计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(12):233−242 [doi: 10.15888/j.cnki.csa.009340]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

相似度约束的交互式分割网络①

付　杰[1], 宣士斌[1,2], 江进宝[1]

[1](广西民族大学 人工智能学院, 南宁 530006)
[2](广西混杂计算与集成电路设计分析重点实验室, 南宁 530006)
通信作者: 宣士斌, E-mail: xuanshibin@gxmzu.edu.cn

摘　要: 交互式图像分割是像素级注释和图像编辑的重要工具. 现存方法大多采取两阶段预测, 首先预测一个粗糙的结果, 在第 2 个阶段细化之前预测的结果来得到更精确的预测, 为了使在硬件资源受限时, 网络模型仍可以使用, 基于此, 在两阶段共享同一个网络, 为了更好地将标记信息传播到未标记区域, 设计了一个相似度约束传播模块, 在训练时使用了一个简单的原型提取模块来使正点击向量高度内聚, 加速网络收敛, 在推理时移除. 在推理阶段通过使用意图感知模块来捕获细节, 使得预测性能进一步提升. 大量实验表明, 该方法在所有流行的基准测试上与最先进的方法最有可比性, 证明了其有效性.

关键词: 交互式分割; 图像分割; 语义分割; 图网络; 深度学习

引用格式: 付杰,宣士斌,江进宝.相似度约束的交互式分割网络.计算机系统应用,2023,32(12):233–242. http://www.c-s-a.org.cn/1003-3254/9340.html

Interactive Segmentation Network with Similarity Constraints

FU Jie[1], XUAN Shi-Bin[1,2], JIANG Jin-Bao[1]

[1](School of Artificial Intelligence, Guangxi Minzu University, Nanning 530006, China)
[2](Guangxi Key Laboratory of Hybrid Computation and IC Design Analysis, Nanning 530006, China)

Abstract: Interactive image segmentation is an important tool for pixel-level annotation and image editing. Most existing methods adopt two-stage prediction: first predicting a rough result, and then refining the previously predicted results in the second stage to obtain more accurate predictions. To ensure the viability of the network model under limited hardware resources, the same network is shared across the two stages. To better propagate labeled information to unlabeled areas, a similarity constraint propagation module is designed. Meanwhile, a simple prototype extraction module is used during training to make forward click vectors highly cohesive, accelerate network convergence, and remove them during inference. At the inference stage, the implementation of intention perception modules to capture details further improves prediction performance. Numerous experiments show that the method is most comparable to the most advanced methods on all popular benchmark tests, demonstrating its effectiveness.

Key words: interactive segmentation; image segmentation; semantic segmentation; graph network; deep learning

交互式图像分割在广泛的视觉任务中发挥着至关重要的作用, 例如图像编辑、医学图像分析和密集图像注释. 随着数据驱动的深度学习技术的普及, 在一些领域, 对掩码级注释的需求急剧增加, 如显著对象检测、语义分割、实例分割、伪装对象检测和图像/视频操作, 为了降低标注成本, 迫切需要高效的交互式分割技术. 因此, 越来越多的研究人员正在这一领域进行广泛的探索. 通过交互交互式图像分割的进展, 可以减少

① 基金项目: 国家自然科学基金 (61866003); 广西民族大学研究生教育创新计划 (gxun-chxs2021063)
收稿时间: 2023-06-19; 修改时间: 2023-07-19; 采用时间: 2023-07-27; csa 在线出版时间: 2023-10-30
CNKI 网络首发时间: 2023-10-31

人力标注成本, 为监督模型提供高精度真实掩码的大数据集成为可能, 由此来提高模型的泛化能力. 交互式分割有着悠久的研究发展历史, 探索了多种交互策略, 包括点[1–4]、涂鸦和边界框等. 虽然基于边界框[5–7]的方法可以快速定位目标对象, 而基于涂鸦的方法[8–11]提供了更丰富的用户输入线索, 但它们通常涉及更多的用户交互. 本文的工作主要关注基于点[12–16]的交互式分割, 它只提供点来指示图像中的前景或背景, 这通常需要人类付出较少的努力来标注图像.

交互式分割的一个关键特征是用户感兴趣区域的多样性. 与具有预定义标签空间的语义分割不同, 交互式分割任务中的每个实例可以根据用户的意图产生不同的前景区域. 因此, 基于点的交互式分割的主要挑战是找出用户意图, 并基于有限数量的用户交互将用户注释 (即点击) 传播到其他未标记的像素区域.

以前的大多数工作都专注于对每个单独的交互步骤进行建模, 并主要依靠稀疏注释的点击来推断用户意图. 这种无记忆策略在捕获前景时往往效率较低, 因为它忽略了先前步骤的预测. 为了缓解这一问题, 文献 [13]将最后一步预测作为推断目标区域的额外输入, 从而实现更有效的前景估计. 然而由于先前步骤中潜在的不准确掩模预测, 这种简单的融合方法可能会产生错误的信息传播. 针对标签传播的挑战, 文献 [17–22]侧重于标注对象边界, 执行基于区域的分割, 旨在更好地利用每个交互步骤中的用户点击, 这些方法通常利用堆叠卷积以隐式方式传播用户注释, 这在捕获长距离依赖性方面能力有限, 常导致前景掩码不完整. 为了更好地适应测试用例, 文献 [23]开发了一种反向传播细化方案, 以纠正测试时间中标记错误的用户点击. 尽管这些方法具有良好的性能, 但它们需要反向传播, 延长了推理时间. Liew 等人[24]试图根据正负点击对细化局部区域. Majumder 等人[25]生成内容感知引导图, 用于利用图像中的分层结构信息.

最近, Chen 等人[12]采用完全连通的图网络[26]开发了 CDNet, 以促进全局和局部的信息传播. 由于极高的计算复杂性, 全连接的图网络只能处理相对低分辨率的特征图, 这可能导致不准确的前景边界. 此外, 两个传播阶段都容易受到依赖于先前预测或颜色相似性的噪声亲和性估计的影响. 与之类似, 文献 [27]提出了将用户点击的向量通过与特征图交互来将用户标注的信息传播到未标注的地方, 使用骨干网络 ResNet[28]输出的两个特征图, 大小分别为原图的 1/2 和 1/4, 它基于两阶段的模型且不共享参数. Lin 等人[29]认为当前点击与先前的点击一起确定全局预测可能会削弱新输入的点击对周围细节的决定性影响, 并反馈不满意的结果, 因此, 开发了 FocusCut[29]将每个点影响的范围裁剪出来, 使用同一网络做进一步的细化, 允许交互式分割网络不仅分割目标对象, 而且修复局部细节. 通过对点击点的局部细节优化来提升结果, 它们使用同一个网络, 降低了参数. FocalClick[30]同样也利用了用户提供的点击信息在同一特征向量空间中以点击点为中心做裁剪做进一步的细化. SimpleClick[31]通过简单的视觉自注意力模型 (ViT) 主干捕获全局信息, 实现精确分割. PiClick[32]同样使用了 ViT 主干, 对相互交互的掩码查询注入目标先验进一步地提升了分割性能.

FocusCut[29]通过将 DeepLabv3+[33]中的第 1 个卷积层的输入通道 3 变为 6, 其余模块没变, 也没有引入任何其他额外的模块, 本文的工作是基于 FocusCut 改进而来的, 提出了一种用于交互式图像分割的新颖用户标记特征传播策略. 其主要想法包括两个方面: 1) 基于稀疏图神经网络 (GNN) 学习点击增强特征表示, 为了得到特征向量与用户提供的点击点向量最接近和差异最大的向量且抑制噪声影响, 引入了相似度最大最小约束; 2) 为了加速收敛, 提出了一个简单的原型提取模块用于加强训练样本中正点击特征的表征, 使得前景特征高度内聚. 没有采用 FocusCut 中的用户点击点影响范围的局部细节优化, 而是使用同一网络相同权重, 不仅可以预测粗糙结果, 也可以对预测的粗糙结果做整体细化. 细化之前用意图感知模块来提取用户感兴趣的区域.

总之, 本文的贡献如下.

(1) 基于一个稀疏图神经网络开发了一种新的点击传播策略, 能够捕获对高分辨率特征图的长依赖, 在相似度最大最小的约束下, 抑制噪声影响, 可以更好地将用户标记信息传播到未标记区域.

(2) 提出了一种提取前景特征向量表征原型的方法, 能够使正点击特征向量高度内聚, 促使模型学习更强的表征, 加速模型收敛.

(3) 通过意图感知模块对用户感兴趣区域放大使网络模型较好的捕获细节, 并且使用同一个网络模型, 减少了模型对显存的需求.

该方法在大多数公共基准上达到了较先进的结果, 证明了设计的有效性.

1 相关工作

1.1 问题设置和方法总览

交互式分割的目标是正确推断用户感兴趣的区域, 并用尽可能少的点击来分割目标对象. 该顺序估计任务通常被转换为一系列前景分割问题, 其中每个问题的目标是: 在已知给定用户的当前点击集的情况下, 在每个交互步骤中尽可能准确地输出前景掩码. 形式上, 假设生成一组图像点击对作为训练数据D, 它包括数据元组$\{(I_i,U_i,Y_i)\}_{i=1}^{|D|}$, 其中$I_i$表示输入图像, $U_i=\{(u_{(i,j)},l_{(i,j)})\}_{j=1}^{M_i}$是一组用户点击, 由其像素索引$u_{(i,j)}$和标签$l_{(i,j)}\in\{pos,neg\}$组成, 其中$M_i$是第$i$副图像的点击总次数, Y_i表示第i副图像的真实掩码, pos表示正点击, neg表示负点击, $u_{(i,j)}$保存的是第i副图像用户提供的第j个点击点的横纵坐标. 当$l_{(i,j)}=pos$表示i副图像$u_{(i,j)}$位置上的像素属于前景, 当$l_{(i,j)}=neg$表示i副图像的$u_{(i,j)}$位置上的像素属于背景. 目标是学习一个基于数据集$\mathcal{D}$的神经网络M_θ, 使损失函数$\mathcal{L}$最小, θ为网络参数, E表示在不同的数据集上做的训练, 由于训练过程中有用随机对数据进行增强处理, 原始数据集一样, 但每次训练时数据集有差别. 可以表示为式 (1):

$$\min_\theta E_{(I_i,U_i,Y_i)\sim D}\,\mathcal{L}(\mathcal{M}_\theta(I_i,U_i),Y_i) \tag{1}$$

交互式分割可以被视为一种特定类型的分割任务, 许多方法都是基于经典的语义分割网络 DeepLab 系列设计的, 尤其是 DeepLabv3+[33]. 该网络体系结构包括骨干网络、ASPP (atrous spatial pyramid pooling) 和解码器 3 部分. 很多交互式分割方法都用 ResNet[28]作为骨干网络; ASPP 部分包含 4 个扩张卷积分支和 1 个全局平均池化分支; 解码器部分通过融合主干的低级别特征来细化 ASPP 模块的输出, 以生成最终预测. 对于交互式分割, 输入应该包含交互的信息, 其中点击位置将转换为两个点击图, 用来表示正点击图和负点击图. 点击图可以为距离、圆盘和高斯图等. 例如点击图为圆盘图 (disk maps), 现在生成正点击图, 首先生成一个全为 0 单通道图, 该单通道图的大小和输入图像一样, 当用户提供的标签为pos时, 根据像素索引在其对应位置置为 1, 如果是负点击图, 那么用户提供的标签为neg时在其相应位置置为 1. 大多数交互式分割工作都修改了网络的输入部分, 并将 RGB 图像、正负点击图和之前预测图在通道维度上拼接起来组成 6 通道图作为输入, 然后通过添加另一个卷积层来将 6 通道映射编码为 3 通道映射, 然后输入 DeepLabv3+做预测. 本文是将 DeepLabv3+的第 1 个卷积层的输入通道 3 变为 6.

1.2 图神经网络

图神经网络 (GNN)[34,35]已被广泛用于建模长依赖关系, 然而, 用于交互式分割任务比较少见. 它的计算公式可以总结为式 (2) 和式 (3), 通过式 (3) 得到的$\widehat{A}$表示依赖矩阵, $\widehat{A}[i,j]$用于测量第i、j位置特征之间的依赖性. 然后计算式 (2) 得到增强后的特征. x代表输入特征. g、θ、ϕ是用 1×1 卷积实现的变换函数.

$$y^{HW\times C}=Softmax(\widehat{A}^{HW\times HW})\times g(x)^{HW\times C} \tag{2}$$

$$\widehat{A}=\theta(x)^{HW\times C}\times\phi(x)^{C\times HW} \tag{3}$$

其中, x中的信息可以在每两个位置之间远距离传播, 这有助于在全局上下文中构建更统一的特征表示. 这里将图像特征中的每一个与其他位置建立关系, 实现不同的点之间的信息相互传播. 上标代表特征图的形状大小, 其中H、W、C分别代表特征图的高度、宽度和通道数. $Softmax$对$\widehat{A}$每一行归一化为一个概率分布向量.

2 相似度约束的交互式分割网络

在本节中, 首先概述了提出模型的架构. 随后分别详细介绍了两个主要组件的设计, 包括相似度约束传播模块和原型提取模块. 最后介绍文献 [27]提出的意图感知模块.

对用户的每一次点击, 提出的模型分两阶段预测, 第 1 阶段的粗粒度网络, 它基于图像、前一步预测和用户的点击输入预测前景掩码, 然后通过意图感知模块自适应地放大前景区域并对预测图和用户提供的点击点坐标做相应变化后送入细粒度网络进行处理, 并将细化后得到的结果拷贝到粗粒度网络预测图中的相应位置, 来得到当前交互最终的预测图. 如图 1 所示. 图中红色圆圈代表前景, 绿色圆圈代表背景. 值得注意的是本文的粗粒度网络和细粒度网络是同一个网络, 并且共享参数, 网络由两部分组成, 如图 2 所示的语义分割模型 (segmentation) 为第 1 部分, 这里选用 DeepLabv3+, 第 2 部分为设计的相似度约束传播模块, 它将语义分割模型输出的特征图作为输入, 将用户提供的点击点信息传播到特征图中的每一个特征向量, 然后通过一个卷积层做预测, 通过上采样输出一个大小为$H\times W\times 1$的预测图, 以及为了快速收敛而设计的原型提取模块.

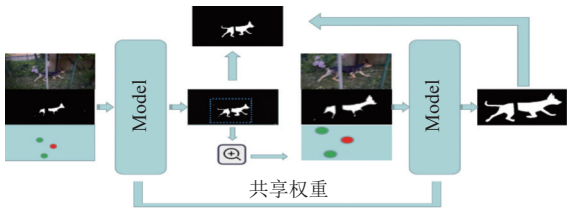


图 1 网络模型的工作流程

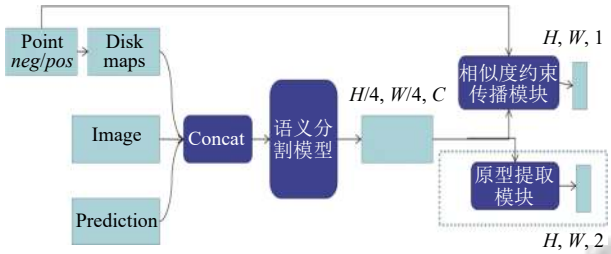


图 2 网络的组成部分

2.1 相似度约束传播模块 (SPM)

相似度约束传播模块是基于文献 [27]中的特征传播模块 FPM 进行改进得到, FPM 可以用式 (4) 和式 (5) 表示.

$$\widehat{f_n} = f_n + \sum_{j=1}^{M} \widehat{\alpha}(f_n, f_{u_j}) W_c^{\mathrm{T}} f_{u_j},\ \forall f_n \in F \tag{4}$$

$$\widehat{\alpha}(f_n, f_{u_j}) = e^{\theta(f_n)^{\mathrm{T}}\phi(f_{u_j})}/Z_n(F) \tag{5}$$

其中, $\widehat{f_n}$是更新后的特征, f_n是语义分割网络提取的特征, f_{u_j}是用户的点击向量, $\widehat{\alpha}$是注意力函数, θ和ϕ是转换函数, W_c是特征的变化权重矩阵, M是用户提供的点击点总数. 简单地说 FPM 将图像特征与用户提供的点击点向量做注意力交互, 详情请参考文献 [27], FPM 将用户提供的点击信息不加区分的用来构建注意力, 隐式建模图像特征与用户点击向量的关系, 使得网络需要多轮训练来学习模型. 我们认为图像的每一个未标记区域只需要查找与它最近的用户提供的点击信息来确定本身是否属于前景或背景, 然而根据空间距离来判断带有偏置, 且效果不理想. 为此, 本文使用余弦相似度来度量两个特征的语义距离, 由此得到相似度约束传播模块 (SPM). FPM 模块将图像特征图中每一个特征向量与用户提供的点击点向量做注意力来使信息融合, 而 SPM 模块是在图像特征向量与用户点击向量做注意力时添加条件也就是约束, 为每个图像特征向量从用户提供的点击点向量中选出两个具有代表性的向量, 也就是余弦相似度最大最小的两个向量做融合来增强向量表征, 这使得模型更容易学习到可判别的特征向量表征, 得到更精确的预测. 下面将进行详细介绍.

形式上, 将语义分割网络输出的特征图记为F, 大小为$H\times W\times C$, 其中H和W分别是特征图的高度和宽度, C是特征的通道数. 为了构建稀疏图, 首先根据用户点击坐标U, 通过双线性采样得到用户点点击向量, 记为F_u, 大小为$M\times C$, M为用户提供的点击点的个数, α表示余弦相似度函数, 如式 (11) 所示, 首先计算特征图F和用户点击向量F_u的相似度矩阵A, 大小为$H\times W\times M$, $A_{i,j,m}$表示第$i\times W+j$个特征图向量与用户提供的第m个点击向量的相似度值, 然后通过式 (9) 选出每一个特征图向量相似度最大的特征向量, 具体操作为通过式 (7) 将相似度最大值赋值为 1, 其余值置为 0 得到的张量与F_u做矩阵乘积后, 与F在通道维度拼接后执行一个1×1的卷积降为原来的通道数, 记为$\widehat{F}_{pos}$, 然后与之类似的可以找出差异最大 (余弦相似度最小) 的用户点击特征向量, 与F在通道维度拼接后执行一个1×1的卷积降为原来的通道数, 记为$\widehat{F}_{neg}$, 将$\widehat{F}_{pos}$、$\widehat{F}_{neg}$和F做逐元素相加, 输入一个简单的前向反馈网络 (FFN) 后, 得到更新后的特征$\widehat{F}$, 前向反馈网络 (FFN) 由一个1×1的卷积、一个ReLU激活和一个1×1的卷积组成. 最后将$\widehat{F}$与F做逐元素相加后输入一个卷积层做前景概率预测. 实际引入了多头机制, 为了使公式简洁, 没有在公式中标出. 该模块的执行过程可用式 (9)–式 (12) 描述, 其中Conv表示一个1×1的卷积, 如图 3 所示. $max_smi(A)\otimes F_u$做矩阵乘法的具体过程为: 先将$max_smi(A)$从$H\times W\times M$变形为$(H\times W)\times M$, 与F_u做矩阵乘法后恢复成$H\times W\times C$. 同样的$min_smi(A)\otimes F_u$之类似.

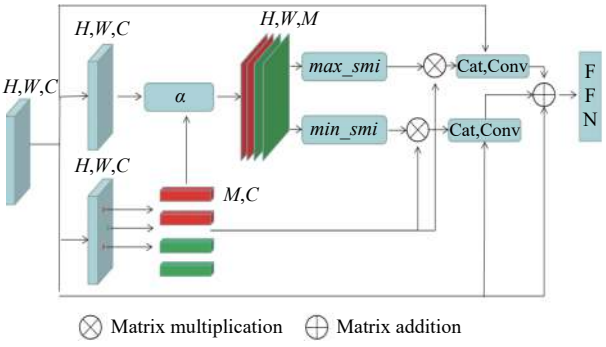


图 3 相似度约束传播模块

$$A_{i,j,m} = \alpha(f_{i,j}, f_{u_m}),\ i\in[1,H],\ j\in[1,W],\ m\in[1,M] \tag{6}$$

$$max\_smi(A) = \begin{cases} B_{i,j,m} = 1, \text{ if } A_{i,j,m} == \max(A_{i,j}) \\ B_{i,j,m} = 0, \text{ if } A_{i,j,m} < \max(A_{i,j}) \end{cases} \tag{7}$$

$$min\_smi(A)=\begin{cases} B_{i,j,m}=1, \text{ if } A_{i,j,m}==\min(A_{i,j}) \\ B_{i,j,m}=0, \text{ if } A_{i,j,m}>\min(A_{i,j}) \end{cases} \tag{8}$$

$$\widehat{F}_{pos}=Conv(Cat(F,max\_smi(A)\otimes F_u)) \tag{9}$$

$$\widehat{F}_{neg}=Conv(Cat(F,min\_smi(A)\otimes F_u)) \tag{10}$$

$$\widehat{F}=FFN(\widehat{F}_{pos}+\widehat{F}_{neg}+F) \tag{11}$$

$$\alpha(f_n,f_m)=\frac{f_n\cdot f_m}{\|f_n\|\times\|f_m\|} \tag{12}$$

在相似度约束传播模块与中, 图像中的每个特征只与用户提供的相似度最大最小的特征做关联, 而文献 [27]中 FPM 是与用户提供的所有特征做关联. 在消融部分, 经实验证明在不损失性能的情况下, 相似度模块可以较快的收敛. 直观地说, 用户提供的特征具有代表性, 通过寻找与用户提供的相似度最大最小的特征来提高特征表征, 理想情况下使得用户感兴趣区域和背景区域在高维空间中线性可分, 然而, 这些增强特征只是经过卷积层与真实掩码做损失, 未能充分的挖掘正点击特征的原型, 为此引入了一个非常简单的原型提取模块.

2.2 原型提取模块 (PEM)

Zhang 等人[36]首次提出了掩码平均池化 *MAP* (masked average pooling) 策略, 来得到支持图像的代表特征, 这里假设$I\in R^{(B\times 3\times W\times H)}$及其二值掩码$G\in\{0,1\}^{(B\times 1\times W\times H)}$, 其中$W$和$H$是图像的宽度和高度, B为批次大小. 设I的输出特征图是$F\in R^{(B\times C\times W\times H)}$. 然后通过式 (13) 来得到一个$B\times C$的张量, 表示$B$副图像的代表特征, 特征维度为$C$, 其中$F\times G$做张量的乘积, $sum_{(h,w)}(\cdot)$表示对高、宽维度求和. 为了获取图像的前景和背景原型这里使用了掩码平均池化 *MAP*, 下面来进行详细介绍.

将每批次经过语义分割网络输出的特征F与真实掩码G做 *MAP* 后送入$mean_b(\cdot)$得到一个大小为$1\times C$的向量, 也就是当前批次正点击的原型, 记为Pro_{pos}, 然后与之前得到的原型Pro_{pos}^{pre}送入式 (18) 做动量更新, $mean_b(\cdot)$表示对批量维度求均值, 在计算完损失后将Pro_{pos}赋值给Pro_{pos}^{pre}. 负点击的原型Pro_{neg}通过将输出的特征与取反后的真实掩码做 *MAP* 得到, 大小为$B\times C$, 它没有做动量更新, 将正负原型与F做向量内积得到$y\hat{\in}R^{(B\times 2\times H\times W)}$, 与真实掩码做交叉熵损失, 正负原型的获取可以用式 (13)–式 (18) 表示:

$$MAP(F,G)=sum_{h,w}(F\times G)/sum_{h,w}(G) \tag{13}$$

$$Pro_{pos}=mean_b(MAP(F,G)) \tag{14}$$

$$Pro_{pos}=EMA(Pro_{pos},Pro_{pos}^{\text{pre}}) \tag{15}$$

$$Pro_{pos}^{\text{pre}}=Pro_{pos} \tag{16}$$

$$Pro_{neg}=MAP(F,1-G) \tag{17}$$

$$EMA(A,B)=\beta\times A+(1-\beta)\times B \tag{18}$$

实验中设β为 0.01, 通过原型提取模块, 这样可以有效地提取正点击的特征, 使得正点击的前景特征向量更内聚. 没有使用卷积层来做约束, 而是直接使用神经网络输出的数据将正点击特征聚拢在某个特定的空间. 交互式分割是一个类不可知的语义分割, 它将输入图像分割成前景和背景, 由于网络会根据用户输入的正点击来确定用户想要的目标, 在经过网络的特征编码后, 图像属于前景的特征向量应该在语义上尽可能的接近, 而图像中属于背景的特征向量由于多样性 (图像中背景相较于前景来说, 它的数量和种类远超前景) 和不确定性 (同一幅图像, 由于用户的不同, 感兴趣区域也不同) 导致不同的图像背景的高维特征不一定相似. 所以对每一幅图像单独提取它的背景, 不进行图像之间的加权平均, 相反由于前景图像是根据用户提供的正点击向量来确定的, 不同图像的前景在高维空间上是很接近的, 可以将不同图像的前景做加权平均, 为了更好地与训练数据中的图像前景特征关联和增强鲁棒性, 保存了前景原型 (正点击原型) 特征向量, 在每批次中与当前原型相加来更新. 通过使用原型提取模块, 使得语义分割网络更容易判别前景和背景, 进而网络模型更快收敛.

2.3 意图感知模块 (IAM)

意图感知模块是由文献 [27]提出的, IAFPN[27]提出的模型包含一个粗粒度网络和一个细粒度网络, 粗粒度网络首先对图像、用户提供的点和上一次预测的前景做处理, 得到前景概率图, 通过阈值二值化后送入意图感知模块, 意图感知模块首先得到前景最小外接包围框, 然后通过自适应边缘来扩展外接包围框, 边缘大小与初始框的大小成反比, 使得大对象具有相对紧凑的边界框, 该边界框排除了分散注意力的背景, 而小对象采用相对宽松的边界框来包括更多的上下文提示. 通过意图感知模块得到包围框后, 对原图像裁剪后送入细粒度网络做细化. 意图感知模块没有任何学习参数, 本文没有采用自适应边缘算法, 而是通过考虑用户提供的点来扩展外接包围框. 具体来说, 在得到前景最

小外接包围框后, 取包围框的左上角、右下角和用户提供的点坐标, 再一次的计算一个最小外接包围框, 为了不使包围框处在用户提供的点上, 将计算的外接包围框向外扩充了 5 个像素. 然后通过扩展后的包围框对图像做细化处理. 文献 [27]所采用的算法如果用户提供的点击点不在包围框内, 会导致细粒度网络不会使用该点提供的信息, 这样处理不会将用户提供的点击点做剔除, 会使用全部的点击信息来增强特征表征, 使得模型更容易识别感兴趣对象.

在用户的每一次交互中, 本文的模型与 IAFPN [27] 都会对图像进行两次处理, 一次为粗粒度预测, 另一次为细粒度预测. 不同的是粗粒度预测和细粒度预测 IAFPN 使用的是不同的网络. 本文使用的是同一个网络. 因此减少了显存的使用. 意图感知模块将用户感兴趣区域裁剪并放大来进行处理, 首先可以对部分背景对剔除, 其次做相应的放大可以使网络捕获感兴趣区域的细节, 预测更准确. 意图感知模块的引入可以使网络不仅预测粗糙的结果, 同时对用户感兴趣的区域做进一步的细化, 使得模型网络预测更精确. 可以在 3.3 消融中看到意图感知模块对预测精度的影响.

3 实验

在本节中, 首先描述了实验设置和实现细节, 然后将模型与现有工作进行比较, 然后进行消融研究以验证每个组件, 最后可视化了几个示例来进一步地说明相似性传播模块的有效性.

3.1 评估和实现细节

实验采用在 ImageNet 上预训练的 ResNet 作为特征提取器. 实验中, 设置训练过程 80 轮, 批次大小为 8, 初始学习率为 1E–3、每轮$\gamma = 0.9$的指数学习率衰减策略. 采取动量为 0.9, 权重衰减为 5E–4 的随机梯度下降进行参数优化, 并在 40 轮的时候将正点击原型置为 0, 来重新获取正点击原型. 使用随机翻转和放大策略来增强数据, 图像大小为384×384. 算法使用 PyTorch 框架实现. 具体硬件设备和软件环境如表 1 所示.

评估指标遵循文献 [37]的工作, 采用了相同的机器人用户来模拟点击. 简单地说, 通过比较真实掩码和预测, 下一次点击将被放置在最大误差区域的中心. 采用点击次数 (NoC) 作为评估指标, 它计算实现预定交并比 (IoU) 所需的平均点击次数. 将目标 IoU 设置为 85% 和 90%, 分别表示为NoC@85和NoC@90, 每个实例的默认最大点击次数限制为 20 次, 并且还将报告无法到达目标 IoU 的失败次数 (NoF). NoC越小性能越好.

表 1 实验所需的硬件软件信息

软硬件	配置详情
处理器	Intel Xeon 4114
内存	128 GB
显卡	RTX8000
操作系统	Ubuntu 20.04
Python	3.8
PyTorch	1.11.1
Cuda	11.4

为了训练提出的深度网络进行交互式分割模型, 首先利用模拟过程从前景分割数据集生成一组图像点击对, 具体细节请参考文献 [13]. 给定训练数据集, 然后采用大多数先前工作[13]普遍使用的训练策略来训练提出的模型.

3.2 实验结果比较

与最先进比较 通过遵循标准评估协议, 在广泛的数据集上评估该方法, 包括 GrabCut[5]、Berkeley[37]、DAVIS[38]和 SBD[39].

GrabCut 是一个典型的交互分割数据集, 它包含 50 幅具有可区分前景和背景的图像. Berkeley 包含来自其测试子集的具有 100 个对象掩码的 96 个图像. DAVIS 最初被引入用于视频分割. 在实验中, 仅使用了 345 个带有精细标记对象的随机采样帧. SBD 包含用于 2 820 个图像的 6 671 个对象掩码.

如表 2 给出了在 4 个数据集上提出方法与其他方法的NoC指标实验结果, NoC越小效果越好. 相较于 FocusCut[29]来说, 本文的模型有较好的性能, 在 ResNet50 的骨干网络下, 在 GrabCut、Berkeley、DAVIS 和 SBD 数据集上比它好 0.10、0.25、019、0.1、0.18 和 0.21, 由于 FocusCut[29]在用户的每一轮交互中都需要将之前确定的点的影响区域放大后输入网络来做预测. 在最坏情况下, 每轮交互下都要进行局部细节优化, 假设用户交互n次, 它的执行复杂度为$O(n^2)$, 本文算法是针对算法预测的前景图和用户点击点的最大包围框下做细化, 运行次数不随点的增多而增多, 所以执行复杂度为$O(n)$, 在理论上, 本文算法运行时间要短一些. 相较于最先进的模型 IAFPN[27], 本文提出的模型在 GrabCut、Berkeley 上优于它, 在 ResNet-101 骨干上分别为 0.45 和 0.12, 而在 SBD, DAVIS 数据集上效果不佳, 这两个测试数据集上真实掩码细节边缘丰富, 由于

模型没有使用更大分辨率更低语义的特征, 对于需要更多细节的情况下, 效果较差.

表 2 与最先进算法比较

Method		GrabCut	Berkeley	DAVIS	SBD
		NoC@90	NoC@90	NoC@85/90	NoC@85/90
DOS[3]	FCN	6.08	8.65	9.03/12.58	9.22/12.80
RIS[24]	—	5.00	6.03	—/—	6.03/—
BRS[23]	DenseNet	3.60	5.08	5.58/8.24	6.59/9.78
CMG[25]	FCN	3.58	5.60	—/—	—/—
CDNet[12]	ResNet-50	2.64	3.69	5.17/6.66	4.37/7.87
FocusCut[29]	ResNet-50	1.78	3.44	5.00/6.38	3.62/5.66
IAFPN[27]	ResNet-50	2.31	3.35	4.52/5.83	3.08/4.98
Ours	ResNet-50	1.68	3.19	4.81/6.28	3.44/5.45
IS+SA[40]	ResNet-101	3.07	4.94	5.16/—	—/—
FCA[14]	ResNet-101	2.14	4.19	—/7.90	—/—
f-BRS[4]	ResNet-101	2.72	4.57	5.04/7.41	4.81/7.73
CDNet[12]	ResNet-101	2.76	3.65	5.33/6.97	4.73/7.66
FocusCut[29]	ResNet-101	1.64	3.01	4.85/6.22	3.40/5.32
FCFI[41]	ResNet-101	1.80	2.84	4.75/6.48	3.26/5.35
FocalClick[30]	segformerB0-S2	1.90	3.14	5.02/7.06	4.34/6.51
IAFPN[27]	ResNet-101	2.15	3.20	4.51/5.8	2.98/4.83
Our	ResNet-101	1.60	2.91	4.71/6.10	3.25/5.12

100 次点击分析: 在文献 [4,42]之后, 报告在具有 ResNet-50 主干的 DAVIS 数据集上, 最大点击次数限制为 100 的NoC@90. NoF_{20}@90表示当用户点击 20 次, 数据集中未达到 IoU=90% 的失败案例总数, NoF_{20}@90和NoF_{100}@90这两个指标可以衡量模型对用户提供点击点的利用情况, 随着点数的增多模型应该会预测的更准确, 失败总数应该减少. 在表 3 中, 本文的模型在 20 次点击的设置下弱于 FcousCut[29]和 IAFPM[27], 但随着点击次数增加到 100, 在NoF_{100}@90指标下, 提出的模型好于 IAFPN, 预测失败的图片比它少 3 张, 用户标记的信息未经过滤传播会引入噪声, 导致性能下降, FoucsCut[29]的局部区域细化一定程度上缓解了随着点数增加引入噪声的问题, 而使分割更精确, 但由于随着点数的增加, 它的运行次数也会相应的增加, 而本文的模型使用相似度最大最小约束来避免噪声的干扰, 可以有效地利用新增加的点击信息, 总的来说, 本文相似度约束传播模块可以有效地利用用户的点击信息来提升分割的精度.

3.3 消融实验

在这一部分, 进行了几个实验来评估方法中每个组件的有效性. 所有消融实验均在 DAVIS 和 SBD 数据集上使用 NoC@90 度量进行评估, DAVIS 和 SBD 相比其他数据集更具挑战性.

表 3 在 100 次点击下的分析

Method	NoF_{20}@90	NoC_{100}@90	NoF_{100}@90
f-BRS[4]	78	0.70	50
CDNet[12]	65	18.59	48
FocusCut[29]	57	17.42	43
IAFPN[27]	60	17.68	46
Ours	61	17.58	43

每个组成部分的有效性: 表 4 中的结果表明, 每个组成部分在整个框架中发挥着关键作用, 并有助于最终结果. 表 4 的 Baseline 是使用表 1 的设置下得出的实验结果. 其中的 Baseline 是基于 DeepLabv3+实现的, 通过将它的第 1 个卷积层输入通道 3 变为 6 来满足输入通道要求. 对于相似度约束传播模块 (SPM), 它分别在 DAVIS 和 SBD 数据集上改进了 0.64 和 1.23. 此外, 通过结合文献 [27]的意图感知模块 (IAF), 在两个数据集上进一步提高了性能, 这表明 IAF 可以更好地在粗略级别上估计前景, 并在精细级别上准确地分割对象. 由于训练中使用了放大策略的数据增强方法, 在每一轮交互中通过简单的意图感知即使使用同一个网络权重, 性能也都得到了提升, 分别为 0.54 和 1.07. 原型提取模块相比相似度约束传播模块, 它虽然可以提升网络的性能, 但提升效果较差, 原因是它没有增加学习参数, 它想要使所有的前景向量极度相似, 而相似度约束传播模块增加了学习参数, 并且让用户标记的向量传播到未标记区域, 弥补了卷积的局部感受野, 通过建立长距离联系来增强特征的表征能力.

表 4 不同组件的影响

Method	Components			DAVIS	SBD
	SPM	PEM	IAM		
Baseline	—	—	—	7.43	7.13
①	√	—	—	6.79	5.90
②	—	√	—	6.99	6.40
③	—	—	√	6.89	6.20
Ours	√	√	√	6.28	5.45

相似度约束传播模块: 为了验证稀疏图设计在相似度约束传播模块 (SPM) 中的有效性, 进行了更多的实验来与 CDNet[12]中的特征扩散图模块 (FDM)、IAFPN[27]中的特征传播模块 (FPM) 进行比较, FDM 实际上是一个完全连接的图网络. FPM 是一个无约束的稀疏图网络, 为了进行公平的比较, 所有实验都建立在 CDNet[12]中采用的基线之上. 表 5、表 6 中带有 Baseline*

的数据是引用了文献 [27]的结果. 其中 Baseline*与表 4 中的 Baseline 结构一致. 如表 5 所示, 需要说明的是 FPM 是需要在原图像的 1/2 和 1/4 的下采样下进行特征传播, SPM 仅需要在原图的 1/4 下采样下做特征传播, 而且在训练的时候发现在 80 轮就已经收敛了. 可以说明基于相似度模块可以快速地使模型收敛. 由于机器, 运行环境的原因, 未能实现文献 [27]的精度, 考虑到采用了高分辨率的特征图会使推理时间加剧. 并且本文与文献 [27]都采用了每轮交互中都需要进行两轮推理, 所以在论文中就没有与比较. 引用它论文中的数据来进行参数和浮点数计算的比较, 在表 6 展示. 在 FocusCut 的基础上增加了相似度约束传播模块 (SPM), 增加了 0.39M 的参数和 3.82G 浮点运算量. 与 FPM 增加的 22.69G 浮点运算量相比, 本文模型对计算资源更友好些.

表 5 图设计分析在 DAVIS 数据集下

Method	NoC_{20}@90	NoC_{100}@90	epoch
Baseline*	6.60	8.42	120
Baseline*+FDM[12]	5.40	7.64	120
Baseline*+FPM[27]	5.05	7.17	120
Baseline*+SPM	5.07	7.19	80

表 6 模型的参数量和浮点运算量

Method	Params (M)	FLOPs (G)
Baseline*	31.4	508.72
Baseline*+FPM[27]	31.5	531.42
FocusCut[29]	40.36	41.01
Ours	40.75	44.83

原型提取模块 为了验证原型提取模块可以加速网络收敛, 进行了简单的消融实验, 统计了每轮的损失值, 如图 4 所示.

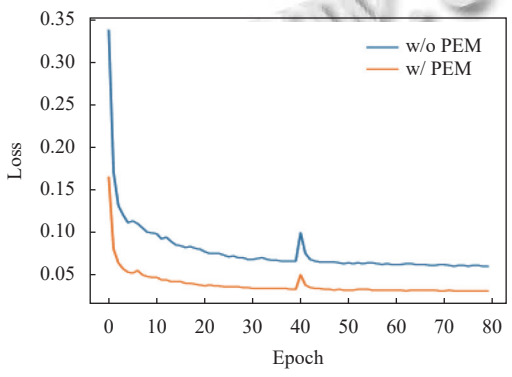
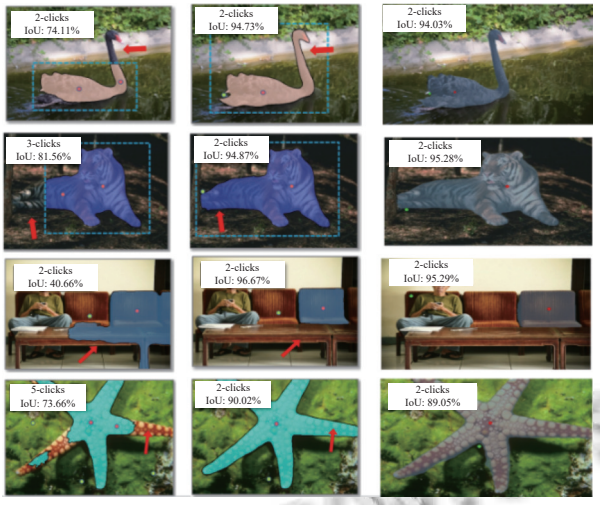


图 4 有无 PEM 的损失

通过引入 PEM, 在相同设置下, 模型收敛的更快. 在第 40 轮时候, 由于将图像之前的预测掩码和原型置为 0, 重新模拟用户的点击信息, 导致 loss 升高. 训练轮数与 FocusCut 一样. 在每一轮中, FocusCut 网络模型需要对图像的整体和感兴趣区域的局部进行监督, 虽然训练轮数为 40 轮, 但如果将网络的输入分为整幅图像和图像中的局部区域, 它的总训练轮数为 80 轮. 为了更好地与 FocusCut 进行比较, 设置了训练轮数为 80 轮.

可视化分析, 相似度约束传播模块 SPM 与 IAFPN[27] 中的特征传播模块 (FPM) 都是基于稀疏图构建的, 为了更直观的比较, 模型预测结果和预测结果与真实掩码的交互比值 (IoU) 都画在了原图上, 如图 5 所示, 引用了文献 [27]中的可视化结果, 图中前两列来自 IAFPN[27], 在相同的点击下进行 IoU 比较, IoU 显示在每幅图的左上角, IoU 越大表示预测的越好. 在图 5 中的第 1 行图中, 基础模型不能捕获全局信息, 导致未能预测出图中鹅的头部, IAFPN 和本文模型可以很好地预测出鹅的主体, 图 5 中的第 2 行图, 在给定两个正点击点的情况下, 基础模型未能预测预测出老虎身体的后半部分, IAFPN 和本文模型却在一个正点击点和一个负点击点的协助下预测出完整老虎的主体. 对图 5 中的第 3 行图, 基础模型未能很好地将负点击点信息利用, 导致仍有部分预测错误, IAFPN 和本文模型很好地将负点击信息传播到其他未标记区域, 很好的预测了用户想要的目标即最左边的沙发. 图 5 中的第 4 行图, 海星主体有很多相似的纹理, 在提供了 5 个用户点的基础上, 基础模型未能捕获图像区域间的联系, 导致预测出真实掩码的部分, IAFPN 和本文模型在相同点击 (两个用户点) 提示的基础上, 准确的捕获了图像区域间的联系, 预测出完整的海星. 本文的模型优于基准模型, 与 IAFPN 一样, 都是将用户提供的信息传播到未标记区域, 区别在于相似度模块有选择地将用户提供的信息做融合, 图 5 中 4 行提出的模型与基准相比, 可以预测出完整的物体, 表明模型有能力长距离的建模图像区域间的联系. 但在相同的点击下得到的 IoU 较弱于 IAFPN, 由于本模型没有使用更低级的特征, 导致预测的边缘细节不如使用了更低级图像特征的 IAFPN, 在第 3.2 节中, 通过在 DAVIS 数据集上 100 次点击分析的比较, 对比 IAFPN 中的特征传播模块 (FPM), 相似度约束传播模块 SPM 可以抑制噪声影响, 更充分的利用用户提供的点击信息. 相较于 IoU 的微弱提升, 更好地利用点击信息使用户在给定的点击次数下达到满意的 IoU 更重要, 因为可以减少用户标注的时间. 总的来说, 相似度约束传播模块有能力长距离

的建模图像区域间的联系, 预测出完整的物体, 并且抑制噪声影响, 可以更好地将用户标记信息传播到未标记区域.



(a) Baseline (b) IAFPN (c) Ours

图 5 可视化比较

4 结论与展望

在本文中开发了一种基于相似度向整个输入图像传播用户提供的稀疏注释的模块, 使用该模块可以在高分辨率下捕获长距离依赖关系, 从而预测更完整的对象. 为了实现使正点击点特征向量高度内聚, 引入了简单的原型提取模块, 缩短训练时间, 加速收敛. 在测试阶段, 引入了感知意图模块, 基于高分辨率下可以更好地捕获细节, 预测的更精准, 使的预测的结果有一定提升, 基于这些改进预测精度有一定的改善. 在几个公共基准上评估了本文的方法可以与最先进的性能相比较. 接下来的研究工作是设计轻量化的模型, 来减少网络推理时间, 满足实时性, 可以部署在低功耗的设备上.

参考文献

1 李国庆. 交互式图像分割方法研究 [博士学位论文]. 武汉: 华中师范大学, 2022.
2 龙建武, 栗童, 朱江洲, 等. 基于超像素和随机游走的交互式分割算法. 计算机应用研究, 2022, 39(6): 1891–1896.
3 Xu N, Price B, Cohen S, *et al*. Deep interactive object selection. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 373–381.
4 Sofiiuk K, Petrov I, Barinova O, *et al*. F-BRS: Rethinking backpropagating refinement for interactive segmentation. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 8620–8629.
5 Rother C, Kolmogorov V, Blake A. "GrabCut": Interactive foreground extraction using iterated graph cuts. ACM Transactions on Graphics, 2004, 23(3): 309–314. [doi: 10.1145/1015706.1015720]
6 Zhang SY, Liew JH, Wei YC, *et al*. Interactive object segmentation with inside-outside guidance. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 12231–12241.
7 Wu JJ, Zhao YB, Zhu JY, *et al*. MILCut: A sweeping line multiple instance learning paradigm for interactive image segmentation. Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 256–263.
8 Bai JJ, Wu XD. Error-tolerant scribbles based interactive image segmentation. Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 392–399.
9 Grady L. Random walks for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(11): 1768–1783. [doi: 10.1109/TPAMI.2006.233]
10 Agustsson E, Uijlings JR, Ferrari V. Interactive full image segmentation by considering all regions jointly. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 11614–11623.
11 Gulshan V, Rother C, Criminisi A, *et al*. Geodesic star convexity for interactive image segmentation. Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco: IEEE, 2010. 3129–3136.
12 Chen X, Zhao ZY, Yu FW, *et al*. Conditional diffusion for interactive segmentation. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal: IEEE, 2021. 7325–7334.
13 Sofiiuk K, Petrov IA, Konushin A. Reviving iterative training with mask guidance for interactive segmentation. Proceedings of the 2022 IEEE International Conference on Image Processing. Bordeaux: IEEE, 2021. 3141–3145.
14 Lin Z, Zhang Z, Chen LZ, *et al*. Interactive image segmentation with first click attention. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 13336–13345.
15 Benenson R, Popov S, Ferrari V. Large-scale interactive object segmentation with human annotators. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 11692–11701.
16 Hao YY, Liu Y, Wu ZW, *et al*. EdgeFlow: Achieving practical interactive segmentation with edge-guided flow.

Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Montreal: IEEE, 2021. 1551–1560.
17 Castrejón L, Kundu K, Urtasun R, *et al*. Annotating object instances with a polygon-RNN. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 4485–4493.
18 Le H, Mai L, Price B, *et al*. Interactive boundary prediction for object selection. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 20–36.
19 Acuna D, Ling H, Kar A, *et al*. Efficient interactive annotation of segmentation datasets with polygon-RNN++. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 859–868.
20 Ling H, Gao J, Kar A, *et al*. Fast interactive object annotation with curve-GCN. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 5252–5261.
21 王涛. 特征度量与信息传递的交互式图论分割方法研究 [博士学位论文]. 南京: 南京理工大学, 2017.
22 罗灵鲲. 迁移学习技术及交互式图像分割相关问题的研究 [博士学位论文]. 上海: 上海交通大学, 2018.
23 Jang WD, Kim CS. Interactive image segmentation via backpropagating refinement scheme. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 5292–5301.
24 Liew JH, Wei YC, Xiong W, *et al*. Regional interactive image segmentation networks. Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017. 2746–2754.
25 Majumder S, Yao A. Content-aware multi-level guidance for interactive instance segmentation. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 11594–11603.
26 Wang XL, Girshick R, Gupta A, *et al*. Non-local neural networks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7794–7803.
27 Zhang CY, Hu CY, Liu YF, *et al*. Intention-aware feature propagation network for interactive segmentation. arXiv:2203.05145, 2022.
28 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778.
29 Lin Z, Duan ZP, Zhang Z, *et al*. FocusCut: Diving into a focus view in interactive segmentation. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022. 2627–2636.
30 Chen X, Zhao ZY, Zhang YL, *et al*. FocalClick: Towards practical interactive image segmentation. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022. 1290–1299.
31 Liu Q, Xu ZL, Bertasius G, *et al*. SimpleClick: Interactive image segmentation with simple vision transformers. arXiv:2210.11006, 2022.
32 Yan CL, Wang HC, Liu J, *et al*. PiClick: Picking the desired mask in click-based interactive segmentation. arXiv:2304.11609, 2023.
33 Chen LC, Zhu YK, Papandreou G, *et al*. Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 833–851.
34 Gori M, Monfardini G, Scarselli F. A new model for learning in graph domains. Proceedings of the 2005 IEEE International Joint Conference on Neural Networks. Montreal: IEEE, 2005. 729–734.
35 Scarselli F, Gori M, Tsoi AC, *et al*. The graph neural network model. IEEE Transactions on Neural Networks, 2009, 20(1): 61–80. [doi: 10.1109/TNN.2008.2005605]
36 Zhang XL, Wei YC, Yang Y, *et al*. SG-One: Similarity guidance network for one-shot semantic segmentation. IEEE Transactions on Cybernetics, 2020, 50(9): 3855–3865. [doi: 10.1109/TCYB.2020.2992433]
37 McGuinness K, O'Connor NE. A comparative evaluation of interactive segmentation algorithms. Pattern Recognition, 2010, 43(2): 434–444. [doi: 10.1016/j.patcog.2009.03.008]
38 Perazzi F, Pont-Tuset J, McWilliams B, *et al*. A benchmark dataset and evaluation methodology for video object segmentation. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 724–732.
39 Hariharan B, Arbeláez P, Bourdev L, *et al*. Semantic contours from inverse detectors. Proceedings of the 2011 International Conference on Computer Vision. Barcelona: IEEE, 2011. 991–998.
40 Kontogianni T, Gygli M, Uijlings J, *et al*. Continuous adaptation for interactive object segmentation by learning from corrections. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 579–596.
41 Wei QQ, Zhang H, Yong JH. Focused and collaborative feedback integration for interactive image segmentation. arXiv:2303.11880, 2023.
42 张华悦, 张顺利, 张利. 基于双阶段网络的交互式目标分割算法. 计算机工程, 2021, 47(2): 300–306. [doi: 10.19678/j.issn.1000-3428.0057071]

(校对责编: 牛欣悦)