控制与决策
Control and Decision

基于多级表征线索注意模型的轻量化抠图方法

刘相良, 张林丛, 朱宏博, 张文波

引用本文:
刘相良,张林丛,朱宏博,张文波. 基于多级表征线索注意模型的轻量化抠图方法[J]. 控制与决策, 2024, 39(1): 87−94.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2022.0585

基于多级表征线索注意模型的轻量化抠图方法

刘相良, 张林丛†, 朱宏博, 张文波

(沈阳理工大学 信息科学与工程学院，沈阳 110159)

摘 要: 面对主流计算平台对框架轻量化的需求,设计一种基于多任务结构的轻量化抠图框架. 将总体任务拆分为两类子任务,其中一类任务用来在语义层面上为高级特征分类,区分前景背景与未知区域的特征;另一类任务用于计算前景与背景图层的线性组合权重. 通过与特征分类任务共享高级特征网络的权值获得精准的前景特征,再与低级别卷积特征相融合. 所提出的模型能够生成精准的抠图掩膜,同时优化卷积神经网络来实现模型轻量化. 在 Composition 1K 数据集上对比不同方法的实验结果: 在分辨率为 640 × 640 的输入条件下,所提方法比 DIM (deep image matting) 和 AdaMatting (adaptation and matting) 方法分别减少 19 % 和 81 % 的空间消耗;对于同样的数据输入,所提方法需要的处理时间只有 DIM 消耗时间的五分之一.

关键词: 数字抠图技术；轻量化；三分图；多任务框架；多级表征线索

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2022.0585

引用格式: 刘相良,张林丛,朱宏博,等. 基于多级表征线索注意模型的轻量化抠图方法 [J]. 控制与决策, 2024, 39(1): 87-94.

A lightweight image matting method based on attentive model for multi-level appearance cues

LIU Xiang-liang, ZHANG Lin-cong†, ZHU Hong-bo, ZHANG Wen-bo

(School of Information Science and Engineering，Shenyang Ligong University，Shenyang 110159，China)

Abstract: A lightweight image matting framework, which is based on multi-task structure, is designed to meet the requirements of the mainstream computing platforms. Concretely, the overall task can be split into two sub-tasks. One sub-task is to classify the higher-level features at the semantic level, and then it distinguishes foreground/background features from the unknown regions. Another task is to calculate the weights of the linear combination for the foreground and background layers. Accurate foreground features are obtained by sharing the weights of high-level feature networks with feature classification tasks, and they are fused with low-level convolution features. The proposed model outputs more accurate mattes. Also, the convolutional neural network is optimized to lightweight the model. On a benchmark dataset of Composition 1K, schemes performance is compared with different architectures. The proposal can reduce the 19 % and 81 % of storage consumption in comparison with DIM (deep image matting) and AdaMatting (adaptation and matting) on 640 × 640 images. For the identical data inputs, the running time of the proposed model is only about $1/5$ of DIM's.

Keywords: digital image matting；lightweight；trimap；multi-task framework；multi-level appearance cues

0 引 言

抠图是图像处理领域近年来的热门研究课题,这一技术的主要目标是从图像中提取前景目标并使其能够与其他背景组合的更加自然,实现拍摄素材的复用,其中最为知名的应用场景当属电影拍摄中的数字图像编辑. 在诸多抠图技术中, alpha-matting[1] 由于其更为简单而有效的既定模型和较好的抠图效果成为了目前抠图应用中的主流技术. 对于 alpha-matting 一般采用人工预标定手段作为抠图约束引导条件. 常见的有三分图 (trimap)[2] 和草图 (scribble)[3] 两种人工添加约束条件方式. 其中,借助 trimap 作为额外约束的抠图方式主要是借助 trimap 图像将输入

收稿日期: 2022-04-11；**录用日期:** 2022-08-09.

基金项目: 国家自然科学基金青年项目 (62102272)；沈阳理工大学引进高层次人才科研支持计划项目 (1010147000911)；辽宁省教育厅青年科技人才“育苗”项目 (LG202027).

责任编委: 张国山.

†通讯作者. E-mail: lincongz@foxmail.com.

的RGB图像分成前景、背景和未知区域3个部分. 这种方式可以提供较为全面的前景和背景特征参考模板,但制作一张高精度trimap图像的过程较为繁琐. 基于scribble的抠图方式则是通过手动标记前景和背景目标为算法指定抠图的目标区域,尽管相对于基于trimap的抠图方式,基于scribble的抠图方式[4-7]用户体验好一些,但是由于没有精确的前景、背景特征可以参考,基于scribble的抠图方式得到的matte精度更低. 因此,大多数传统抠图方法都是基于trimap的抠图方式.

在此基础上,传统的抠图方式主要分为基于采样的抠图方法[1,8-10]和基于传播的抠图方法[11-15]. 其中基于采样的抠图算法就是通过从背景区域和前景区域采集样本对,为给定的前景和背景像素寻找候选颜色. 采样到的候选颜色应该十分接近前景或背景颜色. 一旦确定了前景颜色和背景颜色,便可计算相应的α值. 根据图像的局部平滑假设,相关学者提出了更多基于采样的抠图方式,包括贝叶斯抠图[11]、共享采样抠图、全局采样抠图[2]和稀疏编码抠图[8].

对比基于采样的抠图方法,基于传播模型的抠图方法[8-12]避免了在基于采样的方法中可能遭受的α值(matte)连续问题,从而有效保证了包含空洞的前景的抠图效果逼近真实值. 基于该模型的方法利用相邻像素间的相似度将α值从已知区域传播到未知区域. 其中最为经典的方法是封闭型抠图[4],它通过求解一个稀疏线性方程组的封闭解来找到全局最优的α值. 测试图像上,该方法获得了较好的抠图效果,但其抠图效果较为依赖草图输入标记的准确性. 此外,基于传播的抠图方式还包括泊松抠图[11]、随机游走抠图[2]、KNN抠图[13]和信息流抠图[14].

尽管传统的抠图方法在特定的场景下能够提供较为可观的抠图结果,但其痛点在于通过构建复杂数学模型或以特征工程为切入点求解α值的方式缺乏较强的泛化能力,在大量生图上得到抠图效果难以尽如人意. 深度学习(deep learning)在包括图像抠图在内的各种计算机视觉任务中表现出了良好性能. Cho等[16]提出了一个基于深度卷积网络(deep convolutional neural network, DCNN)的端到端架构,该架构利用闭合型抠图[4]和KNN抠图[13]的结果来更好地预测α值. DIM (deep image matting)[17]率先将深度学习引入到抠图任务的研究中,并获得高精度的抠图结果. 该文提出同一类抠图对象可以由同一种高级的语义特征表示,而一个RGB图像可以提取出高级别的语义特征和低级别的表征线索特征,在抠图任务中高级别的语义特征表示前景的类别,而表征线索特征表示前景的纹理细节. 此后的抠图算法[18-21]大多为前景的高级语意提取设计复杂的结构,并融合来自输入图像或者低级CNN (convolutional neural networks)特征的表征线索. Shen等[22]提出了一种基于CNNs的全自动人像照片抠图系统. Lutz等[23]利用对抗性学习的能力来提取α-matte,从而产生视觉效果很好的合成. Qiao等[24]将注意力机制引入到抠图任务中,取得了令人满意的效果. Wang等[21]表示,基于传播的抠图可以通过深度学习机制学习语义级的成对相似度.

综上所述,基于深度学习的抠图方式存在两个亟待解决的问题: 1) 构建深度学习神经网络需要大量的算力消耗,这意意味着基于深度学习的抠图方式很难布置在边缘设备上; 2) 较浅的网络对背景过滤的能力不足,导致前景目标上与背景相似的边缘易与背景混淆; 3) 此外,目前移动端的抠图框架/应用主要采用离线训练方式,从而无法进行有效的用户个性化网络训练. 因此,如何顺应时代需求设计出一套可在边缘设备上训练,并且可流畅运行的抠图框架就成为研究中的重点. 针对上述问题,本文从以下几个方面展开工作:

1) 提出一种基于多级语义线索注意力机制的数字抠图框架. 针对深度学习抠图算法前景边缘易与背景混淆的问题,引入多头注意力机制,分别从全局线索与局部细节入手生成最优化抠图掩膜.

2) 在框架设计方面,充分地考虑框架各组件的轻量化. 通过引入不对称卷积、轻量化的注意力机制、密集连接的特征融合等框架设计,缩减框架整体计算开销,便于提出框架后续在边缘设备上的部署.

3) 整体结构上,引入生成对抗神经网络,并采用多任务结构共同作用于数字抠图任务,在多个任务的互补下使抠图结果逐渐趋近于最佳.

1 轻量化数字图像抠图框架设计

1.1 数字抠图技术基础

alpha-matting[1]技术的核心思想可以简要描述为:将输入图片看作是前景与背景的线性组合,有

$$I = \alpha F + (1-\alpha)B. \tag{1}$$

其中: I表示输入图像; F表示前景区域; B表示背景区域; α表示透过率,当其为1时表示该区域为前景,为0表示该区域为背景,当介于二者之间则表示该区

域为前后景图像的线性组合. 可以发现上述方程有3个未知量,所以该方程是一个欠约束方程. 已有的大多数抠图方案需要手动添加约束条件,这使得模型需要对输入图像的颜色、纹理有预先的估计和假设.

1.2　网络结构

为了解决目前研究方法中对前景和背景的多尺度语义特征分离能力不足的问题,本文引入通道注意力机制与空间注意力机制,并充分考虑边缘设备上算力紧张的问题引入大量的轻量化策略,设计一种基于生成对抗神经网络的多任务轻量化抠图框架(如图1所示). 从整体上看,本文网络基于对抗生成神经网络设计,框架可分为生成模型和判别模型两部分. 基于深度模型强大的生成能力以及潜变量对图像语义分割的导向作用生成α-matte,用以为输入图像输出备选抠图结果,后将输出结果交由判别器鉴别生成结果并完成优化获得最优的α-matte. 在基于深度学习图像处理领域,面对同样的算力消耗,深而窄的网络要比浅而深的网络更有效. 网络的算力消耗主要是由各网络层的参数造成的,也就是说网络层的宽度直接影响模型的算力消耗,所以需要控制每一层的通道数量,尽量保证在压缩空间消耗的前提下,不严重影响特征提取的效果.

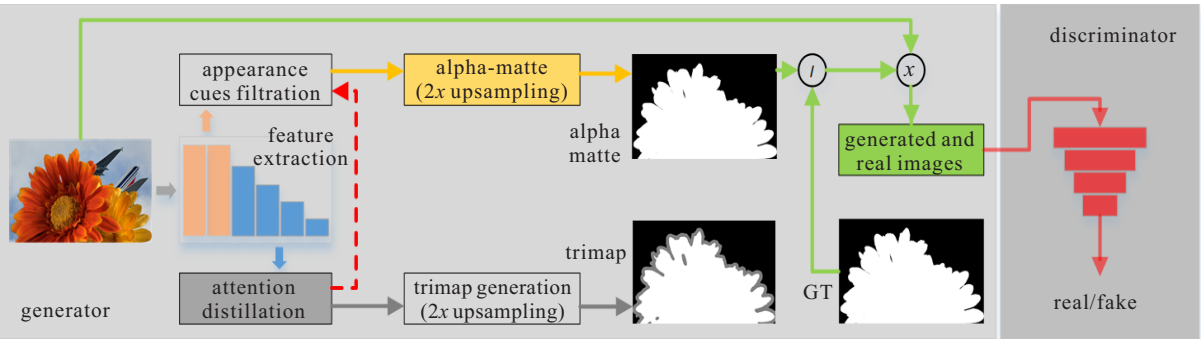


图 1　本文所提出方法的主体框架

残差网络解决了深度神经网络隐藏层过多的网络退化问题,近些年在深度学习领域有着广泛的应用,并获得了优秀的表现. 本文模型参考经典的残差网络结构设计,在模型中分别测试了经典的残差网络架构ResNet与密集连接的轻量化残差网络DenseNet的网络结构在生成模型中的表现. 考虑到特征在由低层网络向深层传递过程中会丢失一些细节信息,有必要将深层特征与浅层特征进行特征融合. U-Net[25]使用Concat操作将深层特征与浅层特征叠加融合,之后的U-Net3+[26]中的每个解码器层都融合了来自编码器的较小和相同尺度的特征图以及来自解码器的较大尺度的特征图. 它们捕获了全尺度下的细粒度语义和粗粒度语义,取得了很好的效果. 本文也尝试这种方式,以求融合多尺度下的特征. 同时,基于不同的特征融合方式,分别设置向前连接与向后连接的方式.

本文采用注意力机制[27]和特征融合相结合的方式,对照图1中的注意力蒸馏(attention fistillation)结构. 从图1中可以看出,本文使用多头注意力机制[9]融合后语义中的隐藏信息输入至多层感知器(multi-layer perceptron, MLP)以提供包含复杂高层特征的全局编码,其后将其与表征线索滤波器(appearance cues filtration)结合提高模型的非线性表达能力. 根据式(1)可知,按照α-matte的取值范围可以将目标图像分为3个区域:透明度取值为0的背景区域B;透明度为1的前景区域F;取值介于$(0,1)$之间的不确定区域,称为未知区域U. 具体定义如下所示:

$$p(x,y)=\begin{cases}B, & \alpha=0;\\ U, & 0<\alpha<1;\\ F, & \alpha=1.\end{cases} \tag{2}$$

由式(2)可知,抠图任务可以分解成两个相关的任务:第1个任务是区分不同区域的像素,本文称之为trimap-task;第2个任务是精确计算未知区域的透明度,本文称之为alpha-task. 在现有研究中,已有学者初步尝试采用双任务结构解决抠图问题,但是该工作使用后处理的方式对alpha-task做精细化处理,这无疑影响了框架处理效率.

本文工作从设计之初就考虑了后处理可能给用户带来的不良使用体验. 对于trimap-task任务,这一模块的主要任务是对语义特征分类. 将包含前景轮廓的深层语义信息输入transformer进行注意力蒸馏,再将筛选后的语义特征输入到trimap-task的上采样模块并输出一个trimap图像. 尽管trimap图像与最终

的α-matte没有直接关系,但是trimap可以区分目标图像中前景、背景和未知区域的像素,这对α值的计算起到了关键性的指导作用. alpha-task任务是为了精确计算前景在未知区域的边缘细节,框架将高级语义同低级特征输入边缘过滤器以过滤背景细节特征,再将过滤后的特征输入alpha-task上采样模块获得更为可靠的matte信息(如图2所示).

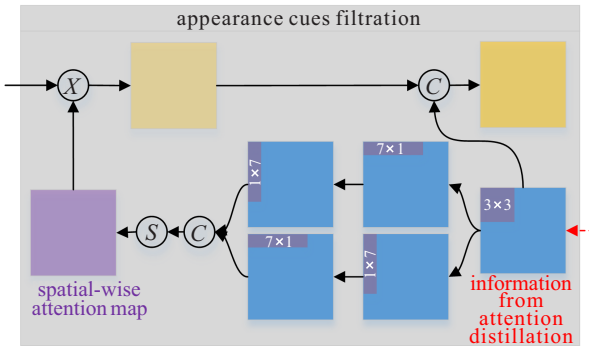


图 2 本文表征线索过滤模型的具体处理流程

多任务模型输出的多个预测α-matte与其真实值经由一个点乘操作输出一组前景图像,将其输出至判别模块(discriminator). 该模块与生成模型共同构成一个GAN[28]结构,基于生成模型强大的生成能力为图像持续提供抠图备选结果. 通过计算生成对抗损失与像素级的损失,逐步调优生成模型的抠图结果至最佳.

1.3 网络轻量化

基于深度学习的图像处理方法往往都依赖庞大的算力消耗,这使得基于深度学习的抠图方式很难部署到边缘设备上. 各卷积层的输出是训练期间空间消耗的最重要因素,文献[29]去掉了全卷积一步检测框架(fully convolutional one-stage object detection, FCOS)系列难以训练的centerness分支的大量卷积,合理地使用插值法代替卷积层修改特征地图大小,减少了检测头的计算开销,并且对深度学习的结构只会产生有限的影响. 对于一个抠图任务而言,不同于检测任务或者分割任务,抠图任务依赖比较精确的特征计算,所以本文在轻量化的过程中只在特征地图缩放的过程中进行不大于2倍的缩放. 在过滤器设计上,本文引入不对称卷积[30],主要目的在于:对于一个输入张量,先进行$n\times 1$卷积再进行$1\times n$卷积,与直接进行$n\times n$卷积的结果是等价的,但是相比于直接使用正方形的卷积核,使用两个一维的卷积核所需要的参数要小很多,可以降低模型的计算量,达到轻量化模型的目的. 对于一个用于图像处理的深度网络,在较深的网络中携带着最高级的信息,通道数量也是最大的. 过早地对深度的语义上采样会造成很大的空间压力,所以本文提出的架构选择只在高层进行复杂的计算,这就需要考虑在特征融合的过程中是选择将中间层的特征与深层特征叠加融合,还是将中间层的特征与浅层特征叠加融合. 具体的融合过程如图3所示.

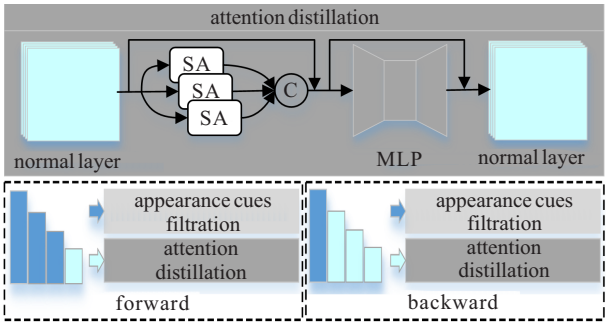


图 3 本文注意力蒸馏模型中给出的前向连接与反向连接

同样出于轻量化方面的考虑,本文在生成器的设计上采用了transformer结构. 其中多头注意力机制融合多级语义信息中的隐藏信息,并交由多层感知机编码高级特征作为最终图像合成阶段的线索信息.

1.4 损失函数

本文使用多任务结构搭建数字图像抠图方法,其中多任务结构意在一个模型解决多个任务. 与单任务相比,多任务结构更为灵活地利用计算资源,提高了计算效率和抠图性能,其可以看作通过互补任务[31]之间共享的域信息以诱导知识转移的方法. 本文将抠图任务分解为3个相关任务,分别是对alpha在像素级上运算的alpha精度计算任务、对特征在语义上分类的trimap任务、基于GAN[28]的语义判别任务. 对应上述3个任务,总体损失函数可以写作

$$L=\lambda_1\mathcal{L}_\alpha+\lambda_2\mathcal{L}_{\rm adv}+\lambda_3\mathcal{L}_{\rm map}. \tag{3}$$

其中:$\mathcal{L}_\alpha$表示生成的α-matte执行抠图所带来的像素级损失,$\mathcal{L}_{\rm adv}$表示生成器G与判别模型D的生成对抗损失[28],$\mathcal{L}_{\rm map}$表示trimap任务的分析损失. 对于像素级损失的评估,使用均方差MSE损失$\ell_{\rm MSE}$与结构相识度(structural similarity, SSIM)损失$\ell_{\rm SSIM}$构建损失函数,有

$$\mathcal{L}_\alpha=\ell_{\rm MSE}(\hat{\alpha},\alpha_{\rm gt})+\ell_{\rm SSIM}(\hat{F},F_{\rm gt}). \tag{4}$$

其中:$\hat{\alpha}$是由生成器G生成得到的备选结果,$\alpha_{\rm gt}$是数据集给定的参考α-matte,$\hat{F}$与F_{gt}分别是根据预测与参考α-matte得到的前景抠图结果,$F=\alpha_*\times I$. $\mathcal{L}_{\rm map}$

同样由两部分组成:二值交叉熵损失$\ell_{\rm bce}$和适应性评估损失$\ell_{\rm est}$,可写作

$$\mathcal{L}_{\rm map}=\ell_{\rm bce}+\ell_{\rm est}, \tag{5}$$

$$\ell_{\rm est}=\frac{1}{\|P\|}\sum_{p\in P}\|\hat{\alpha}(p)-\alpha_{\rm gt}(p)\|, \tag{6}$$

其中$\|P\|$是当前处理子图的像素总数. 损失函数实际上将图像抠图分解为两部分,确保每个解码器分别学习结构语义和光度信息.

2 实　验

2.1 实验数据与数据扩充

实验部分使用Adobe Composition-1K数据集(https:// paperswithcode. com/ dataset/ composition-1k),该据集是一个公开的数据集,包含431张前景图片和与之匹配的前景α-matte. 从Microsoft COCO数据集(http://cocodataset.org/#download)中选取100张背景图片与前景图片组合形成训练集. 测试集包含50张前景图片和相应的α-matte,从PASCAL VOC 2012数据集(http://host.robots.ox.ac.uk/pascal/VOC/voc2012)中挑选1 000张图片作为背景,数据集分割与处理的方式与文献[2]完全一致. 所有的输入图片被随机初始裁剪为640×640、512×512和320×320分辨率,后统一缩放为320×230分辨率. 训练使用的trimap是使用随机形态学操作从标准α-matte生成的. 此外,对所有的图像采用$0.75\sim1.5$之间的随机尺度缩放,并额外添加$-45°\sim45°$之间的随机旋转. 训练数据在每一轮都会被随机打乱.

2.2 评价指标

评价指标使用alpha-matting最为常用的3个指标,分别是绝对误差(SAD)、均方误差(MSE)、结构相似性(SSIM)[32],对本文算法生成的α值与目前的几个经典的alpha-matting算法进行对比评价. 各指标的公式如下:

$$\mathrm{SAD}=\frac{1}{N}\sum_{p\in P}\|x(p)-y(p)\|, \tag{7}$$

$$\mathrm{MSE}=\frac{1}{N}\sum_{p\in P}(x(p)-y(p))^2, \tag{8}$$

$$\mathrm{SSIM}=\frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)}. \tag{9}$$

其中:p是像素索引,N是未知区域像素点的个数,μ_x和μ_y分别是图像x和y的均值,σ_x和σ_y是标准差.

2.3 实验设置

实验设置遵循文献[21],同样使用Poly作为学习率更新策略. 学习率更新策略可以表示为

$$\tau_{t+1}=\tau_t\times\left(1-\frac{t}{T}\right)^p. \tag{10}$$

其中:$t\leqslant T$是当前迭代次数,T是本轮最大迭代次数. 初始学习率和p分别设置为0.000 1和0.9. 梯度优化器为Adam,其动量和权重衰减分别设置为0.9和0.000 1.

实验框架基于Pytorch库构建,计算设备采用Ubuntu 18.04作为操作系统,配有Intel I9 10980 XE,256 G内存,以及一块NVIDIA RTX3090 24 G计算卡. CUDA版本选择的是11.3版本,CUDNN为8.3版本. 对于数字抠图空间和时间消耗实验,使用CPU型号Intel I5 10400F,16 G内存,显卡采用一块Titan X Pascal 12 G计算卡.

在训练过程中,原始输入图像与trimap拼接成为4通道张量并作为网络输入. 模型分别采用Resnet[10]和Densnet[3]作为骨干网络,相应地修改了每一层的特征通道数. 由于抠图任务与大多数分类任务不同,目前实验没有使用预训练模型初始化.

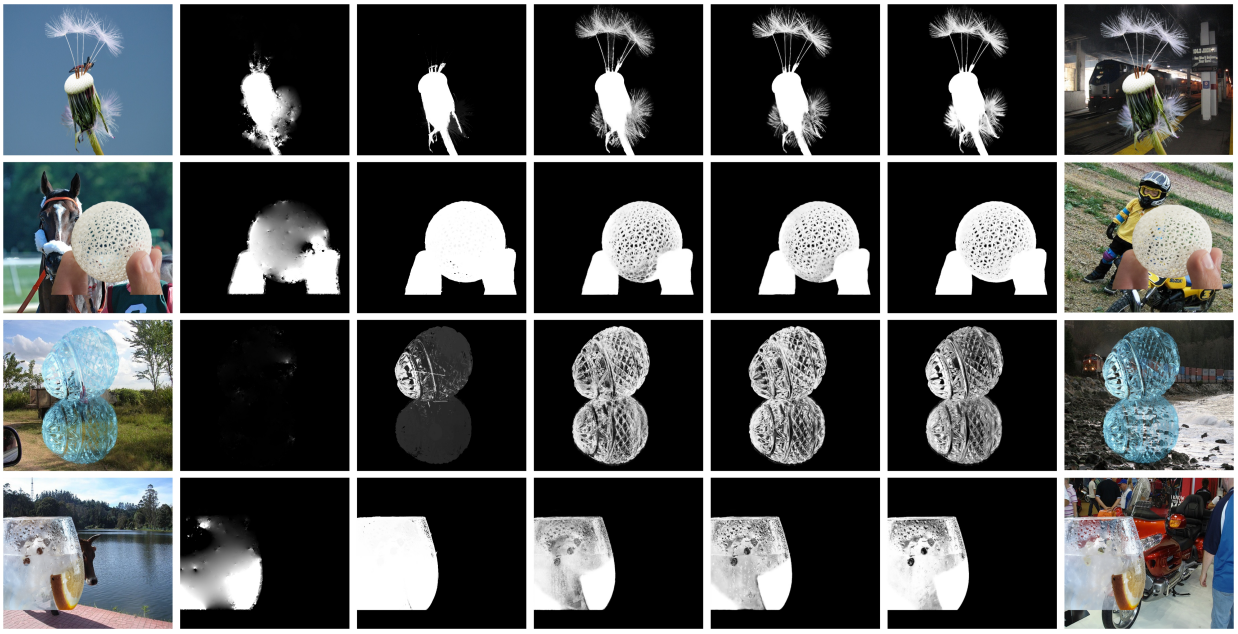
2.4 实验结果

对比实验方面选取较为有代表性的3个抠图方法作为参照,分别是closed-form matting[4],KNN matting[13]和DIM[17]. 其中,close-form matting和KNN matting是两个最具代表性的基于人工特征的抠图方法,分别基于交互式模型构建和传统机器学习. 具体的抠图效果与对应的量化结果见图4和表1.

表1　经90轮训练后不同算法的抠图量化结果

方法	SAD	SSIM	MSE
close-form matting[4]	111.28	–	1.010 3
KNN matting[13]	106.27	–	0.953 4
DIM[17]	10.015	0.937 5	0.007 3
MODNet[33]	10.230	0.937 1	0.008 2
ours (forward-ResNet)	10.490	0.927 0	0.008 0
ours (backward-ResNet)	13.410	0.914 6	0.012 1
ours (forward-DenseNet)	10.280	0.926 3	0.008 1
ours (backward-DenseNet)	9.317	0.939 8	0.006 6

在实验过程中发现了很有趣的现象. 在使用ResNet[10]作为骨干的网络结构的结果中,使用前连接结构的网络获得的抠图精度要远远优于使用后连接的结构的网络. 相反地,在使用DenseNet[34]作为骨干的网络结构中,使用前连接结构的网络获得的抠图精度不及使用后连接的结构的网络. 为便于观察,将训练轮次扩大到120轮,结果如表2所示.



(a) 输入原图 (b) close-form matting (c) KNN matting (d) DIM (e) 本文方法 (f) 真实值 (g) 基于本文抠图结果与随机背景合成的图像

图 4 不同抠图方法效果对比

表 2 经 120 轮训练后不同方案的抠图量化结果

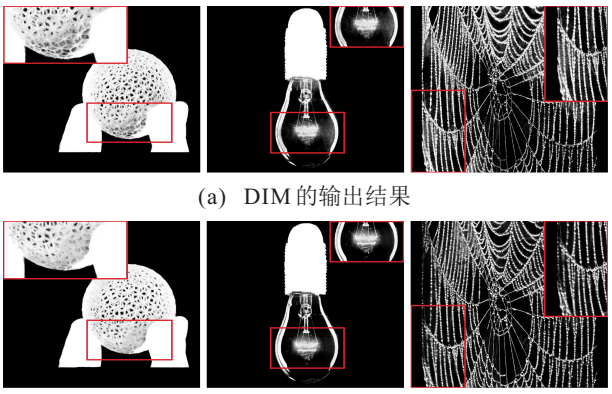
方法	SAD	SSIM	MSE
ours (forward-ResNet)	9.78	0.931	0.007 3
ours (backward-ResNet)	9.04	0.943	0.006 8
ours (forward-DenseNet)	10.26	0.929	0.007 8
ours (backward-DenseNet)	9.13	0.940	0.006 4

从表 2 的数据可以看出,在 120 轮的训练次数下,使用 ResNet[10] 作为骨干网络的结构,向后的连接方式获得了比向前的连接方式更好的抠图结果. 这说明,使用 ResNet 作为骨干网络导致向后连接的网络结构收敛变慢. 同时也可以看出,DenseNet 作为骨干网络的结构,在将训练轮次扩大到 120 轮之后,抠图精度并没有出现大幅度的提升.

通过观察可推测在向前连接的方式中,低级的细节特征和边缘信息没经过边缘过滤,造成上采样结果中有很多与前景混淆的背景噪音. 与之相比,向后连接的方式将高级的前景深度语义同浅层的相融合,融合了低层纹理特征的高级特征经过 transformer 后拥有更复杂的数学表达,并加强了特征的非线性表达能力,如此处理后再经过 trimap-task 强大的分类处理和边缘过滤器的过滤,可以很好地去除无用的背景信息.

对于使用的轻量化的抠图架构,Densnet[34] 提出的向后的密集连接的特征融合方式可以为本文轻量化框架提供更多的前景信息. 同时在实验中还发现,使用 Densnet 做为骨干网络会加剧后期网络的震荡,并且向前连接的方式要比向后的连接方式提前收敛. 在实验过程中使用 DenseNet 作为骨干网络的向前连接方式,甚至出现了连续 23 轮没有更新更优解点现象.

选取 DIM 和本文提出的模型在训练 90 轮之后获得的抠图局部效果作对比 (见图 5),其中第 1 排是 DIM 方法输出的 α-matte 结果,第 2 排是本文给出的抠图 α-matte. 同时,对抠图过程中消耗的时间和空间作对比. 在 Titan X Pascal 上对一次处理 16 张 320 × 320 的图片,不同模型占用的空间和处理时间作了对比,结果如表 3 所示. 其中:存储开销为模型加载并处理输入图像过程中的内存开销,运算时间为模型加载并在处理输入图像所需要的时间.



(a) DIM 的输出结果

(b) 本文所提出方法的输出结果

图 5 不同抠图方法的效果对比

表 3　不同模型在存储空间和计算时间上的开销

方法	输入$(n \times c \times w \times h)$	存储开销/ms	运行时间/ms
DIM[17]	16×4× 320× 320	5 098.4	315.22
AdaMatting[18]	4×4×320×320	5 563.2	19.75
MODNet[33]	16×4×320×320	4 599.8	47.68
ours (forward-resNet)	16×4×320×320	3 554.7	69.10
ours (backward-resNet)		4 229.0	74.88
ours (forward-denseNet)		2 481.1	35.67
ours (backward-denseNet)		2 796.7	37.90

通过实验结果对比可以发现,使用前连接的方式可以在更大程度上节省空间和时间,但是相应地对抠图的精度造成了一定的影响,原因是深度的特征空间没有足够的宽度容纳丰富的语义信息.

3　结　论

本文设计了一种可用于数字图像抠图任务的深度网络架构. 针对在一般基于学习的抠图算法中易出现的前景边缘易与背景混淆的问题,引入了多级表征线索注意力机制、提高了模型对于前景特征的提取能力. 对于边缘设备算力低下的问题,本文设计的抠图框架分别从存储空间和计算时间消耗上对模型进行优化. 通过对比不同网络结构在Composition 1K数据集上的抠图效果,验证了不对称卷积与注意力机制的结合可以在实现轻量化的同时获得较高质量的特征效果. 同时通过算力对比实验表明本文所提出方法在640 × 640的图像上空间消耗比DIM和AdaMatting方法分别减少19 %和81 %. 对于同样的数据输入,本文提出的模型处理所需要的时间仅为DIM消耗时间的12 %～20 %. 本文提出的框架基于生成对抗神经网络,并采用多任务结构,提高了训练效率.

参考文献(References)

[1] Chuang Y Y, Curless B, Salesin D H, et al. A Bayesian approach to digital matting[C]. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Kauai, 2001: 7184431.

[2] Grady L, Schiwietz T, Aharon S, et al. Random walks for interactive alpha-matting[C]. Visualization, Imaging, and Image Processing. Benidorm, 2005: 423-429.

[3] Gastal E S L, Oliveira M M. Shared sampling for real-time alpha matting[J]. Computer Graphics Forum, 2010, 29(2): 575-584.

[4] Levin A, Lischinski D, Weiss Y. A closed-form solution to natural image matting[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(2): 228-242.

[5] Levin A, Rav-Acha A, Lischinski D. Spectral matting[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(10): 1699-1712.

[6] Wang J, Cohen M F. An iterative optimization approach for unified image segmentation and matting[C]. Tenth IEEE International Conference on Computer Vision. Beijing, 2005: 936-943.

[7] Guan Y, Chen W, Liang X, et al. Easy matting-a stroke based approach for continuous image matting[J]. Computer Graphics Forum, 2006, 25(3): 567-576.

[8] Feng X X, Liang X H, Zhang Z L. A cluster sampling method for image matting via sparse coding[C]. European Conference on Computer Vision. Amsterdam, 2016: 204-219.

[9] He K M, Rhemann C, Rother C, et al. A global sampling method for alpha matting[C]. Conference on Computer Vision and Pattern Recognition. Colorado Springs, 2011: 2049-2056.

[10] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 770-778.

[11] Sun J, Jia J Y, Tang C K, et al. Poisson matting[C]. Special Interest Group on Computer Graphics and Interactive Techniques. Los Angeles, 2004: 315-321.

[12] Lee P, Wu Y. Nonlocal matting[C]. Conference on Computer Vision and Pattern Recognition. Colorado Springs, 2011: 2193-2200.

[13] Chen Q F, Li D, Tang C K. KNN matting[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(9): 2175-2188.

[14] Aksoy Y, Aydin T O, Pollefeys M. Designing effective inter-pixel information flow for natural image matting[C]. IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 228-236.

[15] Li X, Wang W, Wu L, et al. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection[J]. Advances in Neural Information Processing Systems, 2020, 33: 21002-21012.

[16] Cho D, Tai Y W, Kweon I. Natural image matting using deep convolutional neural networks[C]. European Conference on Computer Vision. Amsterdam, 2016: 626-643.

[17] Xu N, Price B, Cohen S, et al. Deep image matting[C]. IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 311-320.

[18] Cai S F, Zhang X S, Fan H Q, et al. Disentangled image matting[C]. 2019 IEEE/CVF International Conference on Computer Vision. Seoul, 2019: 8818-8827.

[19] Hou Q Q, Liu F. Context-aware image matting for simultaneous foreground and alpha estimation[C]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, 2019: 4129-4138.

[20] Lu H, Dai Y T, Shen C H, et al. Indices matter: Learning to index for deep image matting[C]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, 2019: 3265-3274.

[21] Wang Y, Niu Y, Duan P Y, et al. Deep propagation based image matting[C]. Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. Stockholm, 2018: 999-1006.

[22] Shen X Y, Tao X, Gao H Y, et al. Deep automatic portrait matting[C]. European Conference on Computer Vision. Amsterdam, 2016: 92-107.

[23] Lutz S, Amplianitis K, Smolic A. AlphaGAN: Generative adversarial networks for natural image matting[J/OL]. 2018, arXiv: 1807.10088.

[24] Qiao Y, Liu Y H, Yang X, et al. Attention-guided hierarchical structure aggregation for image matting[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, 2020: 13673-13682.

[25] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]. Medical Image Computing and Computer-Assisted Intervention. Munich, 2015: 234-241.

[26] Huang H M, Lin L F, Tong R F, et al. UNet 3+: A full-scale connected UNet for medical image segmentation[C]. 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona, 2020: 1055-1059.

[27] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J/OL]. 2017, arXiv: 1706.03762.

[28] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[J/OL]. 2014, arXiv: 1406.2661.

[29] Uhrig J, Cordts M, Franke U, et al. Pixel-level encoding and depth layering for instance-level semantic labeling[C]. German Conference on Pattern Recognition. Hannover, 2016: 14-25.

[30] Peng C, Zhang X Y, Yu G, et al. Large kernel matters — improve semantic segmentation by global convolutional network[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 1743-1751.

[31] Cipolla R, Gal Y, Kendall A. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 7482-7491.

[32] Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: From error visibility to structural similarity[J]. IEEE Transactions on Image Processing: A Publication of the IEEE Signal Processing Society, 2004, 13(4): 600-612.

[33] Ke Z, Li K, Zhou Y, et al. Is a green screen really necessary for real-time portrait matting?[J/OL]. 2020, arXiv: 2011.11961.

[34] Huang G, Liu Z, van der Maaten L, et al. Densely connected convolutional networks[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 2261-2269.

作者简介

刘相良 (1995−), 男, 硕士生, 从事图像处理与机器学习的研究, E-mail: xiangliang.liu@hotmail.com;

张林丛 (1985−), 女, 副教授, 博士, 从事边缘智能计算、体域网等研究, E-mail: lincongz@foxmail.com;

朱宏博 (1986−), 男, 副教授, 博士, 从事图像处理、模式识别与智能计算等研究, E-mail: hombochu@sina.com;

张文波 (1973−), 男, 教授, 博士, 从事移动边缘计算、网络工程等研究, E-mail: zhangwenbo@yeah.net.