计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(7):261−271 [doi: 10.15888/j.cnki.csa.009828] [CSTR: 32024.14.csa.009828]

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

双金字塔式编码的人像语义感知自动抠图网络①

程　艳[1,4], 严志航[2,3,4]

[1](江西师范大学 软件学院, 南昌 330022)
[2](江西师范大学 数字产业学院, 上饶 334000)
[3](东华理工大学 网络与信息中心, 南昌 330013)
[4](江西省智能信息处理与情感计算重点实验室, 南昌 330022)
通信作者: 严志航, E-mail: yanzh@ecut.edu.cn

摘　要: 人像抠图是图像处理领域重要任务之一, 针对现有图像数据中人像前景尺度多样造成的人像抠取粗糙问题, 提出一种双金字塔式编码的人像语义感知自动抠图网络. 双金字塔式编码器包含输入金字塔和特征金字塔, 输入金字塔中输入图像等比例下采样后送入网络以保留原始图像细节, 特征金字塔结合带状卷积组和 5 个层级的编码块充分捕获不同层次的图像特征. 同时, 在双分支解码结构中, 全局分割解码分支上设计了视域扩张模块扩大网络感受范围, 进一步增强全局上下文信息的捕获; 局部细节分支上提出细节感知模块融合编码特征与解码输出, 引导网络关注人像轮廓. 在 3 个数据集上与 6 种人像自动抠图方法进行了对比实验, 所提方法的抠图性能均优于对比方法. 验证了所提方法能改善人像抠取的精细度, 提高了复杂图像数据下人像抠取的鲁棒性.

关键词: 人像抠图; 语义感知; 双金字塔; 双分支解码; 全局上下文信息

引用格式: 程艳,严志航.双金字塔式编码的人像语义感知自动抠图网络.计算机系统应用,2025,34(7):261–271. http://www.c-s-a.org.cn/1003-3254/9828.html

Dual-pyramid Encoded Portrait Semantic-aware Automatic Matting Network

CHENG Yan[1,4], YAN Zhi-Hang[2,3,4]

[1](School of Software, Jiangxi Normal University, Nanchang 330022, China)
[2](School of Digital Industry, Jiangxi Normal University, Shangrao 334000, China)
[3](Information & Network Center, East China University of Technology, Nanchang 330013, China)
[4](Key Laboratory of Intelligent Information Processing and Affective Computing of Jiangxi Province, Nanchang 330022, China)

Abstract: Portrait matting is a significant task in the field of image processing. To address the issue of rough portrait extraction caused by the diverse scales of human figures in existing image data, this study proposes a dual-pyramid encoded portrait semantic-aware automatic matting network. The dual-pyramid encoder consists of an input pyramid and a feature pyramid. In the input pyramid, the input image is proportionally downsampled and fed into the network to preserve the original image details. The feature pyramid combines banded convolution groups and five levels of encoding blocks to fully capture image features at different levels. Meanwhile, in the dual-branch decoder structure, a field expansion module is designed in the global segmentation decoding branch to expand the network's receptive field, further enhancing its ability to capture global contextual information. In the local detail branch, a detail-aware module is proposed to fuse encoded features with decoder output, guiding the network to focus on portrait contours. A comparative analysis is conducted to evaluate the performance of six automatic portrait matting methods on three datasets. The results demonstrate that the proposed method exhibits superior matting performance compared to the other methods. This

① 基金项目: 国家自然科学基金 (62167006); 国家社会科学基金重点项目 (20AXW009); 江西省科技创新基地计划 (2024SSY03131); 江西省主要学科学术和技术带头人培养计划-领军人才项目 (20213BCJL22047); 江西省自然科学基金 (20212BAB202017)
收稿时间: 2024-09-09; 修改时间: 2024-09-30, 2024-10-10, 2024-11-01; 采用时间: 2024-11-08; csa 在线出版时间: 2025-05-29
CNKI 网络首发时间: 2025-05-29

validates the effectiveness of the proposed method in enhancing the precision and robustness of portrait extraction in complex image data.

Key words: portrait matting; semantic-aware; dual-pyramid; dual-branch decoder; global contextual information

1 引言

人像抠图旨在从输入图像或视频帧中预测一个透明度遮罩用以提取人像前景, 在图像编辑[1]、广告制作、电影创造、行业直播[2]等领域有着十分广泛的应用. 相较于其他分割任务, 人像抠图提取的结果更为细腻自然, 是图像处理领域基础且极具挑战的视觉任务.

早先大多数传统的抠图方法需要一个标注好的三元图作为辅助指导输入, 它明确地定义了前景和背景的区域以及抠图方法要求解的未知部分. 传统的方法大致分为基于采样和基于传播两种[3]. 基于采样的方法通过在确定的前景和背景区域内采样像素来建立颜色统计量, 以估计未知部分区域内的透明度值. 而基于传播的方法, 又称基于亲和力的方法, 通过将前景和背景像素的透明度值传播到未知部分以估计该区域的透明度值. 然而, 不论是基于传播还是基于采样的方式, 都依赖于图像中的低级特征信息, 难以处理当今特征分布规律复杂的人像数据.

近年来, 深度学习方法在目标检测、语义分割等领域取得了突破性进展, 基于深度学习的方法在图像抠图领域也被证明是强大的, 但大多数早期的工作仍需要将三元图作为额外的输入. Xu 等人[4]提出的 DIM 较早地将原始图像和三元图连接起来作为网络输入, 并创建了一个大规模抠图数据集, 将带注释的抠图合成到各种背景图像中, 并在该数据集上进行模型训练, 在当时取得了最先进的抠图性能. Luts 等人[5]同样采用三元图作为额外辅助输入, 提出了第 1 个用于自然图像抠图的生成对抗网络 AlphaGAN[6], 通过对抗性训练过程, 生成器学习生成在视觉上与真实标签类似的透明度遮罩, 鉴别器学习更好地区分真实透明度遮罩和生成透明度遮罩, 进一步改善了抠图结果. Sun 等人[7]使用基于图像块状的分类器, 将传统三元图扩展为语义三元图, 对抠图区域进行语义分类, 再将语义三元图和原始 RGB 图像一起作为输入送入网络中进行透明度遮罩预测. Park 等人[8]提出利用 swin Transformer[9]模型结合三元图和先验知识令牌信息. Cai 等人[10]同样基于 swin Transformer 模型并加入了三元令牌的概念, 提高了透明度遮罩预测精度. 尽管添加注释三元图使抠图问题变得更容易处理, 但它对用户来说可能是相当繁重的, 限制了这些方法在许多非交互式应用程序中的可用性.

对此, 研究人员开始思考无三元图输入下的抠图问题. 一类方向是背景输入替代三元图引导的方案, 即用户拍完照片后再拍摄一张无人的背景图. Sengupta 等人[11]利用易获取的背景图和原始图像作为输入, 通过一个编码器-解码器结构得到透明度遮罩以及前景, 并紧接着生成对抗网络对在不同背景生成前景图片的真实性进行自监督的判别. Lin 等人[12]同样基于背景图片提出一个两段的抠图网络, 基础网络段用以生成低分辨率的结果, 微调网络段用以在选中的图块上生成高分辨率的结果, 提升了模型抠取人像细节的能力. 但基于背景抠图的方法在动态的环境中并不适用, 在不能保证背景输入与实际背景一致时很难取得良好的效果. 另一类方向是摆脱任何外部引导, Chen 等人[1]创造性地用两阶段深度学习网络挖掘语义信息而不用输入三元图, 首先生成伪三元图, 然后用作透明度遮罩预测阶段的先验知识. 但分阶段的模型训练容易带来错误的语义信息, 为此 Ke 等人[13]通过任务目标分解和显式监督下的并行优化, 可以实时预测出高质量的透明度遮罩. Li 等人[14]采用一个编码器和两个独立的解码器, 以协作的方式学习不同的任务, 从而实现端到端的自然图像抠图, 在此基础上, Li 等人[15]设计了三重特征融合、浅层双重特征融合和深层双重特征融合, 以对共享编码器和两个解码器之间的交互关系进行进一步建模. 虽然以上方法较好地解决了无辅助输入的图像抠图问题并取得较好的抠图性能, 但随着图像数据多样性增长, 人像前景各异尺度多样, 上述方法不能很好地适应人像尺度的复杂性、多样性, 致使抠图不够精细.

为解决现有图像数据中人像前景尺度的多样性带来的抠图粗糙问题, 本文设计了一种无需辅助输入的双金字塔式编码的人像语义感知自动抠图网络 (dual-pyramid encoded portrait semantic-aware automatic matting network, DPSAM), 用于实现高精度的人像抠

图. 其主要贡献如下.

(1) 提出一种双金字塔编码网络 (dual-pyramid architecture network, DPANet). DPANet 依次对原始输入图像进行下采样操作, 生成不同尺度的输入图像以构成输入金字塔; DPANet 中设计了带状卷积组 (ribbon convolution group, RCG) 以捕获不同尺度的上下文信息, 并结合修改后的 ResNet-34 来提取多层级特征, 构成特征金字塔; 输入金字塔提供原始图像的局部细节, 特征金字塔从每一层的下采样输入中提取复杂的语义特征, 形成一系列互补特征.

(2) 设计了视域扩张模块 (perceptual expansion module, PEM) 增强高层特征的捕获. 该模块引入 4 个并行空洞卷积分支, 并通过级联较低扩张率的空洞卷积以达到更广的感受野, 且另一个分支使用全局平均池化, 以提取丰富的上下文语义信息.

(3) 设计了细节感知单元 (detail-aware unit, DU) 引导优化分割细节. 该模块引入通道注意力机制以调整编码层的输出的通道权重, 接着与来自解码层的输出进行交互, 送入 *Sigmoid* 激活及后续加乘操作, 得到最终注意力特征.

2 方法

2.1 问题定义

在抠图任务中, 给定图像 $I \in R^{H\times W\times C}$, I 是由前景 $F \in R^{H\times W\times C}$ 与背景 $B \in R^{H\times W\times C}$ 构成. 假定图像中第 i 个像素的颜色通过前景与背景颜色之间的线性组合方程表示为:

$$\begin{cases} I_i^r = \alpha_i F_i^r + (1-\alpha_i)B_i^r \\ I_i^g = \alpha_i F_i^g + (1-\alpha_i)B_i^g \\ I_i^b = \alpha_i F_i^b + (1-\alpha_i)B_i^b \end{cases} \tag{1}$$

其中, I_i 代表图像在 i 位置的像素值, 而 F_i 和 B_i 分别表示位置 i 处的前景像素值和背景像素值, 右上角标 r、g、b 分别对应图像的红、绿、蓝通道. 数学上, α_i 表示 F_i 在 I_i 中所占的权重, 而在涉及图像抠图任务时, α_i 则反映了前景在位置 i 的透明度, 透明度取值范围为[0, 1]. 该图像 I_i 所有像素的透明度值构成了一张前景掩码, 本文方法所提取的黑白人像轮廓即为该图像所有像素的透明度值, 构成了一张前景掩码即为透明度遮罩 α. 式 (1) 需要求解 7 个未知量, 但在给定的 RGB 图像中, 每个像素已知的值只有 3 个. 因此, 通常需要使用先验知识或辅助输入, 为抠图问题求解增加限制条件.

2.2 双金字塔式编码的人像语义感知自动抠图网络

本文提出了一种双金字塔式编码的人像语义感知自动抠图网络 (DPSAM), 网络结构如图 1 所示, 包含双金字塔编码网络与双分支解码网络两部分. 输入端仅需一张 RGB 图像, 无需额外辅助信息输入; 双金字塔编码网络包括输入金字塔、融合带状卷积组 (RCG) 的特征金字塔; 双分支解码网络包括全局分割分支和细节提取分支, 并在其中设置了视域扩张模块 (PEM) 与注意力导引模块 (AGM).

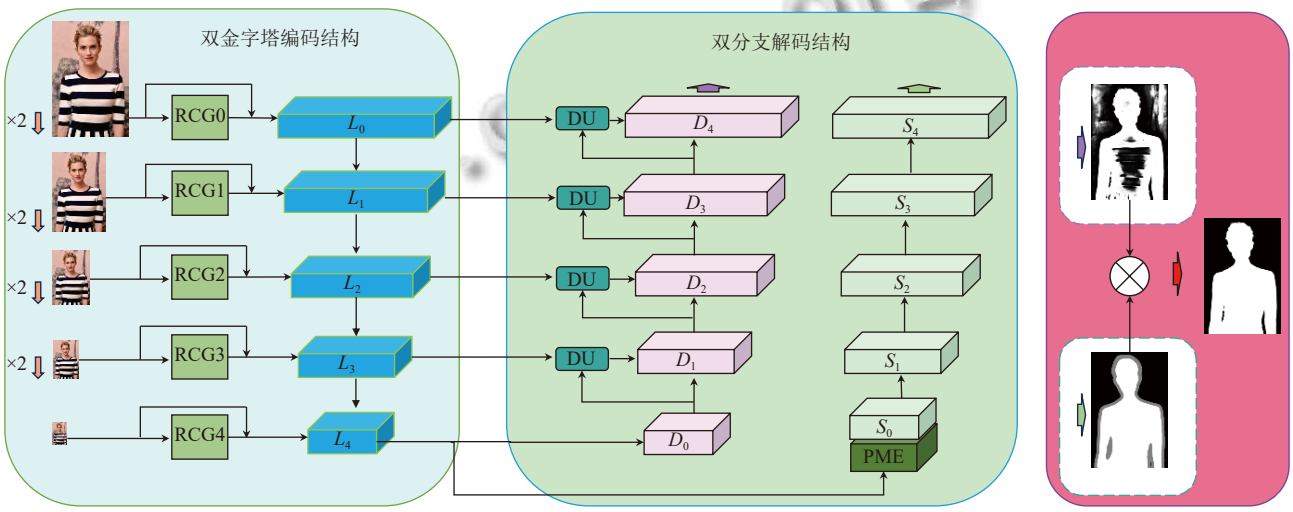


图 1 DPSAM 网络结构图

为捕捉物体的细节特征和感知物体尺度的变化, 本文设计一种双金字塔式的编码网络 DPANet. 其中, 该编码结构包括输入金字塔、融入带状卷积组的特征金字塔. 首先, 原始人像输入通过多次最大池化操作形

成的输入金字塔保留了原始图像的细节特征, 带状卷积组从每一层的输入金字塔中提取不同尺度特征表示, 再将不同尺度的特征表示送入不同层级的编码块, 进一步提取不同层的特征图形成特征金字塔, 结合输入金字塔和特征金字塔, 通过提取人像图像的详细特征和高级特征, 使其形成互补. 视域扩张模块设置在编码部分末端, 进一步扩大全局感受野, 将提取的上下文信息与全局解码分割分支连接. 同时, 细节感知单元指引网络更关注人像轮廓, 帮助网络更好地理解图像中的细节特征. 第 2.2.1 和第 2.2.2 节将详细介绍网络结构.

2.2.1 双金字塔式编码结构

(1) 输入金字塔

不同的人像图像数据中, 人像前景有不同的尺度. 不同的人像图像在人体比例、人体形状和人体位置上都有明显的区别, 为适应人像前景尺度多样性的问题, 设计输入金字塔模块来捕捉人像的细节特征, 本文首先对原始输入图像 $I \in R^{d\times H\times W}$ 进行最大池化下采样操作, 形成输入金字塔 ($I_0 \in R^{d\times H\times W}$, $I_1 \in R^{d\times H/2\times W/2}$,⋯, $I_4 \in R^{d\times H/16\times W/16}$), 它保留了原始图像的细节特征; 然后, 将缩放后的图像送入对应的带状卷积组 (RCG0, RCG1,⋯, RCG4).

(2) 特征金字塔

1) 带状卷积组. 相较于常规方形卷积核, 带状卷积凭借长核形状及多方向特性捕获能力在面对人体类长条状对象时展现出显著优势. 带状卷积组的输入是不同规格的原始输入图像, 与标准卷积方式不同, 带状卷积组以并行的 $1\times k$ 的行向和 $k\times 1$ 的列向带状卷积的组合方式模拟 $k\times k$ 的卷积核, 结构如图 2 所示. 本文选择带状卷积有两方面原因. 一方面, 带状卷积是轻量级的, 标准卷积的参数量是 k^2, 而带状卷积块的参数量为 $4k$. 另一方面, 抠图场景中人体是带状的, 因此, 带状卷积可以作为标准卷积的补充, 有助于提取类带状特征[16].

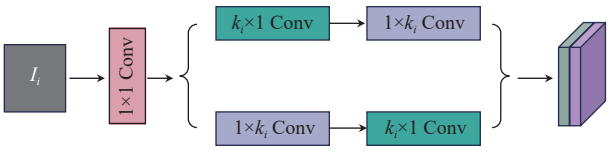


图 2 RCG 结构图

具体地说, 对于输入 $I_i \in R^{d\times H\times W}$ 在带状卷积组其中一个并行分支中, 本文先以列的方向与 $k_i\times 1$ 的内核进行卷积, 然后以行的方向与 $1\times k_i$ 的内核进行卷积; 在另外一个分支中, 则是先以行的方向与 $1\times k_i$ 的内核进行卷积再以列的方向与 $k_i\times 1$ 的内核进行卷积; 然后将两个分支的结果相加得到每个带状卷积块包含丰富上下文信息的特征图. 以上过程用公式表述为:

$$F_M^i = f^{k_i\times 1}(f^{1\times k_i}(I_i)) + f^{1\times k_i}(f^{k_i\times 1}(I_i)) \tag{2}$$

其中, $i\in\{0, 1, 2, 3, 4\}$ 表示 5 种尺度的图像输入、5 种不同内核大小的带状卷积块以及在该大小下的特征输出的索引号; $f^{k_i\times 1}(\cdot)$ 和 $f^{1\times k_i}(\cdot)$ 分别表示内核大小为 $k_i\times 1$ 和 $1\times k_i$ 的带状卷积.

2) 特征编码块. ResNet[17]结构的创新解决了深度网络的梯度消失和退化问题, ResNet-34 是 ResNet 系列模型的其中一个代表, 其网络结构主要由卷积层、批量归一化 (batch normalization, BN) 层和残差块组成. 其中, 残差块是 ResNet 系列模型的核心组件, 其包含了两个卷积层和一个跨层残差连接. 这个跨层残差连接允许信息在不同层之间直接流动, 从而避免了信息的丢失. 在 ResNet-34 中, 残差块的数量较少, 使得模型具有较好的计算效率和泛化能力.

本文选择 ResNet-34 后并对其进行修改作为轻量级骨干网络. 修改后的 ResNet-34 在原始模型的 5 个卷积组进行了改动, 将每个卷积组的第 1 个卷积层的卷积步幅由 2 改为 1, 然后相应的增加最大池化层, 并以此形成 5 个层级的编码块 (L_0, L_1,⋯, L_4). 其中, L_{i-1} 层级编码块的输出 E_{i-1} 和同一层级的带状卷积组的输出 F_M^i 进行拼接, 作为 L_{i-1} 层级编码块的输出, 每一层的编码块提取不同尺度特征表示, 最终生成 5 个层级的特征金字塔 (E_0, E_1,⋯, E_4), 这 5 个层级的特征将与细节分支进一步交互建模. 表 1 为修改后的网络结构.

表 1 5 个层级编码块的网络结构

编码块	输出大小	结构
L_0	256×256	7×7, 64, *stride*=1 最大池化层
L_1	128×128	[3×3, 64, *stride*=1 3×3, 64, *stride*=1] ×2 最大池化层
L_2	64×64	[3×3, 64, *stride*=1 3×3, 64, *stride*=1] ×2 最大池化层
L_3	32×32	[3×3, 64, *stride*=1 3×3, 64, *stride*=1] ×2 最大池化层
L_4	16×16	[3×3, 64, *stride*=1 3×3, 64, *stride*=1] ×2 最大池化层

2.2.2 双分支解码结构

缺失了三元图等先验知识, 以往的抠图方法很难获得正确的语义上下文. 而后续的双分支解码部分消除了这个问题, 全局分割分支用于全局上下文抽取, 细节提取分支用于提取未知区域细节表示, 然后将其与全局分割中的语义图融合, 生成最终的透明度遮罩.

双分支解码结构中包括视域扩张模块、细节感知单元、全局分割分支解码块 (S_0, S_1,⋯, S_4) 以及细节提取分支解码块 (D_0, D_1,⋯, D_4). 其中, 解码块都由若干个 3×3 卷积层、批归一化层、*ReLU* 激活函数和上采样层堆叠而成.

(1) 视域扩张模块

随着网络的加深, 高层的语义信息对于提高分割性能起着至关重要的作用. 受空洞空间金字塔池化模块[18]启发, 本文在 L_4 层后设置了视域扩张模块 (PEM), 结构如图 3 所示.

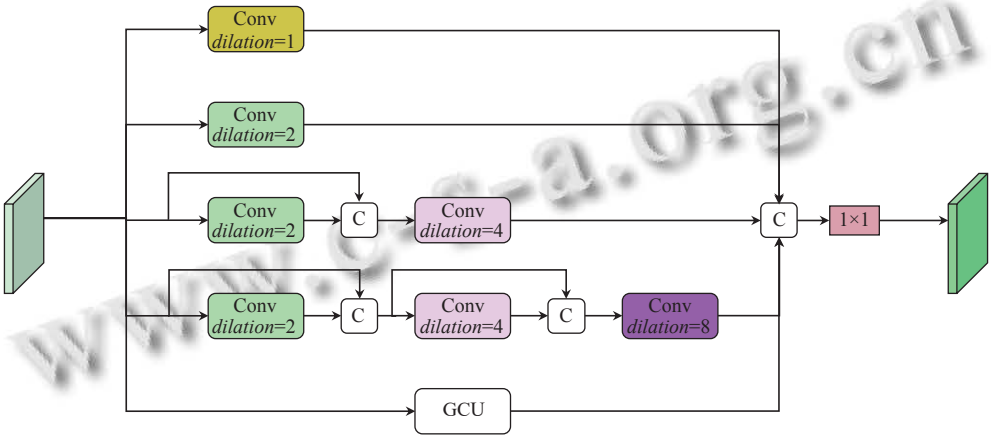


图 3 PEM 结构图

PEM 由两部分组成, 其一是 4 个并行的扩张卷积, 在第 1、2 分支上分别设置扩张率为 1、2 的 3×3 卷积 $DConv_1$、$DConv_2$, 前者相当于普通卷积. 由于随着扩张率增加时, 扩张卷积稀疏的采样输入方式会使提取的信息片段之间的相关性缺失, 同时有效卷积核参数会逐渐减少[19]. 因此, 本文主张在 PEM 中使用低扩张率卷积,并注重初始特征注入. 在第 3 分支上级联扩张率最高为 4 的扩张卷积, 依次为扩张率为 2、4 的 3×3 卷积 $DConv_2$、$DConv_4$, 并在两个扩张卷积中设置残差连接以保存原始输入信息. 在第 4 个分支上级联扩张率最高为 8 的扩张卷积, 依次扩张率为 2、4、8 的 3×3 卷积 $DConv_2$、$DConv_4$、$DConv_8$, 同样在相邻两个扩张卷积间设置了残差连接. 另一分支为全局平均池化分支, 该分支由全局平均池化、1×1 卷积、上采样层组成. 而后将 5 个分支的输出拼接后再次经过 1×1 卷积调整, 最终输出. 计算过程表示见式 (3)、(4).

$$\begin{cases} O_1 = DConv_1(E_4) \\ O_2 = DConv_2(E_4) \\ O_3 = DConv_4(Concat(DConv_2(E_4), E4)) \\ O_4 = DConv_8(Concat(DConv_4(Concat(DConv_2(E_4), \\ \quad E4)), Concat(DConv_2(E_4), E4)))) \\ GCU_i(E_4),\ i = 5 \end{cases} \tag{3}$$

$$O = Concat(O_1, O_2, O_3, O_4, O_5) \tag{4}$$

其中, E_4 为最高层编码输入, O 为 PEM 输出, $DConv_i$ 表示 4 个分支上不同的扩张卷积设置, *GCU* 表示为全局平均池化、卷积及上采样操作, $Concat(\cdot)$ 为维度上的拼接操作.

(2) 细节感知单元

为抑制干扰噪声以更好地从浅层特征中提取人像轮廓细节, 在编码分支、细节提取分支间提出了细节感知单元 (DU), 结构如图 4 所示. DU 融合了通道注意力模块以提高对特征输入的深层感知. 细节感知单元有两条路径输入, 一条为经过通道注意力计算的编码特征输入 E_i, 一条为经过 CNR (卷积、归一化、*ReLU*) 的解码特征输入 D_i, 其中 $i \in \{1, 2, 3, 4\}$.

通道注意力模块, CA 结构如图 5 所示. 首先对编码特征输入 E_i 分别进行全局最大池化 (global max pooling, GMP) 和全局平均池化 (global average pooling, GAP), 得到两个 1×1×C 的特征图; 接着将两个 1×1×C 的特征图送入包含池化、全连接层、激活函数的多层感知机 (multilayer perceptron, MLP) 之中; 最后将 MLP 的输出结果相加, 经过 *Sigmoid* 函数处理, 最终得到通道注意力权重值, 赋予各通道不同值. 计算过程表

示如下:

$$E_{i'} = \sigma(MLP(GAP(E_i)) + MLP(GMP(E_i))) \quad (5)$$

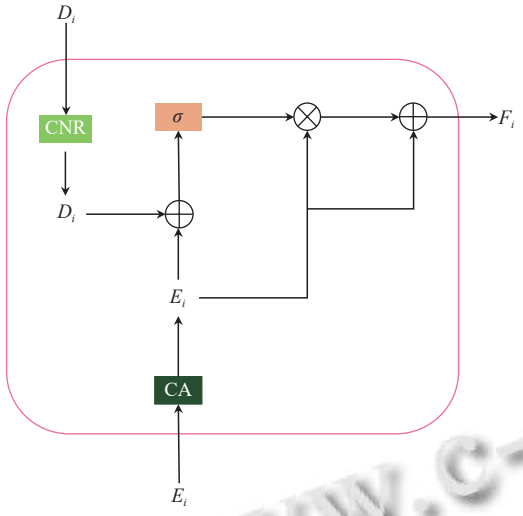


图 4 DU 结构图

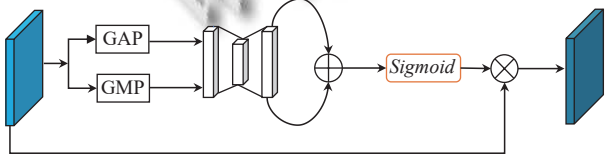


图 5 通道注意力 (CA) 结构图

在另外一条路径上, 解码层的特征输入 D_i 先送入 CNR 模块, 经过归一化、1×1 卷积、非线性激活函数 *ReLU* 增强特征表示并调整通道大小, 利用跳跃连接融合输入特征 D_i 后经过上采样操作使其大小与 E_i 一致, 得到 $D_{i'}$. 接着与另一路径中经过 CA 注意力加权的 $E_{i'}$ 相加, 送入 *Sigmoid* 函数激活得到感知注意力图 M, 以感知人像边缘细节, 最后将 $E_{i'}$与注意力图 M 相乘生成最终注意力特征, 与 $E_{i'}$相加后而输出 F_i. 计算过程为:

$$D_{i'} = Norm(Conv(ReLU(Conv(Norm(D_i))))) + D_i \quad (6)$$

$$F_i = (E_{i'} \otimes \sigma(D_{i'} \oplus E_{i'})) \oplus E_{i'} \quad (7)$$

其中, F_i 表示细节感知单元的最终输出, i 为不同层感知单元的索引; $Norm(\cdot)$ 表示归一化操作; $Conv(\cdot)$ 为 1×1 卷积; $ReLU(\cdot)$ 为非线性激活, 结构如图 6 所示.

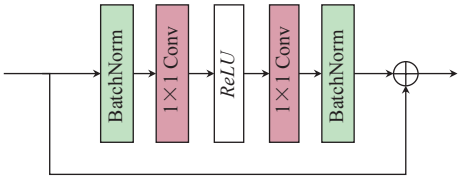


图 6 CNR 结构图

2.3 损失函数

本文网络的损失计算综合全局分割分支、局部细节提取分支以及最后融合部分的总和 L_{total}, 由于本文的网络架构与文献[15]的网络架构都为双分支解码结构, 本文在损失函数的选取参考了文献[15], 并以此进行端到端的训练:

$$L_{\text{total}} = \lambda_1 L_S + \lambda_2 L_D + \lambda_3 L_F \quad (8)$$

其中, $\lambda_1, \lambda_2, \lambda_3$ 为各分支的损失权重, 权重设置参照文献[15].

全局分割分支的任务实质是一个多分类任务, 即预测图像的前景、背景和未知区域, 为更好地监督模型学习有效分类决策边界, 该分支损失 L_S 采用交叉熵损失函数惩罚分割分支预测差异, 其定义如下:

$$L_S = -\sum_{c=1}^{3} S_g^c \log(S_p^c) \quad (9)$$

其中, $S_g^c \in \{0, 1\}$表示像素点的真实标签, $S_p^c \in [0, 1]$表示像素点是第 c 类的预测概率值.

局部细节分支重点关注的是人像轮廓的细节, 该分支的训练损失由L_α^{T}和$L_{\text{lap}}^{\text{T}}$两部分组成[20]. L_α^{T}定义为未知部分真值标签α_i与预测透明度遮罩α_i^F的绝对差值计算, $L_{\text{lap}}^{\text{T}}$定义为计算真值标签的拉普拉斯金字塔与预测值的拉普拉斯金字塔在 5 个尺度上的 $L1$ 距离, 并乘以相应的权重, 即:

$$L_\alpha^{\text{T}} = \frac{\sum_i \sqrt{((\alpha_i - \alpha_i^F) \times W_i^{\text{T}})^2 + \varepsilon^2}}{\sum_i W_i^{\text{T}}} \quad (10)$$

$$L_{\text{lap}}^{\text{T}} = \sum_i W_i^{\text{T}} \sum_{i=1}^{5} \left\| Lap^k(\alpha_i) - Lap^k(\alpha_i^F) \right\|_1 \quad (11)$$

$$L_D = L_\alpha^{\text{T}} + L_{\text{lap}}^{\text{T}} \quad (12)$$

其中, i 表示像素索引数, $W_i^{\text{T}} \in \{0, 1\}$表示像素 i 是否属于未知部分, ε是一个小的正数, $Lap^k(\alpha_i)$表示真值的第 k 层拉普拉斯金字塔, $Lap^k(\alpha_i^F)$表示预测值的第 k 层拉普拉斯金字塔.

融合部分的训练损失由预测损失L_α^{T}、未知部分的损失$L_{\text{lap}}^{\text{T}}$和融合损失$L_{\text{OT}}$组成[15], L_{OT}定义为 α 真值标签和预测的α_i^{ot}融合图像的绝对差值计算, 其数学表达式如下:

$$L_F = L_\alpha^{\text{T}} + L_{\text{lap}}^{\text{T}} + L_{\text{OT}} \quad (13)$$

$$L_{\mathrm{OT}}=\frac{\sum_i\sqrt{(\alpha_i-\alpha_i^{\mathrm{ot}})^2+\varepsilon^2}}{N} \tag{14}$$

其中, N 为图像中的像素总数.

3 实验与分析

3.1 实验环境与训练细节

实验设备的操作系统为 Ubuntu 20.04, 显卡为 NVIDIA RTX 3090, 24 GB. CPU 为 Intel i7-11700, 实验在基于 PyTorch 的深度学习框架下进行模型训练和测试, 采用的 Python 3.6 版本. 本文采用了 Adam 优化器来对整个网络的权重进行优化, 学习率为 1E–4, 并将训练批次大小设置为 8, 训练 180 个 epoch 后模型基本收敛, 各分支损失不再下降, 一次训练总计耗时达 64 h 左右, 损失收敛曲线如图 7 所示. 为了进一步增加数据的多样性, 本文还采用了随机旋转、剪裁等数据处理方法, 以提升网络的鲁棒性.

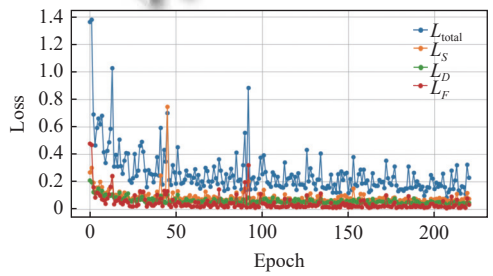


图 7 训练损失

3.2 实验数据集

P3M-10k[15]: 为包含 10 421 张高质量各种背景各种姿态下的人体图像的大型数据集, 为专业人员精心注释. 其中还包含具有 500 幅人脸模糊图像的测试集 P3M-500-P 和 500 幅人脸清晰图像的测试集 P3M-500-NP. 选取 P3M-10k 的 9 421 张图像作为训练集, 以 P3M-500-P 和 P3M-500-NP 作为测试数据集.

PPM-100: 包含精细注释的 100 张具有多样化背景的人像图像测试数据集 PPM-100[13], 标注中设定将前景人物手持的小物件视为前景的一部分, 更加贴近实际应用场景. PPM-100 中的数据拥有更加自然的背景以及更为丰富的姿态变化.

3.3 评价指标

本文采用绝对误差值之和 (sum of absolute difference, *SAD*)、均方误差 (mean square error, *MSE*)、平均绝对差值 (mean absolute difference, *MAD*)、梯度误差 (gradient, *Grad*) 和连通度误差 (connectivity, *Conn*) 作为人像抠图的评价指标, 计算方法如下所示:

$$SAD=\sum_{i=1}^{N}\left|\alpha_i-\widehat{\alpha}_i\right| \tag{15}$$

$$MSE=\frac{1}{N}\sum_{i=1}^{N}(\alpha_i-\widehat{\alpha}_i)^2 \tag{16}$$

$$MAD=\frac{1}{N}\sum_{i=1}^{N}\left|\alpha_i-\widehat{\alpha}_i\right| \tag{17}$$

$$Grad=\frac{1}{N}\sum_{i=1}^{N}\left\|\nabla\alpha_i-\nabla\widehat{\alpha}_i\right\| \tag{18}$$

$$Conn=\frac{1}{N}\sum_{i=1}^{N}(\varphi(\alpha_i,\Omega)-\varphi(\widehat{\alpha}_i,\Omega)) \tag{19}$$

其中, α_i 是像素 i 的预测值; $\widehat{\alpha}_i$ 是像素 i 的真实值; N 是图像中的像素个数; ∇ 表示梯度计算, 该数值由高斯一阶导计算; Ω 表示真值标签与预测结果中的最大连通区域, $\varphi(\cdot)$ 表示像素 i 与 Ω 间连接性程度计算. 此外, 为更好地进行实验结果对比, 本文将 *MSE*、*MAD* 放大 10^3 倍. 评价指标的值越低, 说明网络抠图性能越好.

3.4 对比实验

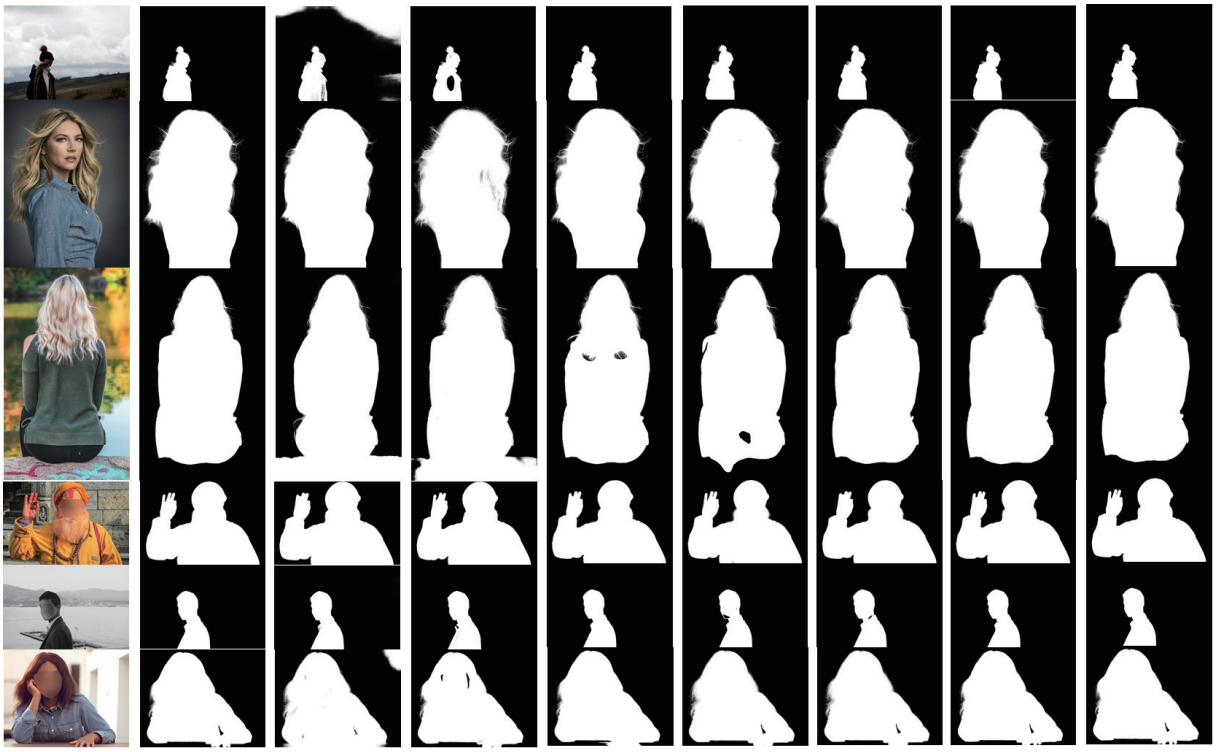
为客观评估本文提出方法的抠图性能, 本文同样选取 6 个自动抠图 (无额外辅助输入) 方法 HATT[21]、SHM[1]、LFM[22]、MODNet[13]、GFM[14]、P3M[15]在 P3M-500-NP、P3M-500-P 和 PPM-100 这 3 个测试数据集上进行实验并与本文方法对比, 实验结果如表 2 所示, 表中加粗数字为最优值, 符号“↓”表示值越低性能越好. 由表 2 可知, 7 个抠图方法都具有良好的指标表现. 在 P3M-500-NP 数据集上, P3M 在 6 个所选取的对比方法中表现最佳, 与 P3M 方法相比, 本文方法所有指标都存在明显降低, 本文方法在所有指标上都优于对比的 6 个自动抠图方法, *SAD* 等 5 个指标分别降低 3.83、1.92、2.15、1.10 和 3.70. 在 P3M-500-P 数据集上, 本文方法同样在所有指标上都有所降低, 均优于其他对比方法, 与 P3M 方法相比, *SAD* 等 5 个指标分别降低 1.32、0.56、0.77、0.25 和 2.20. 在标注更加贴近实际应用场景的 PPM-100 数据集上, 可以观察到所有方法的评价指标都存在较大波动, 但本文方法在所比较方法中 5 个评估指标均为最低, 对复杂的环境相对能够更好地适应, 展现出不错的鲁棒性. 综合来看, 实验结果验证了本文抠图网络结构设计的有效性, 能够更好地提取图像特征, 提高人像抠图的精度.

为直观感受本文人像自动抠图方法的效果, 本文

在 P3M-500-P、P3M-500-NP 数据集上选取了不同的人体自然图像, 对 6 个基准方法进行抠图测试并与本文方法抠图结果进行定性分析, 其中 GT 为图片透明度的真实值 (ground truth), 实验结果如图 8 所示. 图 8 所展示的人像轮廓是本文方法直接获得的人像透明度遮罩, 而人像透明度遮罩实际上是对式 (1) 中 α 的求解, 根据所得透明度遮罩进行后处理操作抠取人像前景. 在第 1 组测试结果中 LFM、SHM 难以准确判别人像前景, HATT、GFM、MODNet 和 P3M 在与帽子交接的发丝处预测模糊, 在镜子细节处的抠取不如本文方法细腻、精确; 在第 2 组测试结果中, 其他方法在额头上沿、头后及胸前发丝处的抠取略微粗糙, 存在漏抠、误抠, 本文方法对发丝的抠取较为细腻; 在第 3 组测试结果中, LFM、SHM 方法错误地抠取了地面, 其他方法要么难以准确抠全人像、要么边缘抠取粗糙, 而本文方法精细地预测出人像及边缘轮廓细节. 在第 4 组测试结果中, 可以清晰地观察到本文方法在分散的手指处抠取效果最好; 在第 5 组测试结果中, 对比方法对人像中部的脖子、下巴存在判别错误的情况; 在第 6 组测试结果中, 其余 6 种方法容易错误地将手指与桌子交界处预测为前景, 而本文方法抠取的更加准确更加精细. 可视化结果表明, 本文所提出的方法提高了人像抠取的精度, 综合抠图效果优于其余方法.

表 2　DPSAM 与其他方法的比较

方法	P3M-500-NP					P3M-500-P					PPM-100				
	SAD↓	*MSE*↓	*MAD*↓	*Grad*↓	*Conn*↓	*SAD*↓	*MSE*↓	*MAD*↓	*Grad*↓	*Conn*↓	*SAD*↓	*MSE*↓	*MAD*↓	*Grad*↓	*Conn*↓
HATT	30.53	9.18	17.63	27.42	19.88	25.97	7.21	15.42	25.29	14.91	123.91	14.45	15.12	73.75	121.38
SHM	23.63	10.61	13.68	15.17	28.52	21.95	9.93	12.76	18.17	27.06	150.07	17.03	14.86	63.21	147.78
LFM	40.71	16.34	23.72	41.36	17.63	31.65	12.76	16.81	30.29	18.74	148.14	24.64	16.39	60.28	145.63
MODNet	15.68	6.14	9.23	13.63	15.29	12.93	4.57	7.54	13.31	12.38	85.96	8.14	12.76	64.26	94.28
GFM	14.98	9.04	8.65	17.57	14.68	13.09	5.12	7.68	17.34	12.63	107.38	18.27	10.65	45.59	107.89
P3M	12.88	4.63	7.42	12.85	12.31	10.29	3.64	5.98	13.69	10.92	97.21	14.23	17.61	49.35	112.92
本文方法	**9.05**	**2.71**	**5.27**	**11.75**	**8.61**	**8.97**	**3.08**	**5.21**	**13.44**	**8.72**	**78.47**	**7.32**	**9.37**	**37.64**	**72.38**



(a) 原图 (b) GT (c) LFM (d) SHM (e) HATT (f) GFM (g) MODNet (h) P3M (i) 本文

图 8　DPSAM 与其他方法的测试结果

3.5 消融实验

为验证 RCG、PEM、DU 这 3 个模块在网络中的有效性, 采用 DPANet 作为编码网络, P3M-500-NP 作为实验数据集, 共设计 4 组对比实验, 结果如表 3 所示. 在表 3 中, 符号“√”表示应用该模块, “×”表示不使用该模块. 可以看出, 在不添加 3 个模块时, 抠图网络的性能最差; 在最高层编码块 L_4 添加 PEM 模块后, 感受野的扩大以及多尺度的池化操作有益于丰富深层语义信息; 在前一设置基础上, 增加 RCG 模块后, 带状卷积一定程度上与标准卷积形成互补, *SAD*、*MSE*、*MAD* 这 3 个指标依次降低 0.83、0.46、0.73, 这表明 RCG 模块的有效性; 在添加所有模块后, *SAD*、*MSE*、*MAD* 这 3 个指标进一步降低 1.47、1.11、1.04, 说明抠图性能进一步得到提升, 这也验证了 DU 模块有益于引导模型感知人像细节轮廓, 提升抠图效果. 综上, 3 个模块有益于模型预测更准确更高质量的透明度人像遮罩.

表 3 各模块消融实验

PEM	RCG	DU	*SAD*↓	*MSE*↓	*MAD*↓
×	×	×	13.92	5.42	8.32
√	×	×	11.35	4.28	7.04
√	√	×	10.52	3.82	6.31
√	√	√	**9.05**	**2.71**	**5.27**

为验证 DPANet 的有效性, 本文在相同训练策略下对本文模型及其变体进行训练, P3M-500-NP 作为实验数据集, 实验结果如表 4 所示. 表 4 列出已去除输入金字塔部分的 DPANet*、ResNet-34、DenseNet-121[23] 以及 DPANet (本文) 作为编码部分的测试结果. 可以看出, 当以 DPANet 作为编码结构时, 网络抠图性能达到最优, DPANet*由于去除了输入金字塔结构, *SAD* 等抠图指标均有不同程度的回升. DPANet 作用在于保留不同规格原始输入的细节特征, 并利用带状卷积和金字塔编码块提取多尺度特征, 提高模型更全面的特征表征能力. DPANet 较其余编码结构取得更低的评价指标值, 验证了该网络更能提取复杂多样图像数据中的人像特征, 预测得到更精细的透明度遮罩.

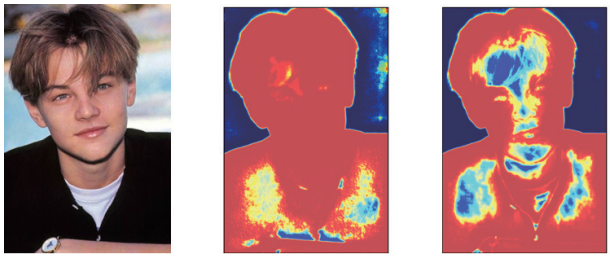
表 4 DPANet 消融实验

编码器	*SAD*↓	*MSE*↓	*MAD*↓
DPANet*	13.12	5.82	8.34
ResNet-34	14.27	5.93	8.21
DenseNet-121	12.19	4.89	7.03
DPANet (本文)	**9.05**	**2.71**	**5.27**

为验证 PEM 设计的有效性, 在 P3M-500-NP 数据集中进行了消融实验, 实验结果见表 5. 第 1 个实验中去除了级联扩张卷积的设置, 即使用单个扩张卷积; 第 2 个实验维持本文设置, 结果如表 5 所示. 当去除级联扩张卷积的设置后, *SAD*、*MSE*、*MAD* 均有不同程度的降低, 依次为 1.19、1.05、0.84; 当维持原有设置时, 抠图性能得到了提升. 实验结果表明 PEM 级联扩张卷积设置的可行性. 同时为进一步验证 PEM 的作用效果, 本文在有 PEM 作用和无 PEM 作用下展示了输出热力图, 如图 9 所示. 可以看到, 当没有 PEM 模块时, 网络容易忽略人像前景, 而当在 PEM 作用下时, 网络对人像语义更为敏感, 对人像前景的识别更为充分. 由此可见, PEM 模块可以有效地感知人像语义上下文, 提高网络抠图精度.

表 5 PEM 消融实验

设置	*SAD*↓	*MSE*↓	*MAD*↓
–级联扩张卷积	10.24	3.76	6.11
+级联扩张卷积	**9.05**	**2.71**	**5.27**



(a) 输入 (b) 有 PEM 输出 (c) 无 PEM 输出

图 9 PEM 消融实验

为验证 DU 设计的有效性, 在 P3M-500-NP 数据集中对通道注意力模块进行了消融实验, 实验结果见表 6. 第 1 个实验中去除了通道注意力计算, 第 2 个实验增加了 SE[24]注意力模块, 第 3 个实验增加了 CBAM[25] 注意力模块, 第 4 个实验增加了 CA 通道注意力模块, 结果如表 6 所示. 当去通道注意力模块的设置后, *SAD*、*MSE*、*MAD* 均有不同程度的降低; 而当分别添加 3 种注意力模块时, 添加 CA 注意力模块时, 抠图性能达到最佳. 实验结果表明通道注意力设置的必要性, 验证了 DU 设计的有效性.

表 6 DU 消融实验

设置	*SAD*↓	*MSE*↓	*MAD*↓
–CA	10.46	3.63	6.21
+SE	9.72	3.41	5.84
+CBAM	10.86	3.74	6.25
+CA	**9.05**	**2.71**	**5.27**

3.6 实例应用

为更直观地展现本文方法在抠图方面的实际应用,

本文选取某抠图软件与本文方法进行抠图测试结果对比. 使用配备的计算机运行模型对现实生活场景照片进行了人像抠取测试, 实现了不同背景的替换, 如图 10 所示, (b) 为本文方法抠图结果, (c) 为某抠图软件抠图结果, (d) 为透明度遮罩预测结果.

(a) 原图 (b) DPSAM (c) 抠图软件

(d) 透明度遮罩 (e) 背景 (f) 合成结果

图 10 抠图测试

由图 10 可见, 抠图软件的抠图结果存在人物前景轮廓预测粗糙、前景误判等问题; DPSAM 仅输入单张 RGB 图像提取精细的透明度遮罩, 然后以此提取发丝级的人像前景. 同时, 利用 DPSAM 所提取的透明度遮罩, 将前景物体合成到新的背景中, 人物与新的背景自然地结合在一起, 呈现出较好的视觉效果 (见图 10 (e)、(f)). 测试结果表明, 得益于 DPSAM 对高层语义的捕获及细节信息的引导, 本文方法在提取人像轮廓时能够实现细腻的人像抠图效果, 特别是在发丝和细节的处理上表现突出. 通过该方法的自动化抠图与背景合成, 可以在实际应用中提供较高的实用性和美观度, 更迎合用户的高质量使用需求.

4 结论

针对人像数据复杂性和多样性导致人像抠图粗糙的问题, 本文设计了一种双金字塔编码结构和双分支解码结构的人像自动抠图方法. 首先, 本文构建的双金字塔编码结构有效地增强了抠图网络感知人像尺度变化的能力, 并通过双分支解码结构实现无辅助输入抠图. 同时, 本文所提出的细节感知单元及视域扩张模块针对性增强了网络对高层语义的感知、对细节捕获的引导, 从而提高抠图方法的预测精度. 在数据集 P3M-500-NP、P3M-500-P 和 PPM-100 上进行实验, 验证了本文方法在人像抠图任务具有不错的视觉效果. 在今后的工作中, 将进一步探索在保证抠图网络结构轻量化的同时提升抠图精度.

参考文献

1 Chen Q, Ge TZ, Xu YY, *et al*. Semantic human matting. Proceedings of the 26th ACM International Conference on Multimedia. Seoul: ACM, 2018. 618–626. [doi: 10.1145/3240508.3240610]

2 Kim SH, Tai YW, Park J, *et al*. Multi-view object extraction with fractional boundaries. IEEE Transactions on Image Processing, 2016, 25(8): 3639–3654. [doi: 10.1109/TIP.2016.2555698]

3 Huang LT, Liu XP, Wang XL, *et al*. Deep learning methods in image matting: A survey. Applied Sciences, 2023, 13(11): 6512. [doi: 10.3390/app13116512]

4 Xu N, Price B, Cohen S, *et al*. Deep image matting. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 311–320. [doi: 10.1109/CVPR.2017.41]

5 Lutz S, Amplianitis K, Smolic A. AlphaGAN: Generative adversarial networks for natural image matting. Proceedings of the 2018 British Machine Vision Conference. Newcastle: BMVA Press, 2018.

6 Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 4396–4405. [doi: 10.1109/CVPR.2019.00453]

7 Sun YN, Tang CK, Tai YW. Semantic image matting. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 11115–11124. [doi: 10.1109/CVPR46437.2021.01097]

8 Park G, Son S, Yoo J, *et al*. MatteFormer: Transformer-based image matting via prior-tokens. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 11696–11706. [doi: 10.1109/CVPR52688.2022.01140]

9 Liu Z, Lin YT, Cao Y, *et al*. Swin Transformer: Hierarchical vision Transformer using shifted windows. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 10012–10022. [doi: 10.1109/

ICCV48922.2021.00986]

10 Cai HQ, Xue FL, Xu LL, *et al*. TransMatting: Enhancing transparent objects matting with Transformers. Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022. 253–269. [doi: 10.1007/978-3-031-19818-2_15]

11 Sengupta S, Jayaram V, Curless B, *et al*. Background matting: The world is your green screen. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 2288–2297. [doi: 10.1109/cvpr42600.2020.00236]

12 Lin SC, Ryabtsev A, Sengupta S, *et al*. Real-time high-resolution background matting. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 8758–8767. [doi: 10.1109/CVPR46437.2021.00865]

13 Ke ZH, Sun JY, Li KC, *et al*. MODNet: Real-time trimap-free portrait matting via objective decomposition. Proceedings of the 36th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2022. 1140–1147. [doi: 10.1609/aaai.v36i1.19999]

14 Li JZZ, Zhang J, Maybank SJ, *et al*. Bridging composite and real: Towards end-to-end deep image matting. International Journal of Computer Vision, 2022, 130(2): 246–266. [doi: 10.1007/s11263-021-01541-0]

15 Li JZZ, Ma SH, Zhang J, *et al*. Privacy-preserving portrait matting. Proceedings of the 29th ACM International Conference on Multimedia. ACM Press, 2021. 3501–3509. [doi: 10.1145/3474085.3475512]

16 Peng C, Zhang XY, Yu G, *et al*. Large kernel matters-improve semantic segmentation by global convolutional network. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 1743–1751. [doi: 10.1109/CVPR.2017.189]

17 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. [doi: 10.1109/cvpr.2016.90]

18 Chen LC, Papandreou G, Kokkinos I, *et al*. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834–848. [doi: 10.1109/TPAMI.2017.2699184]

19 Chen LC, Papandreou G, Schroff F, *et al*. Rethinking atrous convolution for semantic image segmentation. arXiv:1706.05587, 2017.

20 Hou QQ, Liu F. Context-aware image matting for simultaneous foreground and alpha estimation. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 4130–4139. [doi: 10.1109/ICCV.2019.00423]

21 Qiao Y, Liu YH, Yang X, *et al*. Attention-guided hierarchical structure aggregation for image matting. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 13673–13682. [doi: 10.1109/CVPR42600.2020.01369]

22 Zhang YK, Gong LX, Fan LB, *et al*. A late fusion CNN for digital matting. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 7461–7470. [doi: 10.1109/CVPR.2019.00765]

23 Huang G, Liu Z, Van Der Maaten L, *et al*. Densely connected convolutional networks. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 2261–2269. [doi: 10.1109/CVPR.2017.243]

24 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141. [doi: 10.1109/CVPR.2018.00745]

25 Woo S, Park J, Lee JY, *et al*. CBAM: Convolutional block attention module. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 3–19. [doi: 10.1007/978-3-030-01234-2_1]

(校对责编: 王欣欣)