首先基于草图中的笔画信息构建图结构, 草图中的每条笔画作为图的节点, 笔画之间的时空邻近关系组成了图的边. 节点特征由笔画的纹理特征和几何特征组成, 边特征通过手工提取的方式获得. 然后图 Transformer 特征编码器通过多头自注意力机制和边特征增强机制传递笔画之间的信息来提高笔画的特征表征. 最后将节点特征输入到全连接分类层, 通过最小化交叉熵损失对笔画进行多分类, 完成对场景草图的语义分割. 本节针对 GT-4S 的实现细节进行详细介绍.

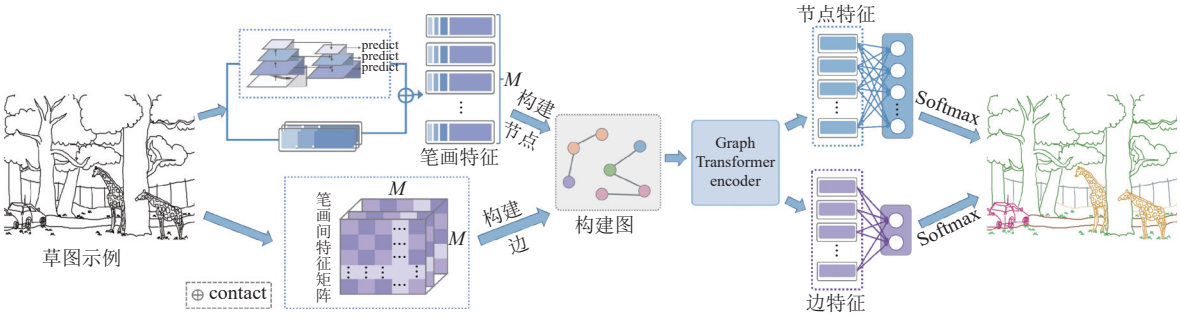

图 1 基于图 Transformer 的场景草图语义分割框架图

2.1 构建笔画图结构

本文假定训练集 D 中包含了 N 张标注后的场景草图: $D=\{D^n\}_{n=1}^N$. 场景草图 D^n 由 M 条笔画组成: $D^n=\{(S_m,c_m)|c_m\in 1,\ldots,C\}_{m=1}^M$, 其中 S_m 为草图 D^n 中的第 m 条笔画, c_m 为笔画 S_m 所属的类别标签, $\{1,\ldots,C\}$ 为数据集中所有类别的集合. 每条笔画由一系列有序的轨迹点 (x,y) 组成.

给定一张场景草图 D^n, 构建草图笔画图结构 $G_n=(V_n,E_n)$, 其中 $V_n=\{v_i|i=1,\ldots,M\}$ 表示包含 M 条笔画的节点集合, $E_n=\{e_{ij}|i=1,\ldots,M\ \&\ j=1,\ldots,M\}$ 表示由邻近矩阵定义的笔画边集合. 节点 v_i 表示笔画 S_i 的特征向量. 如果笔画 S_i 和 S_j 在时序或者空间上是足够邻近的, 边 e_{ij} 表示它们之间邻接边的特征向量.

节点特征提取: 现有的一些草图表示方法直接将笔画采样点的相对坐标[17,56]或绝对坐标[19]作为输入, 另外一些结构化草图的相关方法也使用手工设计的特征进行表示. 由于场景草图相较于单物体草图和结构化草图更加复杂多样, 传统的笔画表示方法难以充分表征笔画特征. 因此本文使用预训练的特征金字塔网络[57]提取笔画对齐位置的纹理特征, 然后使用一个线性层将特征映射到目标维度, 提取的笔画纹理特征表示为 f^{cnn}. 本文也编码了笔画几何特征包含: 长度 f^l、绘制持续时间 f^t、归一化的笔画包围框位置坐标 f^b. 将纹理特征和几何特征拼接起来组成笔画初始特征, 作为图节点的输入特征 $f_i=\text{concat}(f_i^{\text{cnn}},f_i^l,f_i^t,f_i^b)$.

边特征提取: 参考文献 [58], 本文分别构建了图结构的时序边和空间边. 在绘制过程中, 时序上靠近的笔画有

更大的概率属于相同类别. 因此本文分别选择了每条笔画 S_i 在时序上向前和向后的 L_t 条最近邻的笔画组成时序边. 笔画与空间上邻近的笔画组合成的图形传递了草图所要表达的信息, 对于每条笔画 S_i, 本文选择了笔画包围框的中心位置在欧氏距离上最近的 L_s 条笔画组成空间边.

2.2 图 Transformer 特征编码器

具体而言, Transformer 结构是 Vaswani 等人[44]提出的多头自注意力模型, 基于 Transformer 的模型以及变种网络[59]在自然语言处理上取得了良好的性能. 近年来, 许多工作 (如 ViT[46], SegFormer[48], SETR[47], Segmenter[60]) 也将 Transformer 引入到图像语义分割中. 与 CNN 的方法不同, Transformer 方法将图像分割成块并映射到一个线性嵌入序列, 通过捕捉全局上下文信息, 在图像语义分割上带来了很大的提升. 草图兼具有文本和图像的特性. 草图在绘制过程中与文本类似, 笔画按照时序顺序进行输入. 同时绘制完成的笔画所组成图形传递的语义信息与图像像素在空间上的关联类似. 如图 2 所示, 本文采用了边增强的图 Transformer 网络作为笔画的特征提取器, 能够充分挖掘笔画之间的时序上下文信息和空间关联信息, 提高分割网络的性能. 在构建笔画图结构之后, 本文将笔画 S_i 所提取的初始特征输入到一个全连接层来获取隐藏特征 $h_i^{(0)}$, 并将 $h_i^{(0)}$ 作为图 Transformer 的初始输入. 由于注意力机制对位置不敏感, Transformer 通常使用绝对位置编码或者相对位置编码来捕获输入序列数据中 token 的位置信息. 草图笔画同样属于序列数据, 所以本文使用了绝对位置正余弦曲线来编码笔画的位置信息, 将其与节点特征相加作为图 Transformer 模块的输入.

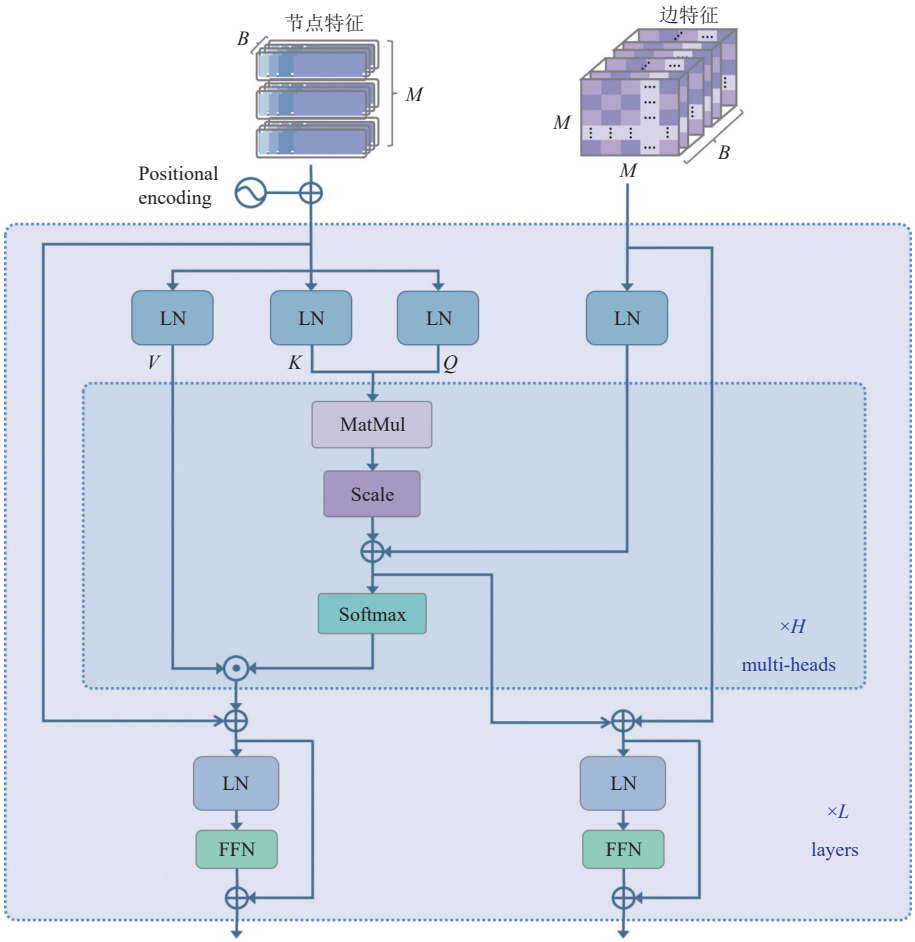

图 2 图 Transformer 网络

$$h_i^{(0)} = f_i + pe(v_i) \tag{1}$$

其中, $pe(\cdot)$ 表示节点 v_i 在笔画序列中的位置编码.

2.2.1 时空特征编码

注意力矩阵与邻接矩阵均可表明对图中节点特征的增强方式. 与输入的图结构不同, 注意力机制生成的是动态图[61]. 因而, Transformer 结构只能通过自注意机制使节点特征互相学习. 为了将包含时序和空间信息的边特征编码到注意力层, 本文采用了公式 (2) 中的方式使边特征参与到节点特征增强的过程中. $h_i^{(0)}$ 和 $e_{ij}^{(0)}$ 是图 Transformer 结构初始层节点和边的特征输入, 第 l 层的迭代计算过程如下:

$$s_{ij}^{(l)} = \frac{(h_i^{(l)}\mathbf{W}_Q^{(l)})(h_j^{(l)}\mathbf{W}_K^{(l)})^{\mathrm{T}}}{\sqrt{d_o}} + \sigma(e_{ij}^{(l)}) \tag{2}$$

$$a_{ij}^{(l)} = \frac{\exp(s_{ij}^{(l)})}{\sum_k^N \exp(s_{ik}^{(l)})} \tag{3}$$

$$o_i^{(l)} = \sum_{j=1}^N a_{ij}^{(l)}(h_j^{(l)}\mathbf{W}_V^{(l)}) \tag{4}$$

其中, $\mathbf{W}_Q^{(l)}, \mathbf{W}_K^{(l)}, \mathbf{W}_V^{(l)} \in \mathbf{R}^{d_l \times d_o}$ 是 3 个可学习的参数矩阵, $\sigma(\cdot)$ 为 Sigmoid 激活函数, $o_i^{(l)} \in \mathbf{R}^{d_o}$ 是自注意力模块的输出, d_o 是输出的维度.

本文通过连接多个自注意力模块的输出, 将单头注意力转化为多头注意力机制 (MHA), 并且堆叠 L 层的多头注意力模块. 本文也采用了层归一化 (LN)、前馈神经网络 (FFN) 和残差来提升训练效果.

$$h'^{(l)} = \mathrm{MHA}(\mathrm{LN}(h^{(l-1)}), \mathrm{LN}(e^{(l-1)})) + h^{(l-1)} \tag{5}$$

$$h^{(l)} = \mathrm{FFN}(\mathrm{LN}(h'^{(l)})) + h'^{(l)} \tag{6}$$

2.2.2 边学习

直观上笔画在时序或空间上邻近, 则属于相同类别的可能性更大. 但是对于时序上邻近绘制物体所属的笔画或者空间上邻近物体接触位置处的笔画都难以进行有效的分割, 因此提高边分类的准确率有助于提高物体时空边缘位置笔画分割的效果. 图结构中的边包含了笔画之间的时序和空间邻接关系, 边特征也可以跟随节点特征学习而增强. 为了使边特征能够更好地表征笔画之间的上下文信息, 本文将笔画之间的注意力权重融合到边特征中, 通过层归一化和前馈神经网络产生当前层的边特征, 并随着图 Transformer 的每一层进行参数更新.

$$e'^{(l)}_{ij} = e_{ij}^{(l-1)} + s_{ij}^{(l)}\mathbf{W}_E^{(l)} \tag{7}$$

$$e_{ij}^{(l)} = \mathrm{FFN}(\mathrm{LN}(e'^{(l)}_{ij})) + e'^{(l)}_{ij} \tag{8}$$

其中, $\mathbf{W}_E^{(l)}$ 为可学习的参数矩阵.

2.3 目标函数

本文使用交叉熵损失函数对节点类别进行训练:

$$L_{\mathrm{Node}} = -\frac{1}{M}\sum_{m=1}^{M} w_c \cdot y_m \cdot \log(\hat{y}_m) \tag{9}$$

其中, y_m 为标注的真实值, $\hat{y}_m$ 为分割模型的预测值. 针对数据集中类别分布高度不均衡的问题, 本文对不同类别 c 赋予了不同的权值 w_c (w_c 为数据集中所有类别出现频率的中位数与类别 c 出现频率的比值).

本文将笔画之间的边分为两个类别: 若一条边连接的两条笔画属于同一实例则表示为 1, 否则表示为 0.

$$L_{\mathrm{Edge}} = -\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{M} Y_{ij} \cdot \log(\hat{Y}_{ij}) \tag{10}$$

其中, $Y_{ij} \in \{0,1\}$ 是边 e_{ij} 标注的真实值, $\hat{Y}_{ij}$ 为边 e_{ij} 的预测值.

通过结合节点分类损失和边分类损失, 最终的笔画语义分割损失函数定义如下:

$$L = \lambda L_{\text{Node}} + (1-\lambda) L_{\text{Edge}} \tag{11}$$

其中, λ 为加权参数.

3 实 验

3.1 实验数据

本文采用 SFSD[22]场景草图数据集来进行实验评测. SFSD 数据集中一共包含了 40 个物体类别, 其中 27 个前景类别、12 个背景类别和 1 个其他类别. 该数据集从 COCO2017 中挑选部分图像作为参考图像, 通过用户自由手绘的方式绘制草图, 保留了用户绘制过程中的笔画时序信息, 并对草图实例进行了细粒度标注. 该数据集包括了 12 100 张自由手绘的场景草图. 本文对该数据集进行随机划分, 将其中 9 100 张草图作为训练集, 剩余的 3 000 张作为测试集.

3.2 评价指标

本文使用了草图语义分割中常用的 5 种评价指标[19−21,58]来衡量模型分割场景草图的质量. 其中总体笔画分类准确率 SCA_o 和每类笔画分类准确率 SCA_c 是笔画级的评价指标, 可以直观地反映笔画分割的质量. 由于部分对比方法是对图像格式的场景草图进行语义分割, 为了客观与公平的对比, 本文也采用了像素级的评价指标, 包括: 总体像素分类准确率 PA_o、每个类别像素分类准确率 PA_c、平均交并比 $MIoU$. 具体定义如下:

(1) 总体笔画分类准确率 SCA_o (overall stroke classification accuracy):

$$SCA_o = \frac{\sum_{n=1}^{N}\sum_{m=1}^{M_n}\phi(\hat{y}_m^n = y_m^n)}{\sum_{n}^{N} M_n} \tag{12}$$

(2) 每个类别笔画分类准确率 SCA_c (per-class stroke classification accuracy):

$$SCA_c = \frac{\sum_{n=1}^{N}\sum_{m=1}^{M_n}\phi(y_m^n = c)\phi(\hat{y}_m^n = y_m^n)}{\sum_{n}^{N}\sum_{m=1}^{M_n}\phi(y_m^n = c)} \tag{13}$$

其中, N 为草图数, M_n 为第 n 张草图包含的笔画数, $\hat{y}_m^n$ 和 y_m^n 分别为笔画 S_m 的预测值和真实值. $\phi(\cdot)$ 为指示函数, 条件为真时取值为 1, 否则取值为 0. $c \in \{1,\cdots,C\}$ 是笔画 S_m 所属的类别. 参考文献 [19,62], 对于像素级的评价指标, 当组成笔画 75% 以上的像素都分类正确, 则此笔画分类正确.

(3) 总体像素分类准确率 PA_o (overall pixel accuracy):

$$PA_o = \frac{\sum_{n}^{N}\sum_{k}^{K_n}\phi(\hat{p}_k^n = p_k^n)}{\sum_{n}^{N} K_n} \tag{14}$$

(4) 每个类别像素分类准确率 PA_c (per-class pixel accuracy):

$$PA_c = \frac{\sum_{n}^{N}\sum_{k}^{K_n}\phi(p_k^n = c)\phi(\hat{p}_k^n = p_k^n)}{\sum_{n}^{N}\sum_{k}^{K_n}\phi(p_k^n = c)} \tag{15}$$

(5) 平均交并比 $MIoU$ (mean intersection over union):

$$MIoU = \frac{1}{C}\sum_{c=1}^{C}\frac{\sum_{n=1}^{N}\sum_{k}^{K_n}\phi(p_k^n = c)\phi(\hat{p}_k^n = c)}{\sum_{n=1}^{N}\sum_{k}^{K_n}[\phi(p_k^n = c) + \phi(\hat{p}_k^n = c) - \phi(p_k^n = c)\phi(\hat{p}_k^n = c)]} \tag{16}$$

其中, K_n 为第 n 张草图掩膜非空白区域的像素总数, $\hat{p}_k^n$ 和 p_k^n 分别为草图掩膜中第 k 个像素的预测值和真实值.

3.3 实现细节

本文使用特征金字塔作为编码器来提取 256 维的笔画纹理特征, 同时提取了 1 维的笔画长度, 1 维的笔画绘制提取时间, 4 维的笔画包围框左上角和右下角的位置坐标, 共 262 维作为笔画的初始特征. 在图结构中, 本文设

置最近邻时序边数 L_t 为 2, 最近邻空间边数 L_s 为 5. 通过对 SFSD 场景草图数据集中所有草图笔画数进行统计, 图结构的最大节点数设置为 700 个. 图 Transformer 模型的层数设置为 4, 每层包含了 8 个自注意力头, 节点输入特征为 262 维, 边输入特征为 19 维.

本文的模型代码使用开源 PyTorch 1.8.2 深度学习框架实现, 在一块 NVIDIA GeForce RTX3090 GPU 上进行模型的训练与测试, CUDA 版本为 11.1. 模型在训练过程中, 采用 Adam 优化器进行参数优化, 其中初始学习率为 0.000 1, 动量和衰减率分别为 0.5 和 0.000 5. 实验中, 每个批处理输入 16 张场景草图, 遍历训练集 100 轮, 学习率每 30 轮衰减一次, 衰减率为 0.5. 实验中, 加权参数 $\lambda = 0.6$.

3.4 与现有方法对比

为了验证本文所提出的 GT-4S 模型在场景草图语义分割方面的优势, 本文与 6 个现有的语义分割方法进行了对比, 分别为: U-Net[63]、FPN[57]、DeepLabv3+[64]、SketchGNN[19]、LDP[21]、S^{3}NN[22]. 本文按照模型输入的草图格式 (图像或矢量草图) 对这些方法进行分类. 其中, U-Net、FPN 和 DeepLabv3+是 3 个经典的自然图像语义分割方法, 而 LDP 针对草图特性在 DeepLabv2 基础上进行了改进的草图图像语义分割方法. 这 4 个方法都是根据草图语义标签生成图像掩膜, 并将矢量草图转化为图像格式作为模型的输入, 从图像像素的角度对场景草图进行分割. 与本文方法相同, S^{3}NN 也是以草图笔画为输入直接进行语义分割. SketchGNN 直接使用草图笔画采样点的绝对坐标作为输入, 因此本文对数据集中每张草图的采样点进行了重采样, 使每张草图 (包含 2 048 个采样点) 的采样点数保持一致.

从表 1 中可以看到, 本文所提基于图 Transformer 的模型 GT-4S 在性能上优于其他现有的语义分割算法. 由于草图具有稀疏性, 大部分区域都是空白的, 自然图像语义分割方法所引入的特征金字塔和上下文信息对草图分割没有显著的提升, U-Net、FPN 和 DeepLabv3+ (都使用 ResNet50 作为基础网络) 在所有评价指标上都远低于本文所提出的方法, 本文方法在 SCA_o、PA_o 和 $MIoU$ 评价指标上比表现最好的 DeepLabv3+分别超出了 6.07%、6.44%、6.58%. 针对手绘草图和自然图像存在的本质差异, LDP[21]充分利用笔画的低层次特征来提高局部细节识别能力, 相较于自然图像语义分割方法有一定的性能提升, 但还是不能完全解决草图稀疏且抽象的问题. 本文方法在 3 个评价指标比 LDP 分别提高 2.9%、3.52% 和 1.67%. SketchGNN 将草图采样点构建图结构作为输入表示, 结合动态和静态图卷积模块学习草图笔画特征, 在单物体草图进行部位分割任务上取得良好的性能, 而场景草图相对于单物体草图具有笔画更多、类别更多、语义更复杂等特点, 该方法在场景草图数据集上表现较差. S^{3}NN 同样是基于笔画的场景草图语义分割方法, 作为对比方法中表现最好的方法, 在 3 个评价指标上比本文方法分别降低 1.34%、1.62% 和 1.67%. 实验结果证明了本文方法使用图 Transformer 能有效地提升笔画特征的表达能力, 并获得最佳的实验效果.

表 1 SFSD 数据集场景草图语义分割结果对比 (%)

方法	SCA_o	PA_o	$MIoU$
U-Net	69.65	65.18	31.31
FPN	74.03	70.83	37.41
DeepLabv3+	76.73	73.83	41.50
SketchGNN	57.83	58.84	25.45
LDP	79.90	76.75	46.36
S^{3}NN	81.56	78.65	46.41
GT-4S	**82.80**	**80.27**	**48.08**

在表 2 中, 对比了本文所提方法 GT-4S、基于像素表现最好的语义分割方法 LDP、基于笔画表现最好的语义分割方法 S^{3}NN 在 SFSD 数据集中全部类别的表现. 本文所提出方法在每个类别笔画准确率 SCA_c 和像素准确率 PA_c 的均值上都优于对比方法, 比 LDP 分别提高了 5.19% 和 6.74%, 比 S^{3}NN 分别提高了 4.21% 和 5.01%. 本文所提出的方法在绝大部分类别上也都优于对比方法. 表 2 的结果表明本文方法具有良好的泛化性能.

表 2　在 SFSD 数据集上各类别的准确率 (%)

类别	LDP		S^3NN		GT-4S	
	SCA_c	PA_c	SCA_c	PA_c	SCA_c	PA_c
airplane	**94.62**	**93.94**	88.84	90.85	92.79	92.72
backpack	1.60	1.07	1.22	1.05	**1.71**	**1.12**
baseball bat	18.71	20.82	11.80	17.98	**18.81**	**21.66**
baseball glove	10.72	10.07	**12.96**	**11.56**	10.52	10.46
bear	53.75	47.31	49.32	47.44	**59.69**	**61.27**
bicycle	59.12	60.05	62.47	60.58	**64.28**	**63.86**
bird	47.98	48.28	58.16	55.39	**61.38**	**61.73**
boundary	26.23	31.14	54.27	56.42	**58.83**	**59.99**
bus	**83.85**	**85.32**	74.32	77.51	78.56	80.89
car	41.07	40.21	48.61	46.76	**50.18**	**50.95**
cloud	94.91	95.43	94.03	94.34	**95.16**	**96.12**
cow	**63.45**	**59.83**	53.86	54.01	56.18	54.22
dog	31.17	**34.37**	23.17	22.78	**37.33**	30.84
elephant	86.23	86.09	87.33	86.19	**87.81**	**87.14**
fence	45.14	45.40	61.18	62.57	**66.30**	**72.15**
frisbee	**20.00**	**23.09**	11.94	15.45	14.84	22.80
giraffe	94.79	93.96	95.44	94.73	**97.24**	**96.67**
grass	88.43	90.73	94.15	94.29	**94.78**	**94.71**
horse	59.02	60.28	47.23	47.11	**61.81**	**63.23**
house	60.17	59.36	63.47	60.21	**65.78**	**63.01**
kite	28.34	21.04	11.74	10.72	**31.98**	**36.52**
motorcycle	**82.11**	**81.85**	76.09	77.75	76.32	76.89
mountain	**46.32**	49.55	44.08	**56.82**	42.39	53.42
person	93.95	93.49	96.50	96.38	**97.13**	**97.32**
playground	36.17	38.42	**52.44**	**56.38**	50.37	54.91
river	66.10	65.45	73.16	71.71	**76.52**	**76.71**
road	45.08	46.74	53.76	57.30	**55.76**	**59.60**
sheep	57.22	52.72	71.47	66.51	**74.93**	**74.13**
skateboard	57.50	55.28	58.13	59.02	**63.06**	**63.16**
skis	33.20	36.69	39.05	46.13	**39.52**	**47.60**
snowboard	**21.47**	**21.26**	14.72	15.11	15.73	17.69
snowfield	78.91	69.32	84.10	76.81	**87.98**	**83.99**
sports ball	34.15	37.54	21.25	21.65	**34.49**	**40.27**
stone	47.07	45.40	61.58	55.38	**65.53**	**61.57**
surfboard	12.38	10.86	3.47	4.48	**17.82**	**20.48**
tennis racket	48.23	51.84	56.25	66.37	**60.10**	**66.44**
tree	87.19	89.03	**89.16**	**90.22**	89.03	90.04
truck	55.94	52.67	48.24	46.50	**59.76**	**60.80**
zebra	96.79	97.20	98.18	98.25	**98.67**	**98.94**
mean	54.07	53.93	55.05	55.66	**59.26**	**60.67**

3.5　定性分析

图 3 比较了本文提出方法与对比方法的语义分割可视化结果 (图 3 下方为图中所包含物体类别对应的颜色). 图中结果反映了本文方法可对包含多个前景物体、背景物体且语义复杂的场景草图进行有效的语义分割. 接着, 从局部细节和整体全局进行比较. 在局部细节处, 以图像为输入的方法在不同类别物体邻近位置的笔画都存在分类错误的情况. 例如第 2–5 行第 2 列中“giraffe”和“tree”邻近或遮挡部位的笔画都存在分割错误, 第 3 列“truck”后轮下方的“grass”均分割错误. 虽然 S^3NN 同样是对笔画特征进行编码, 但是不能很好地平衡笔画的时空关系. 例如

在第 6 行第 4 列中, “truck”中的部分笔画在时序上与“car”的笔画邻近, 受到“car”时序信息的影响, 多条笔画分割错误. 在草图中占比较大的物体, 需要模型对草图整体感知才可以正确分割. 第 3 列中只有 GT-4S 对“truck”进行了正确的分割. 由于图像分割方法对整体感知能力较弱, “truck”所包含的笔画被错误分割为多个类别. 而 S^3NN 将“truck”整体错误地分割为类别“car”. 从整体的分割可视化结果可以看出, GT-4S 对场景局部、全局都感知良好, 在场景草图语义分割上具有优势.

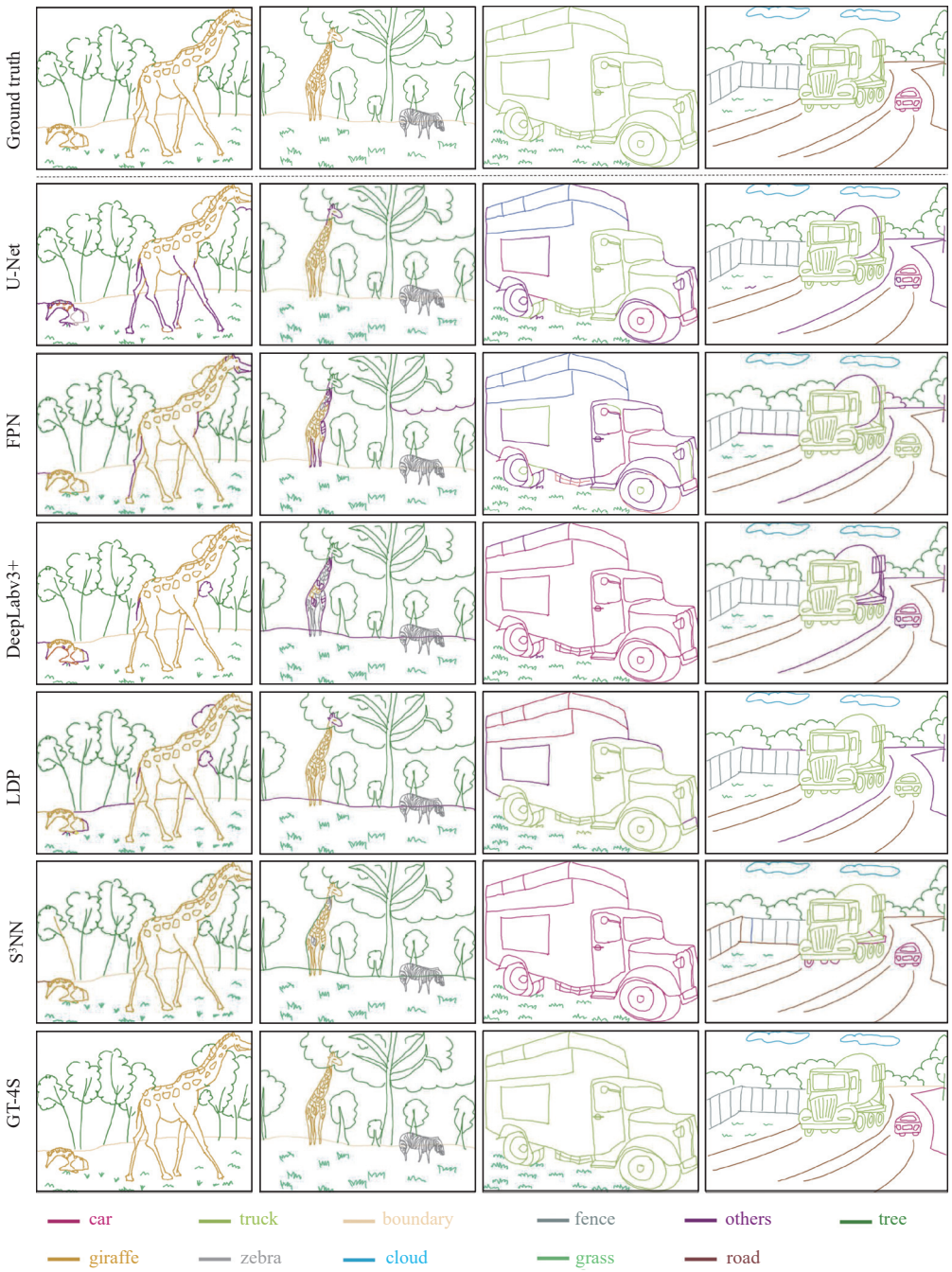

图 3　本文方法与其他方法在 SFSD 数据集上分割结果可视化图

3.6 消融实验

在表 3 中, 本文做了消融实验来验证所提出 GT-4S 模型的优势 (包括几何特征、纹理特征、时序边、空间边、类别权值, 模型的编号为模型 1–5). 为了验证不同笔画初始特征的有效性, 在模型 1 中本文将几何特征和笔画采样点的相对坐标拼接起来, 在模型 2 中将几何特征与纹理特征相拼接. 模型 1 比模型 2 在 SCA_o 上下降了 28.27%, 这表明通过 CNN 提取的纹理特征能够更好地对笔画进行表征. 从模型 3 和模型 4 可知, 在 Transformer 自注意力的基础上, 融合笔画图结构的时序边和空间边特征能够进一步提升模型的性能. 从模型 5 中可知, 在损失函数中对出现频率低的类别赋予更大的类别权值有助于模型性能的提升.

表 3 每个组成模块的消融分析

编号	几何特征	纹理特征	时序边	空间边	类别权值	SCA_o (%)
1	√	—	—	—	—	52.19
2	√	√	—	—	—	80.46
3	√	√	√	—	—	81.25
4	√	√	√	√	—	82.18
5	√	√	√	√	√	82.80

3.7 讨 论

从表 2 可知, 本文所提方法 GT-4S 以及对比方法 LDP、S^3NN 在 SFSD 场景草图数集中部分类别 (如 backpack、baseball glove、surfboard、snowboard、frisbee、baseball bat 等) 的准确率都低于 25%. 图 4 中展示了分割错误的结果 (图 4 下方为图中所包含物体类别对应的颜色), 分析其错误的原因包括: (1) 类别分布不均, 这些类别出现的频率很低; (2) 邻近的小物体与大物体笔画重合度高, 且特征不明显, 容易被错误分为大物体所属的类别; (3) 由于草图的抽象性和稀疏性, 部分类别物体的几何形状相似度很高, 不像图像可以从像素颜色获取信息, 因此易于被互相误分. 从图 4 中第 2 和 4 列可以看出, 类别“backpack”和“baseball glove”都是“person”上的配件, 与“person”笔画重合度很高, 在时序和空间上受到“person”的影响而被错误分类. 在第 2 列中, “snowboard”和“surfboard”在形状上基本类似, 只能从场景语义约束上进行区分.

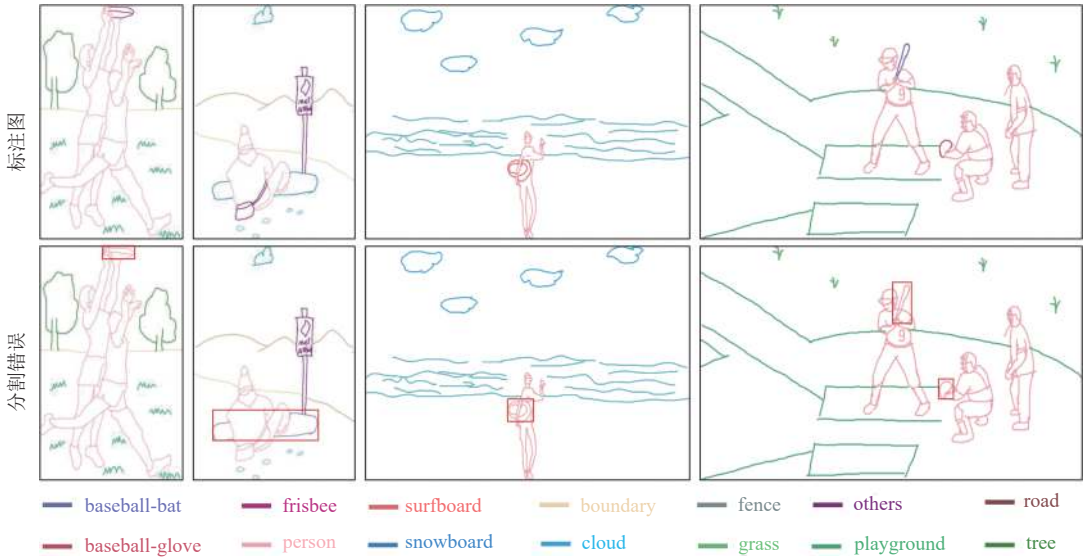

图 4 分割错误结果展示

4 总 结

本文提出了一种创新的图 Transformer 模型来解决基于笔画的场景草图语义分割问题. 该方法首先提取了笔

画的纹理特征和几何特征, 并构建笔画之间的时序边和空间边, 组成图结构. 然后通过边增强的图 Transformer 模型对笔画进行时空表征学习. 最后对笔画进行多分类完成场景草图的语义分割. 通过在 SFSD 自由手绘场景草图数据集上进行实验对比, 证明了本文所提出的 GT-4S 模型能够对复杂、多样的场景草图笔画进行有效的时空编码, 提高场景草图语义分割的准确率.

本文所提出的方法虽然在大部分类别能够取得较好的分割效果, 但是对于类别分布不均衡、物体笔画重合度高、不同类别相似性高以及小物体分割难度大等问题不能被有效解决. 下一步工作拟通过结合单物体部位分割、场景类别条件约束以及交互式引导等方式来进一步提高场景草图语义分割的性能. 为了进一步挖掘场景草图更细粒度的语义信息, 未来还会围绕场景草图的实例分割展开研究.

References:

[1] Wang F, Lin SJ, Wu HF, Li HH, Wang RM, Luo XN, He XJ. SPFusionNet: Sketch segmentation using multi-modal data fusion. In: Proc. of the 2019 IEEE Int'l Conf. on Multimedia and Expo. Shanghai: IEEE, 2019. 1654–1659. [doi: 10.1109/ICME.2019.00285]

[2] Wang SX, Wang SX, Wang GF, Gao MT. Segmentation of online freehand stroke using geometrical feature. Journal of Computer-aided Design & Computer Graphics, 2015, 27(9): 1686–1693 (in Chinese with English abstract). [doi: 10.3969/j.issn.1003-9775.2015.09.012]

[3] Deng ZG, Lyu J, Liu X, Hou YK, Wang S. StyleGAN-based sketch generation method for product design renderings. Packaging Engineering, 2023, 44(6): 188–195 (in Chinese with English abstract). [doi: 10.19554/j.cnki.1001-3563.2023.06.020]

[4] Huang F, Canny JF. Sketchforme: Composing sketched scenes from text descriptions for interactive applications. In: Proc. of the 32nd Annual ACM Symp. on User Interface Software and Technology. New Orleans: ACM, 2019. 209–220. [doi: 10.1145/3332165.3347878]

[5] Zhang JH, Chen YL, Li L, Fu HB, Tai CL. Context-based sketch classification. In: Proc. of the 2018 Joint Symp. on Computational Aesthetics and Sketch-based Interfaces and Modeling and Non-photorealistic Animation and Rendering. Victoria: ACM, 2018. 3. [doi: 10.1145/3229147.3229154]

[6] Yang JK, Wang GZ, Fan T. Recognition and matching of hand-drawn sketch based on neural network. Intelligent Computer and Applications, 2021, 11(6): 148–152 (in Chinese with English abstract). [doi: 10.3969/j.issn.2095-2163.2021.06.028]

[7] Song JF, Yu Q, Song YZ, Xiang T, Hospedales TM. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 5552–5561. [doi: 10.1109/ICCV.2017.592]

[8] Chen J, Bai C, Ma Q, Hao PY, Chen SY. Adversarial training triplet network for fine-grained sketch based image retrieval. Ruan Jian Xue Bao/Journal of Software, 2020, 31(7): 1931–1942 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5934.htm [doi: 10.13328/j.cnki.jos.005934]

[9] Zhou BL, Lapedriza A, Khosla A, Oliva A, Torralba A. Places: A 10 million image database for scene recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2018, 40(6): 1452–1464. [doi: 10.1109/TPAMI.2017.2723009]

[10] Gao CY, Liu Q, Xu Q, Wang LM, Liu JZ, Zou CQ. SketchyCOCO: Image generation from freehand scene sketches. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 5173–5182. [doi: 10.1109/CVPR42600.2020.00522]

[11] Liu F, Zou CQ, Deng XM, Zuo R, Lai YK, Ma CX, Liu YJ, Wang HA. SceneSketcher: Fine-grained image retrieval with scene sketches. In: Proc. of the 16th European Conf. Glasgow: Springer, 2020. 718–734. [doi: 10.1007/978-3-030-58529-7_42]

[12] Chowdhury PN, Bhunia AK, Sain A, Koley S, Xiang T, Song YZ. SceneTrilogy: On human scene-sketch and its complementarity with photo and text. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 10972–10983. [doi: 10.1109/CVPR52729.2023.01056]

[13] Ye YX, Lu YJ, Jiang H. Human's scene sketch understanding. In: Proc. of the 2016 ACM on Int'l Conf. on Multimedia Retrieval. New York: ACM, 2016. 355–358. [doi: 10.1145/2911996.2912067]

[14] Wang JY, Jeon S, Yu SX, Zhang X, Arora H, Lou Y. Unsupervised scene sketch to photo synthesis. In: Proc. of the 2023 European Conf. on Computer Vision. Tel Aviv: Springer, 2023. 273–289. [doi: 10.1007/978-3-031-25063-7_17]

[15] Wang JX, Zhu ZL, Deng XM, Ma CX, Wang HA. Survey on sketch segmentation algorithm based on deep learning. Ruan Jian Xue Bao/Journal of Software, 2022, 33(7): 2729–2752 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6299.htm [doi: 10.13328/j.cnki.jos.006299]

[16] Wu XY, Qi YG, Liu J, Yang J. SketchSegNet: A RNN model for labeling sketch strokes. In: Proc. of the 28th IEEE Int'l Workshop on Machine Learning for Signal Processing. Aalborg: IEEE, 2018. 1–6. [doi: 10.1109/MLSP.2018.8516988]

[17] Qi YG, Tan ZH. SketchSegNet+: An end-to-end learning of RNN for multi-class sketch semantic segmentation. IEEE Access, 2019, 7: 102717–102726. [doi: 10.1109/ACCESS.2019.2929804]

[18] Kaiyrbekov K, Sezgin M. Deep stroke-based sketched symbol reconstruction and segmentation. IEEE Computer Graphics and Applications, 2020, 40(1): 112–126. [doi: 10.1109/MCG.2019.2943333]

[19] Yang LM, Zhuang JJ, Fu HB, Wei XZ, Zhou K, Zheng YY. SketchGNN: Semantic sketch segmentation with graph neural networks. ACM Trans. on Graphics, 2021, 40(3): 28. [doi: 10.1145/3450284]

[20] Zou CQ, Yu Q, Du RF, Mo HR, Song YZ, Xiang T, Gao CY, Chen BQ, Zhang H. SketchyScene: Richly-annotated scene sketches. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 438–454. [doi: 10.1007/978-3-030-01267-0_26]

[21] Ge C, Sun HF, Song YZ, Ma ZY, Liao JX. Exploring local detail perception for scene sketch semantic segmentation. IEEE Trans. on Image Processing, 2022, 31: 1447–1461. [doi: 10.1109/TIP.2022.3142511]

[22] Zhang ZM, Deng XM, Li JY, Lai YK, Ma CX, Liu YJ, Wang HA. Stroke-based semantic segmentation for scene-level free-hand sketches. The Visual Computer, 2023, 38(12): 6309–6321. [doi: 10.1007/s00371-022-02731-8]

[23] Oono K, Suzuki T. Graph neural networks exponentially lose expressive power for node classification. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2019.

[24] Alon U, Yahav E. On the bottleneck of graph neural networks and its practical implications. In: Proc. of the 9th Int'l Conf. on Learning Representations. OpenReview.net, 2021.

[25] Morris C, Ritzert M, Fey M, Hamilton WL, Lenssen JE, Rattan G, Grohe M. Weisfeiler and Leman go neural: Higher-order graph neural networks. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. Honolulu: AAAI Press, 2019. 4602–4609. [doi: 10.1609/aaai.v33i01.33014602]

[26] Xu P, Joshi CK, Bresson X. Multigraph Transformer for free-hand sketch recognition. IEEE Trans. on Neural Networks and Learning Systems, 2022, 33(10): 5150–5161. [doi: 10.1109/TNNLS.2021.3069230]

[27] Zheng YX, Xie JY, Sain A, Ma ZY, Song YZ, Guo J. ENDE-GNN: An encoder-decoder GNN framework for sketch semantic segmentation. In: Proc. of the 2022 IEEE Int'l Conf. on Visual Communications and Image Processing. Suzhou: IEEE, 2022. 1–5. [doi: 10.1109/VCIP56404.2022.10008880]

[28] Ribeiro LSF, Bui T, Collomosse J, Ponti M. Sketchformer: Transformer-based representation for sketched structure. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 14141–14150. [doi: 10.1109/CVPR42600.2020.01416]

[29] Tian JL, Xu X, Shen FM, Yang Y, Shen HT. TVT: Three-way vision Transformer through multi-modal hypersphere learning for zero-shot sketch-based image retrieval. In: Proc. of the 36th AAAI Conf. on Artificial Intelligence. AAAI Press, 2022. 2370–2378. [doi: 10.1609/aaai.v36i2.20136]

[30] Sezgin TM, Stahovich T, Davis R. Sketch based interfaces: Early processing for sketch understanding. In: Proc. of the 2007 ACM SIGGRAPH Courses. San Diego: ACM, 2007. 1–8. [doi: 10.1145/1281500.1281548]

[31] Kim DH, Kim MJ. A curvature estimation for pen input segmentation in sketch-based modeling. Computer-aided Design, 2006, 38(3): 238–248. [doi: 10.1016/j.cad.2005.10.006]

[32] Sun ZB, Wang CH, Zhang LQ, Zhang L. Free hand-drawn sketch segmentation. In: Proc. of the 12th European Conf. on Computer Vision. Florence: Springer, 2012. 626–639. [doi: 10.1007/978-3-642-33718-5_45]

[33] Schneider RG, Tuytelaars T. Example-based sketch segmentation and labeling using CRFS. ACM Trans. on Graphics, 2016, 35(5): 151. [doi: 10.1145/2898351]

[34] Qi YG, Song YZ, Xiang T, Zhang HG, Hospedales T, Li Y, Guo J. Making better use of edges via perceptual grouping. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 1856–1865. [doi: 10.1109/CVPR.2015.7298795]

[35] Wang F, Lin SJ, Li HH, Wu HF, Cai T, Luo XN, Wang RM. Multi-column point-CNN for sketch segmentation. Neurocomputing, 2020, 392: 50–59. [doi: 10.1016/j.neucom.2019.12.117]

[36] Jiang JK, Wang RM, Lin SJ, Wang F. SFSegNet: Parse freehand sketches using deep fully convolutional networks. In: Proc. of the 2019 Int'l Joint Conf. on Neural Networks. Budapest: IEEE, 2019. 1–8. [doi: 10.1109/IJCNN.2019.8851974]

[37] Gori M, Monfardini G, Scarselli F. A new model for learning in graph domains. In: Proc. of the 2005 IEEE Int'l Joint Conf. on Neural Networks. Montreal: IEEE, 2005. 729–734. [doi: 10.1109/IJCNN.2005.1555942]

[38] Lafferty JD, McCallum A, Pereira FCN. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. of the 18th Int'l Conf. on Machine Learning. Williamstown: Morgan Kaufmann, 2001. 282–289.

[39] Bucher M, Vu TH, Cord M, Pérez P. Zero-shot semantic segmentation. In: Proc. of the 33rd Conf. on Neural Information Processing Systems. Vancouver: NeurIPS, 2019. 466–477.

[40] Lu Y, Chen YR, Zhao DB, Chen JX. Graph-FCN for image semantic segmentation. In: Proc. of the 16th Int'l Symp. on Neural Networks. Moscow: Springer, 2019. 97–105. [doi: 10.1007/978-3-030-22796-8_11]

[41] Chen YP, Rohrbach M, Yan ZC, Yan SC, Feng JS, Kalantidis Y. Graph-based global reasoning networks. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 433–442. [doi: 10.1109/CVPR.2019.00052]

[42] Zhang L, Li XT, Arnab A, Yang KY, Tong YH, Torr PHS. Dual graph convolutional network for semantic segmentation. In: Proc. of the 30th British Machine Vision Conf. Cardiff: BMVA Press, 2019. 254.

[43] Liu J, Bao YQ, Ying WZ, Wang HC, Gao Y, Sonke JJ, Gavves E. Few-shot semantic segmentation with support-induced graph convolutional network. arXiv:2301.03194, 2023.

[44] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010. [doi: 10.5555/3295222.3295349]

[45] Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with Transformers. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 213–229. [doi: 10.1007/978-3-030-58452-8_13]

[46] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai XH, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. In: Proc. of the 9th Int'l Conf. on Learning Representations. OpenReview.net, 2021.

[47] Zheng SX, Lu JC, Zhao HS, Zhu XT, Luo ZK, Wang YB, Fu YW, Feng JF, Xiang T, Torr PHS, Zhang L. Rethinking semantic segmentation from a sequence-to-sequence perspective with Transformers. In: Proc. of 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 6877–6886. [doi: 10.1109/CVPR46437.2021.00681]

[48] Xie EZ, Wang WH, Yu ZD, Anandkumar A, Alvarez JM, Luo P. SegFormer: Simple and efficient design for semantic segmentation with Transformers. In: Proc. of the 35th Conf. on Neural Information Processing Systems. NeurIPS, 2021. 12077–12090.

[49] Liu Z, Lin YT, Cao Y, Hu H, Wei YX, Zhang Z, Lin S, Guo BN. Swin Transformer: Hierarchical vision Transformer using shifted windows. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 9992–10002. [doi: 10.1109/ICCV48922.2021.00986]

[50] Tripathi A, Mishra A, Chakraborty A. Query-guided attention in vision Transformers for localizing objects using a single sketch. In: Proc. of the 2024 IEEE/CVF Winter Conf. on Applications of Computer Vision. 2024. 1083–1092.

[51] Chen CQ, Wu YS, Dai QY, Zhou HY, Xu MT, Yang SB, Han XG, Yu YZ. A survey on graph neural networks and graph Transformers in computer vision: A task-oriented perspective. arXiv:2209.13232, 2022.

[52] Wu ZH, Jain P, Wright MA, Mirhoseini A, Gonzalez JE, Stoica I. Representing long-range context for graph neural networks with global attention. In: Proc. of the 35th Conf. on Neural Information Processing Systems. NeurIPS, 2021. 13266–13279.

[53] Rong Y, Bian YT, Xu TY, Xie WY, Wei Y, Huang WB, Huang JZ. Self-supervised graph Transformer on large-scale molecular data. In: Proc. of the 34th Conf. on Neural Information Processing Systems. Vancouver: NeurIPS, 2020. 12559–12571.

[54] Kreuzer D, Beaini D, Hamilton WL, Létourneau V, Tossou P. Rethinking graph Transformers with spectral attention. In: Proc. of the 35th Conf. on Neural Information Processing Systems. NeurIPS, 2021. 21618–21629.

[55] Ying CX, Cai TL, Luo SJ, Zheng SX, Ke GL, He D, Shen YM, Liu TY. Do Transformers really perform bad for graph representation? In: Proc. of the 35th Conf. on Neural Information Processing Systems. NeurIPS, 2021. 28877–28888.

[56] Li K, Pang KY, Song YZ, Xiang T, Hospedales TM, Zhang HG. Toward deep universal sketch perceptual grouper. IEEE Trans. on Image Processing, 2019, 28(7): 3219–3231. [doi: 10.1109/TIP.2019.2895155]

[57] Lin TY, Dollár P, Girshick R, He KM, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 936–944. [doi: 10.1109/CVPR.2017.106]

[58] Yun XL, Zhang YM, Yin F, Liu CL. Instance GNN: A learning framework for joint symbol segmentation and recognition in online handwritten diagrams. IEEE Trans. on Multimedia, 2021, 24: 2580–2594. [doi: 10.1109/TMM.2021.3087000]

[59] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional Transformers for language understanding. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics, 2019. 4171–4186. [doi: 10.18653/v1/N19-1423]

[60] Strudel R, Garcia R, Laptev I, Schmid C. Segmenter: Transformer for semantic segmentation. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 7242–7252. [doi: 10.1109/ICCV48922.2021.00717]

[61] Hussain S, Zaki MJ, Subramanian D. Global self-attention as a replacement for graph convolution. In: Proc. of the 28th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining. Washington: ACM, 2022. 655–665. [doi: 10.1145/3534678.3539296]

[62] Li L, Fu HB, Tai CL. Fast sketch segmentation and labeling with deep learning. IEEE Computer Graphics and Applications, 2019, 39(2): 38–51. [doi: 10.1109/MCG.2018.2884192]

[63] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Proc. of the 18th Int'l Conf.

on Medical Image Computing and Computer-assisted Intervention. Munich: Springer, 2015. 234–241. [doi: 10.1007/978-3-319-24574-4_28]

[64] Chen LC, Zhu YK, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 833–851. [doi: 10.1007/978-3-030-01234-2_49]

附中文参考文献:

[2] 王淑侠, 王守霞, 王关峰, 高满屯. 基于几何特征的在线手绘草图分割. 计算机辅助设计与图形学学报, 2015, 27(9): 1686–1693. [doi: 10.3969/j.issn.1003-9775.2015.09.012]

[3] 邓正根, 吕健, 刘翔, 侯宇康, 王帅. 基于 StyleGAN 的草图生成产品设计效果图方法研究. 包装工程, 2023, 44(6): 188–195. [doi: 10.19554/j.cnki.1001-3563.2023.06.020]

[6] 杨金凯, 王国中, 范涛. 基于神经网络的手绘草图的识别与匹配. 智能计算机与应用, 2021, 11(6): 148–152. [doi: 10.3969/j.issn.2095-2163.2021.06.028]

[8] 陈健, 白琮, 马青, 郝鹏翼, 陈胜勇. 面向细粒度草图检索的对抗训练三元组网络. 软件学报, 2020, 31(7): 1931–1942. http://www.jos.org.cn/1000-9825/5934.htm [doi: 10.13328/j.cnki.jos.005934]

[15] 王佳欣, 朱志亮, 邓小明, 马翠霞, 王宏安. 基于深度学习的草图分割算法综述. 软件学报, 2022, 33(7): 2729–2752. http://www.jos.org.cn/1000-9825/6299.htm [doi: 10.13328/j.cnki.jos.006299]

张拯明(1993－), 男, 博士生, 主要研究领域为人机交互, 计算机视觉.

邓小明(1980－), 男, 博士, 研究员, CCF 高级会员, 主要研究领域为计算机视觉, 人机交互.

郭燕(2000－), 女, 硕士生, 主要研究领域为人机交互, 计算机视觉.

王宏安(1963－), 男, 博士, 研究员, 博士生导师, 主要研究领域为自然人机交互, 实时智能计算.

马翠霞(1975－), 女, 博士, 研究员, 博士生导师, CCF 杰出会员, 主要研究领域为人机交互, 媒体大数据可视分析.