

基于显著性目标检测的图像前景提取算法研究

贾小云¹, 杨振英^{1†}, 邵帆², 罗豪才¹

(1. 陕西科技大学电子信息与人工智能学院, 陕西 西安 710021; 2. 西安高新第一中学, 陕西 西安 710075)

摘要:为解决自然场景中深度学习抠像算法依赖于人机交互且前景提取效果不佳的问题,提出了通过检测显著性目标的方式实现图像前景提取算法。该算法由显著性目标检测和 alpha 估计两个网络组成,前者在基准网络的基础上引入混合损失函数和深监督策略提高图像前景提取的完整度,后者通过增加上下文注意力引导模块的方式优化基准网络,实现对前景图像细节的精确恢复。改进算法以自然场景中人像作为待提取前景进行研究,在自制数据集上训练,公开数据集 Composition-1K 上测试。改进后算法在均方误差(MSE)指标上低至 2.329×10^{-4} 平方像素,梯度误差指标上相较基准算法减小近 60%。实验表明,改进后算法提升了头发、玻璃等半透明区域的 alpha 值估计的准确度从而提高图像中抠像精度,并且具有较强的鲁棒性与泛化能力,可应用于图像编辑等应用场景。

关键词:图像前景提取;显著性目标检测;alpha 估计;混合损失函数;深监督;上下文引导

中图分类号:TP391

文献标识码:A

Research on Foreground Image Extraction Based on Saliency Object Detection

JIA Xiaoyun¹, YANG Zhenying^{1†}, SHAO Fan², LUO Haocai¹

(1. School of Electronic Information and Artificial Intelligence, Shaanxi University of Science & Technology, Xi'an, Shaanxi 710021, China; 2. Xi'an High Tech No. 1 Middle School, Xi'an, Shaanxi 710075, China)

Abstract: To solve the problem of deep learning image matting algorithms in natural scenes relying on manual interaction and poor foreground extraction performance, a method of detecting salient targets is proposed to implement image foreground extraction algorithms. This algorithm consists of two networks: salient object detection and alpha estimation. The former introduces a mixed loss function and deep supervision strategy on the basis of the baseline network to improve the completeness of image foreground extraction, while the latter optimizes the baseline network by adding a context attention guidance module to achieve accurate restoration of foreground image details. The improved algorithm is studied using human portraits in natural scenes as the foreground to be extracted, trained on a self-made dataset, and tested on the publicly available dataset Composition-1K. The improved algorithm has the mean square error (MSE) index as low as 2.329×10^{-4} square pixels, and the gradient error index reduced by nearly 60% compared to the benchmark algorithm. The experiment shows that the improved algorithm improves the accuracy of alpha value estimation in semi-transparent areas such as hair and glass, thereby improving the accuracy of image matting. It also has strong robustness and generalization ability, and can be applied in application scenarios such as image editing.

Key words: foreground image extraction; saliency object detection; alpha estimates; mixed loss function; deep supervision; context guided

自然场景中自动抠像是指针对包含任意场景的图像进行前景提取,是图像和视频编辑中的关键任务。其中,常见的待提取对象有:包含头发丝的人像、透明的玻璃物体等。近年来,研究人员逐渐转向探索基于深度学习的图像前景提取算法。

大多数基于深度学习的抠像算法^[1]需要额外的三分图或者涂鸦作为先验信息,三分图标注了图片的前景、背景和不确定区域,涂鸦则标注了图片的前景和背景,而标注大规模数据成本很高,而且对于数据标注的精确度有较强的依赖性。Sengupta 等^[2]提出的算法虽能解决上述问题,但需要输入对齐的背景图像,实际应用难度较大。此外,随着深度学习的发展,更多的先验知识^[3]被引入抠像中致其对人机交互的强依赖。SHM^[4]则在处理人体全身像或者手持物体人像抠像的问题时,会存在前景提取不完整等现象。由此可见,抠像领域仍存在数据集场景不够丰富、深度学习算法前景提取效果不佳等问题,因此有必要对其进行研究。

为此,本文将人像作为待提取前景进行了以下方面的改进:

(1)针对数据集不足的问题,构建人像抠像领域数据集 AutoCrop;

(2)针对标准二值交叉熵评价局限造成的前景提取效果不佳的问题,引入混合损失函数,并采用深监督训练策略,提高图像前景提取的完整度;

(3)针对基准网络半透明区域 alpha 估计置信度不够高的问题,增加上下文注意力引导模块以实现图像前景细节的精确恢复。

实验结果显示,与基准算法效果相比,本文设计的改进算法具有一定先进性。

1 图像抠像算法原理与网络结构

图像抠像问题的数学模型是前景和背景颜色的线性混合表示,核心问题是解 PORTER 和 DUFF^[5]提出的如下的欠约束数学公式:

$$I_i = \alpha_i F_i + (1 - \alpha_i) B_i \quad (1)$$

其中,未知像素 I_i 表示前景像素 F_i 和背景像素 B_i 的线性组合,透明度 α_i 。对于给定的图像 I ,其中的前景像素 F 和背景像素 B 以及 α 值都是未知的,所有图像抠像算法的难点在于如何准确估计像素的 α 值。

本文通过图像前景提取实现人像自动抠像,将图像前景提取任务划分为两阶段,第一阶段的任务是改进 PoolNet^[6] 显著性检测网络实现前景的粗

分割,第二阶段的任务是改进 MGMatting^[3] 算法实现像素级的边界细化。总体图像前景提取流程如图 1 所示。

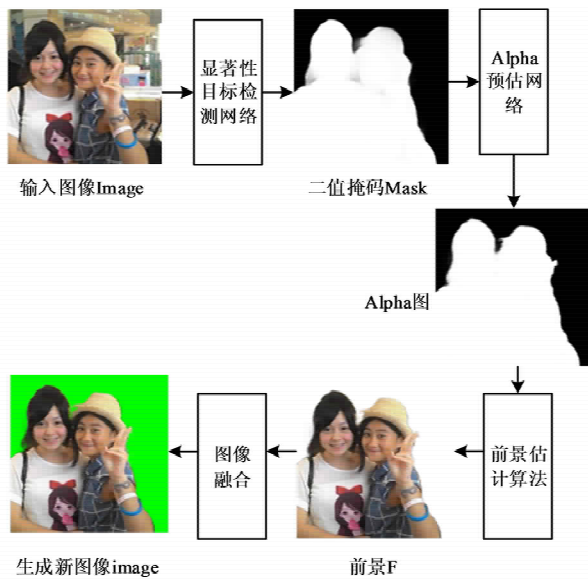


图 1 图像前景自动提取流程

首先将任意一张自然场景的图片 Image(此处为人像),输入到显著性目标检测网络,输出图像的二值掩码 Mask;然后将 Mask 和 Image 两者同时作为先验信息,输入到 alpha 估计网络,输出像素取值在 0 和 1 范围内的 alpha 图;最后采用前景估计算法,计算图像的前景 F ;前景 F 和任意背景 B 融合,得到新图像 image。

1.1 显著性目标检测网络

本文采用的 PoolNet 利用池化技术、全局引导模块(Global Guidance Module, GGM)特征聚合模块(Feature Aggregation Module, FAM)达到锐化显著物体细节以及提升检测速度的效果。但是该算法采用标准二值交叉熵的损失函数,仅通过比较网络的预测值与真实标签之间的差异来衡量网络的性能,未考虑到图像的全局信息、细节信息的融合,我们引入混合损失函数缓解这一问题。

1.1.1 网络结构

改进后的显著性检测网络采用类似于特征金字塔(Feature Pyramid Networks, FPN)的网络结构,将 ResNet-50 作为主干网络,金字塔池化模块 PPM(Pyramid Pooling Module, PPM)位于自底向上的路径与自顶向下的路径之间,两个路径采用 1×1 的卷积核横向连接, FAM 位于自顶向下路径的特征融合之前,全局引导流(Global Guiding Flows, GGFs)将 PPM 收集的语义信息传递给 FUSE 操作进行特征融合。显著性检测网络结构

如图 2 所示。其中,C2,C3,C4,C5 分别表示 Res-Net-50 的 4 个残差块输出;P1,P2,P3,P4,P5 则表示 FPN 金字塔结构的五层输出;⊕表示 FUSE 操作;虚线箭头代表全局引导流 GGFs;GT 代表图像对应的真实标签。

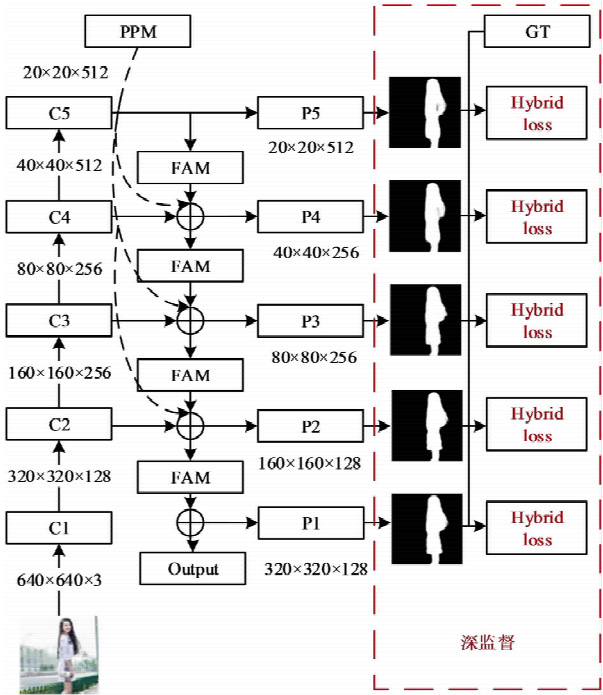


图 2 显著性目标检测网络结构

1.1.2 混合损失函数

本文在指导模型训练时借鉴混合损失函数^[7] (Hybrid Loss),并对其进行改进,具体公式表示为:

$$L_{\text{hybrid}} = L_{\text{focal}} + L_{\text{iou}} + L_{\text{ms-ssim}}$$

(2)

其中, L_{focal} , $L_{\text{ms-ssim}}$, L_{iou} 分别表示 Focal Loss、MS-SSIM Loss、IoU Loss。相较于基准网络中采用的 BCE Loss,Focal Loss 引入衰减因子以降低易分类样本的权重,提高难分类样本的权重,从而缓解类别不平衡问题。MS-SSIM Loss 则在 SSIM Loss 的基础上考虑了不同尺度上的结构信息,能够衡量抠像结果与原始图像的结构相似性。IoU Loss 则能够衡量模型预测的目标与真实目标之间的重叠程度。由于模型有 5 个边路输出,所以分别计算 5 层的以上 3 种外加混合损失函数共 4 种损失函数以及各损失函数之和,得到 6×4 的数组用于模型训练。

另外,本文采用深监督策略对特征图的混合损失函数之和进行梯度下降训练。这一训练策略使得模型的不同特征提取层级 P1、P2、P3、P4、P5 均可以同时接收来自损失函数的反馈信号,这样可以使模型更快地收敛到较优的解;同时使模型更好地

学习到数据的分布特征,提高模型泛化性。

1.2 alpha 估计网络

对于玻璃、头发等半透明区域前景的 alpha 估计,基于像素相似性的传统网络计算量大,不适合对高分辨率图像进行 alpha 估计。加入上下文信息虽在解决这一问题中略显优势,计算性能下降明显。

对此,本文对基准网络 MGMatting 改进生成 alpha 估计网络,改进方式如下:在基准网络上增加特征提取模块和注意力引导模块组成的上下文注意力引导模块,进一步实现前景和背景融合程度的精确估计。

1.2.1 alpha 估计网络结构

改进的 alpha 估计网络结构如图 3 所示,该网络以原图和粗分割掩码共同作为四通道输入,网络框架采用带有残差块的编解码结构。编码器通过 5 次 Shortcut Block 实现特征提取;解码器则通过 ShortCut Block 对齐编码器特征通道并在上采样之前组合编码器特征。上下文注意力引导^[8] 模块 (Guided Contextual Attention, GCA) 被添加在编码器与解码器中,且在解码器中,渐进细化模块 (Progressive Refinement Module, PRM) 选择性地融合前一级别和当前级别的 alpha 输出,通过一个 ShortCut Block 将原始输入转发到最后一个卷积层,从而更加关注细节纹理特征。

1.2.2 上下文注意力引导模块

上下文注意力引导模块是一种基于相似性的 alpha 值传输方法,由低级图像特征提取模块和用于传播上下文信息的注意力引导模块组成,利用图像的低级特征作为引导,基于相似度进行全局高级 alpha 特征传输。

在 MGMatting 网络结构的编码器和解码器分别加入了上下文注意力引导模块,具体来说,用图像特征提取模块从原图中提取 3×3 的区域作为卷积核,采用卷积计算未知区域和图像特征区域之间的相关性,进而计算每个区域的注意力引导分数,同样从 alpha 特征图提取 3×3 的区域作为卷积核,计算注意力引导分数和 alpha 特征图的反卷积,从而预测不确定区域的 alpha 值。

2 实验过程及其分析

2.1 数据集

为了解决图像前景提取场景数据集丰富度不够的问题,本文生成了场景更加丰富的数据集

AutoCrop,总计 98711 张,数据集组成如表 1 所示。

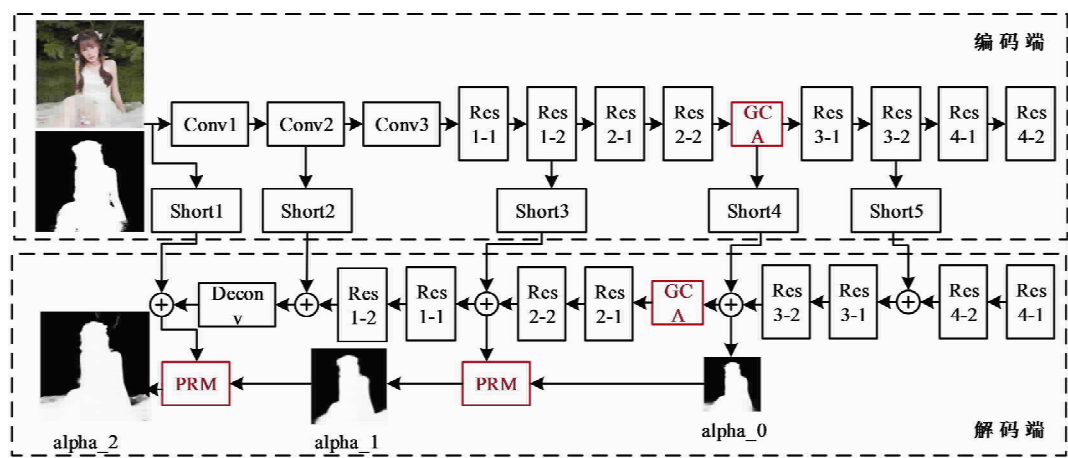


图 3 alpha 估计网络结构

表 1 数据集构成		
数据集名称	原数量(张)	本文用量(张)
AIM-500 ^[9]	95	95
RealWorldPortrait-636 ^[3]	636	636
Human2K ^[10]	2100	2100
Deep AutomaticPortrait Matting ^[11]	2000	20000
PPM-100 ^[12]	100	100
Composition-1K ^[13]	202(训练集) 362(训练集)	20200(合成) 36200(合成)
Distinctions646 ^[14]	596(训练集)	596
自制数据集	18784	18784

以上数据集均为人像抠像数据集。其中, AIM-500 为自然图像,包含多种不同类型前景且有可用的 alpha 图标注;RealWorldPortrait-636 数据集来自真实世界;Human2K 填补了高分辨率人像研究领域的空白;Deep Automatic Portrait Matting 标注精细,且涵盖不同光照条件等;PPM-100 优势在于图像涵盖全身/半身及各种姿势的丰富场景且注释精细;Composition-1K 由于其丰富的物体种类及背景区域成为最通用的抠像检测数据集。

各数据集处理方式如下:对于 Composition-1K、Distinctions646、Human2K 数据集,分别随机从中选取 100 张、100 张、10 张背景图像与 MS COCO 数据集^[15]训练集的人像前景合成,合成工具使用图像处理软件 Adobe Photoshop;其他数据集所有图片均作为本文数据集的组成部分。其中,Deep Automatic matting 数据集为算法生成数据集;另外,从互联网搜集 18784 张人像数据集,调用某公司人像图像前景提取接口生成相应的 alpha 标签,数据样例如图 4 所示,文中称为“自制数据集”。

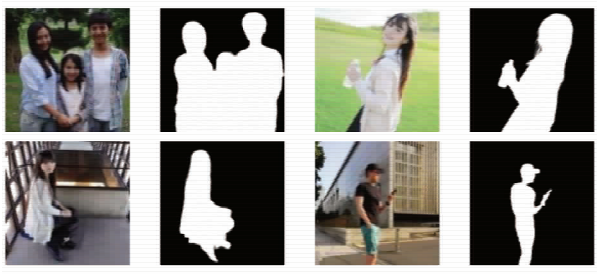


图 4 自制数据集样例

2.2 实验设置

算法借助 Python 语言,采用 PyTorch 深度学习框架实现,实验环境基于 Ubuntu18.04 操作系统,四卡 3090 GPU,内存为 32 GB。实验数据集采用 8 : 1 : 1 比例进行训练集、测试集和验证集划分。显著性目标检测网络采用 Adam 优化器,总共训练 24 个周期。alpha 估计网络采用 Adam 优化器,训练持续 100000 次迭代。其中,显著性目标检测网络在单卡训练耗时 75 h,alpha 估计网络在四卡训练需 33 h。

2.3 评价指标

与 Chen 等^[16]的研究相同,抠像算法常用的评价指标如下:绝对误差(Sum of Absolute Difference, SAD)、均方误差(Mean Square Error, MSE)、梯度误差(Gradient Error, Grad)、连通(Connectivity Error, Coon)以及计算效率(Computational Efficiency)。其中,SAD 与 MSE 考虑单个像素点的透明度遮罩值;Coon 用来体现抠像算法预测的透明度遮罩与真实值之间共有的连通域与周围像素的关系差值;Grad 则更接近人类视觉的主观评价结果。计算效率使用一组图片的抠像算法时间除以总图片个数求出的值。五个指标

皆是值越小,说明预测的准确度越高。

2.4 实验结果与分析

为了公平比较,所有对比算法均利用自制数据集 AutoCrop 重新训练,基于 Composition-1K 测试集进行效果比较。最佳结果加粗表示,次佳结果则下横线表示。

2.4.1 对比实验

这一部分设计实验比较传统图像前景提取算法(KNN)及基于深度学习图像前景提取算法(MG-Matting)的性能差异,还设计实验探索深度学习中 PoolNet 和 U2Net^[17] 两种显著性检测算法的性能差异。

表 2 对比实验结果表

算法	SAD/ 10^{-4}	MSE/ 10^{-4}	Grad/ 10^{-3}	Coon/ 10^{-3}	Computational Efficiency
原 PoolNet+KNN	31.523	61.695	39.970	16.263	2.14
原 PoolNet+原 MGMatting	28.061	51.528	22.154	10.981	2.07
U2Net+原 MGMatting	26.671	54.122	30.672	8.507	1.67
改进 PoolNet+改进 MGMatting	12.329	33.805	13.106	4.326	1.76

2.4.2 消融实验

这一部分我们设计实验验证所提出算法改进思路的有效性。验证 PoolNet 网络改进时,需利用其输出的 Mask 生成三分图输入 KNN 算法比较抠像效果;验证 MGMatting 网络改进时,则使用 PoolNet 算法的输出 Mask 作为先验信息。

前四条数据表明,引入 Hybrid Loss 后,算法在 Grad 和 MSE 指标均有显著改善,抠像结果与原始图像之间差异更小,图像中的边缘和纹理等局部更加平滑,这表明损失函数的优化对显著性检测有积极影响,且这种影响适用于传统/深度学习抠

在 Composition-1K 测试集的测试效果如表 2 所示。前两条数据显示,相较于深度学习抠像算法 MGMatting,传统抠像算法 KNN 在 SAD 和 MSE 方面往往无法达到相同的准确度,而且需要输入辅助的先验信息,不利于自动抠像。第二、三条数据表明与原 PoolNet 相比,U2Net 显著性检测效果更好,SAD 和 MSE 指标表现出色,使得分割结果足够好,而且分割结果与真实标签可以对应起来;但在 Grad 和 MSE 方面略微逊色,导致分割结果的边缘质量较差。改进 PoolNet+改进 MGMatting 组合在 4 个评价指标上都取得了最优,表明该算法在显著性目标检测和抠像任务中具有显著潜力。

像算法,验证了 PoolNet 网络改进的有效性。第三、四条数据中,采用 MGMatting 作为细分模块,性能进一步提高,各项指标都有所改善,表明该算法在前景分割任务中具备一定的优势。第三与第五条数据则表明,在 MGMatting 网络引入上下文注意力引导模块后,SAD、MSE、Grad 和 Coon 等指标都取得了显著改进,证明了这一模块在提高显著性检测效果方面的有效性。改进 PoolNet+改进 MGMatting 组合在所有四个指标上都表现出色,特别是在 SAD 和 Connectivity 方面,表明该算法在显著性目标检测和抠像任务中具有显著潜力。

表 3 消融实验结果表

算法	SAD/ 10^{-4}	MSE/ 10^{-4}	Grad/ 10^{-3}	Coon/ 10^{-3}
原 PoolNet+KNN	31.523	61.695	39.970	16.263
改进 PoolNet +KNN	30.381	54.598	30.107	12.631
原 PoolNet+原 MGMatting	28.061	51.528	22.154	10.981
改进 PoolNet+原 MGMatting	23.343	45.736	18.619	5.344
原 PoolNet+改进 MGMatting	18.273	40.203	17.158	7.237
改进 PoolNet+改进 MGMatting	12.329	33.805	13.106	4.326

2.5 检测效果可视化

为主观展示模型的鲁棒性以及泛化能力,采用 110 张真实场景图片进行效果展示,部分结果如图 5 所示。以手捧花束的女生为例,原 PoolNet 只能检测出人像不能检测出花束,U2Net 则能够检测

出小范围靠近人像的花束,改进 PoolNet 能够检测的花束区域更大,但检测区域连通度较低。改进的 PoolNet 与 MGMatting 分别应用时检测出的区域更加聚集,效果明显提升,全部应用则能够完整提取出手捧花束的人像,其他图像均是如此。



图 5 自制数据集样例

3 结 论

以人像为主要研究对象进行图像前景提取。首先,自制 AutoCrop 数据集对现有抠像数据集进行扩充;再者,针对目前抠像算法人机交互依赖度较强的问题,本文提出基于显著性目标检测的自然场景自动抠像算法。经实验验证,该算法可以有效降低抠像过程中用户工作量,仅需输入单张图像即可自动提取前景,获得精确且稳定的自然场景抠像效果。改进后算法可用于平面广告设计、短视频制作等实际场景,极具商业价值。

由于本文采用显著性目标检测网络和 alpha 估计网络结合的方法, alpha 抠像网络输出结果的好坏依赖于显著性检测网络的输出,因此,后续可以采用精确度更高的显著性检测网络对算法进行优化,提高自动抠像模型的泛化能力。

参考文献

- [1] 苏常保,龚世才. 基于深度学习的人物肖像全抠图算法[J]. 图学学报, 2022, 43(2): 247—253.
- [2] SENGUPTA S, JAYARAM V, CURLESS B, et al. Background matting: the world is your green screen[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2020: 2291—2300.
- [3] YU Q, ZHANG J, ZHANG H, et al. Mask guided matting via progressive refinement network[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2021: 1154—1163.
- [4] CHEN Q, GE T, XU Y, et al. Semantic human matting[C]//Proceedings of the 26th ACM international conference on Multimedia. ACM, 2018: 618—626.
- [5] PORTER T, DUFF T. Compositing digital images[C]//Proceedings of the 11th Annual Conference on Computer

Graphics and Interactive Techniques. ACM, 1984: 253—259.

- [6] LIU J J, HOU Q, CHENG M M, et al. A simple pooling-based design for real-time salient object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2019: 3917—3926.
- [7] QIN X, ZHANG Z, HUANG C, et al. BASnet: boundary-aware salient object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2019: 7479—7489.
- [8] LI Y, LU H. Natural image matting via guided contextual attention[C]//Proceedings of the AAAI Conference on Artificial Intelligence. AAAI, 2020, 34(7): 11450—11457.
- [9] LI J, ZHANG J, TAO D. Deep automatic natural image matting[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 53(2): 273—288.
- [10] LIU Y, XIE J, SHI X, et al. Tripartite information mining and integration for image matting[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE, 2021: 7555—7564.
- [11] SHEN X, TAO X, GAO H, et al. Deep automatic portrait matting[C]//European Conference on Computer Vision. Springer, 2016: 92—107.
- [12] KE Z, SUN J, LI K, et al. Modnet: real-time trimap-free portrait matting via objective decomposition[C]//Proceedings of the AAAI Conference on Artificial Intelligence. AAAI, 2022, 36(1): 1140—1147.
- [13] XU N, PRICE B, COHEN S, et al. Deep image matting[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2017: 2970—2979.
- [14] QIAO Y, LIU Y, YANG X, et al. Attention-guided hierarchical structure aggregation for image matting[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2020: 13676—13685.
- [15] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: common objects in context[C]//European Conference on Computer Vision. Springer, 2014: 740—755.
- [16] CHEN Q, LI D, TANG C K. KNN matting[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(9): 2175—2188.
- [17] QIN X, ZHANG Z, HUANG C, et al. U2-Net: going deeper with nested U-structure for salient object detection[J]. Pattern Recognition, 2020, 106: 107404.