基于注意力机制和多尺度特征的伪装目标检测

蔡俊敏,孙　涵
(南京航空航天大学 计算机科学与技术学院/人工智能学院,江苏 南京 211106)

摘　要:针对伪装目标结构多样、尺度不一和目标边界与其背景具有高度相似性的情况,提出了一种基于注意力机制和多尺度特征的伪装目标检测算法。该算法主要分为两个部分,分别是基于多尺度特征的混合尺度解码器和基于反向注意力机制的注意力引导模块。混合尺度解码器通过级联的特征融合单元,融合高层特征的语义信息与低层特征的空间细节信息,对特征编码器生成的特征金字塔进行解码,得到初步的检测结果;之后引入反向注意力机制,通过擦除图像中已经识别到的目标区域,来引导网络挖掘新的伪装线索,最终得到识别位置更准确、更完整的伪装目标。实验中采用 COD10K 数据集、四种评价指标,与现有的十三种算法进行了对比。实验结果表明,该伪装目标检测算法具有更好的性能表现。

关键词:伪装目标检测;注意力机制;多尺度特征;深度学习;卷积神经网络
中图分类号:TP391　　**文献标识码**:A　　**文章编号**:1673-629X(2023)08-0131-06
doi:10.3969/j.issn.1673-629X.2023.08.019

Camouflaged Object Detection Based on Attention Mechanism and Multi-scale Features

CAI Jun-min, SUN Han
(School of Computer Science and Technology / Artificial Intelligence, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)

Abstract: An algorithm for detecting camouflaged objects based on attention mechanism and multi-scale features is proposed for the situation that camouflaged objects have diverse structures, different scales and the object boundaries are highly similar to their backgrounds. The proposed algorithm is mainly divided into two parts, which are a mixed-scale decoder based on multi-scale features and an attention-guiding module based on the reverse attention mechanism. The mixed-scale decoder fuses the semantic information of high-level features with the spatial detail information of low-level features through a cascaded feature fusion unit to decode the feature pyramid generated by the feature encoder and obtain the preliminary detection results. After that, the reverse attention mechanism is introduced to guide the network to mine new camouflage cues by erasing the already recognized object regions in the image, and finally obtain a more accurate and complete camouflage object. The COD10K dataset and four evaluation metrics are used in the experiments, and the comparison is conducted with thirteen existing algorithms. The experimental results show that the proposed algorithm has better performance.

Key words: camouflaged object detection; attention mechanism; multi-scale feature; deep learning; convolutional neural network

0　引　言

在计算机视觉领域,目标检测一直是一个非常热门的课题,各种各样的目标检测模型层出不穷,并不断刷新各个性能榜单。在 Zhao 等人的研究[1]之后,可以将目标检测大致分为三类,分别是:显著性目标检测、通用目标检测和伪装目标检测。显著性目标检测旨在识别图像中最引人注目的目标,并对它们的轮廓进行分割。通用目标检测往往伴随着语义分割或者全景分割等任务,需要在识别图像中目标对应的区域,并且为之分配可能的标签和相应的分数。而伪装目标检测则要求识别图像中被隐藏的目标。伪装目标是指目标自身的形状、纹理或者颜色特征等特性导致其与周围的背景相近的物体。其中伪装目标检测由于目标和其背景具有高度相似性,所以检测起来更加困难。

伪装图像大致可分为两类,天然伪装和人为伪装的图像。昆虫、头足类等动物的天然伪装是一种可以

收稿日期:2022-09-15　　**修回日期**:2023-01-16
基金项目:中央高校基本科研业务费专项资金资助项目(NZ2019009)
作者简介:蔡俊敏(1999-),男,硕士研究生,CCF 会员(L3554G),研究方向为计算机视觉;通讯作者:孙　涵,博士,副教授,CCF 会员(33361S),研究方向为图像处理、计算机视觉。

避免被天敌察觉的生存技巧。而人为伪装通常被应用于视频检测过程;还会出现在产品制造的时候(即产品瑕疵检测);也可以用于游戏或艺术中帮助隐藏信息。

与类别相关的语义分割任务不同,伪装目标检测任务与类别无关。伪装目标检测的任务简单且易于定义。给定一张图像,该任务需要一个伪装目标检测算法来为每个像素 i 分配一个置信度 $\text{Label}_i \in \{0,1\}$,其中 Label_i 表示像素 i 的概率值。0 表示该像素不属于伪装目标,而 1 表示该像素完全属于伪装目标。

在伪装目标检测领域,如何利用提取到的特征来区分伪装目标和背景是一个至关重要的问题。为此,文中提出了基于注意力机制和多尺度特征的伪装目标检测算法,该算法主要由两部分组成:混合尺度解码器(MSD)和注意力引导模块(AG)。混合尺度解码器对多尺度特征进行解码得到伪装目标的初步检测结果,之后引入反向注意力机制得到最终的伪装目标检测结果。

该文的主要贡献如下:

(1)提出了基于多尺度特征的伪装目标检测算法,通过提取到的多尺度特征有效区分伪装目标和背景。

(2)设计了混合尺度解码器和注意力引导模块,通过级联的特征融合单元对多尺度特征进行解码,之后引入反向注意力机制得到最终的伪装目标检测结果。

(3)在 COD10K 这一数据集上与十三个经典的深度学习方法进行比较,证明了该方法的有效性。

1 相关工作

1.1 传统的伪装目标检测算法

在传统的伪装目标检测算法中,大部分方法都是基于图像的低级特征以及一些人为设计的特征(例如图像的纹理、色彩、亮度、强度)。Galun 等人[2]以纹理分割技术为基础来检测伪装物体。鲜晓东等人[3]借助图像的颜色和纹理信息实现伪装目标检测。周静等人[4]提出基于光流场分割的伪装运动目标检测方法,面对更加复杂的场景,这些方法在性能和泛化性上还有待提高。

1.2 基于神经网络的伪装目标检测算法

近年来,卷积神经网络的发展给目标检测带来了很大的提升。基于神经网络的深度学习方法主要分为三类:结合上下文信息的伪装目标检测模型、结合注意力机制的伪装目标检测模型和结合边缘信息的伪装目标检测模型。

大多数模型致力于挖掘图像特征的上下文信息以提升模型的性能。Le 等人[5]提出了一个端到端的神经网络——Anabranch Network。该网络结合了多任务的学习框架,将图片分类和图像分割整合到一个模型中。Zheng 等人[6]提出了一个密集的反卷积网络。为了能够有效地提取高级特征中包含的丰富的语义信息,该网络采用短连接的方式融合多尺度的高级特征。Fan 等人[7]受动物的捕食过程(先发现猎物,再确定猎物的具体位置)所启发,提出了 SINet。SINet 由搜索模块(SM)和识别模块(IM)组成,搜索模块通过多个感受野(RF)组件来模仿人眼感知系统的感受野。

近些年来,引入注意力机制来提升伪装目标检测性能的方法也有很多。在显著性目标检测方面,Zhao 等人[8]提出的 PFAN 模型认为分辨率较大通道数较少的低级特征图保有丰富的细节和结构信息,适合采用空间注意力加以过滤不需要的空间信息,而分辨率较小通道数较多的高级特征图具有丰富的语义特征,适合采用通道注意力来选取有效的语义信息。Chen 等人[9]提出的 RANet 则设计了反注意力模块,通过擦除每个侧输出特征中的当前预测区域来引导整个网络顺序发现互补对象区域及细节。

将这些方法迁移到伪装目标检测方向,Sun 等人[10]的 C2F-Net 借助注意力引导的跨级融合模块,将多级特征与信息注意系数相结合。

对于边缘信息,许多研究证明了边缘检测在一定程度上可以促进伪装目标检测性能的进步。在显著性目标检测方面,Zhao 等人[11]提出的 EGNet 将提取的显著目标结构特征和显著目标边缘特征两个部分相结合,来提升显著目标的检测效果。伪装目标检测作为与显著性目标检测相似的任务,同样可以借鉴。Ji 等人[12]提出的 ERRNet 利用了边缘信息,旨在对生物的视觉感知系统进行建模并实现有效的边缘先验和潜在伪装区域与背景之间的交叉比较。

2 网络结构

2.1 模型网络结构

网络的整体结构如图 1 所示。主要由两部分组成:混合尺度解码器和注意力引导模块。

2.2 混合尺度解码器

对于输入图像,借助由特征提取网络和通道压缩模块组成的特征编码器帮助提取图像的多层特征。之后,文中构造了基于多尺度特征的混合尺度解码器,以自上而下的形式整合多级特征。混合尺度解码器由多个级联的特征融合单元组成,特征融合单元的结构如图 2 所示。特征融合单元通过分组迭代增强不同通道之间的信息交互,从而提升网络的性能表现。特征融合单元的输入 $\hat{f}_k$ 定义如下:

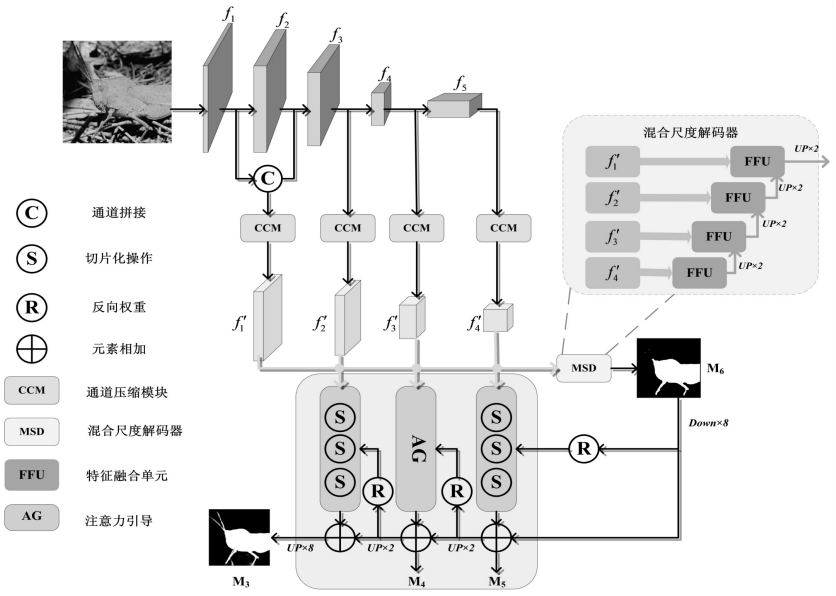

图 1　网络结构

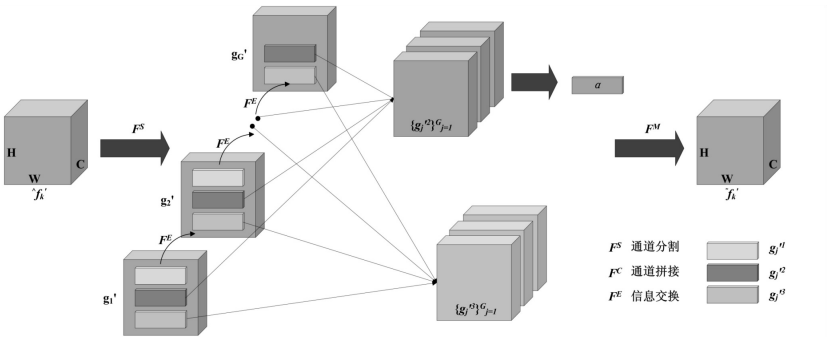

图 2　特征融合单元

$$\hat{f}_k = f_k + \mu(\hat{f}_{k+1}) \tag{1}$$

式中,$\hat{f}_{k+1}$ 表示深层融合单元的输入,$\mu(\cdot)$ 表示特征融合操作。

在特征融合单元中,该文先采用 1×1 卷积来扩展特征图 $\hat{f}_k$ 的通道数。然后沿通道维度将特征划分为 G 个组 $\{g_j\}_{j=1}^{G}$。组之间的特征交互以迭代的方式进行。

通过四个级联的特征融合单元和几个卷积层,组成基于多尺度特征的混合尺度解码器,得到一个单通道的映射图。最后通过 sigmoid 函数得到初步的检测结果 M_6 并输出。

2.3　注意力引导模块

在得到伪装目标的粗略位置之后,设计了基于反向注意力机制的引导模块,通过擦除前景目标的方式逐步挖掘伪装区域。为了节省计算资源,该文选择对三个高层特征的输出分支进行引导。如图 1 所示,从解码器得到的粗略伪装图 M_6 开始,通过从旁路输出特征中擦除当前预测的伪装区域,引导整个网络逐步地发掘补充的目标区域和相关细节。首先对得到的初步伪装图像 M_6 做计算得到反向注意力权重 A_k,然后借助该权重引导网络挖掘有效的伪装区域。受文献[13]启发,文中对特征进行切片化处理,能够更加高效地利用之前得到的反向注意力权重。如图 3 所示,切片化处理的具体过程可以通过以下的式子表示:

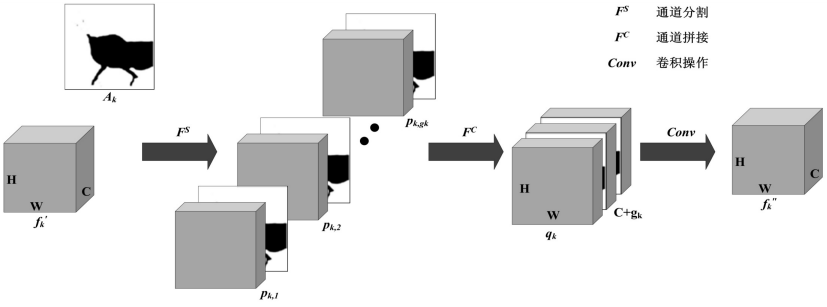

图 3　切片化处理

$$\text{stepI}: F^S(f_k') \rightarrow \{p_{(k,1)}, \cdots, p_{(k,j)}, \cdots, p_{(k,g_k)}\}$$
$$\text{stepII}: F^C(\{p_{(k,1)}, A_k\}, \cdots, \{p_{(k,g_k)}, A_k\}) \rightarrow q_k \quad (2)$$

式中，F^S、F^C 分别表示通道分割和通道连接函数，q_k 表示加入反向注意力引导之后的特征。

之前的研究[14-15]表明，多阶段的细化可以提高模型的性能。因此，文中组合了多个反向注意力模块，来引导网络学习不同的特征金字塔，逐步细化粗略的预测。每个注意力引导模块，均由三个切片化处理组成。通过三个注意力引导模块由深到浅，分别对特征编码器提取到的三组高级特征 $f_k', k \in \{3,4,5\}$ 进行引导，最后得到最终的伪装目标检测结果图 M_3。

3 实验结果与分析

3.1 实验数据集

该文提出的算法所采用的训练集为 COD10K+CAMO，训练完成之后在 COD10K 的测试集上做测试。CAMO 数据集是 Le 等人[5]提出的伪装目标数据集，一共包含 2 500 张图像，其中 2 000 张图作为训练集，500 张图像作为测试集，涵盖了八个类别。COD10K 是 Fan 等人[7]提出的一个大型伪装目标数据集，一共包含 10 000 张图像，其中 6 000 张图像作为训练集，4 000 张作为测试集。这 10 000 张图像划分为 10 个超类和 78 个子类，包括水生、飞行、两栖和陆地等等。此外，该数据集还对伪装图像提供了丰富的标签，包括目标类别、边界框、对象级标注、实例级标注和具有挑战性的属性，如图 4 所示。

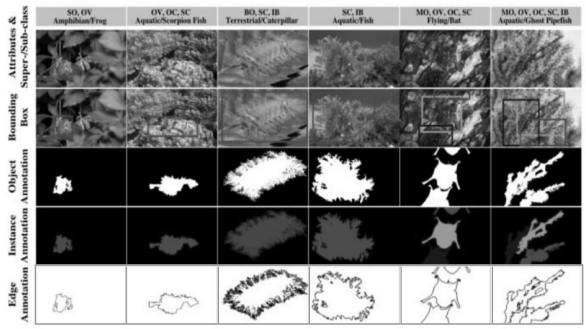

图 4 COD10K 数据集示例

3.2 实验细节及超参数设置

文中提出的算法以 Res2Net50[16]为骨干网络，并使用在 ImageNet 训练好的权重进行初始化。模型的训练集采用 CAMO+COD10K，一共是 4 040 张图像。在模型输入端，统一将输入图像调整为 352×352。在模型训练过程中，使用 Adam 优化器进行训练。批处理大小设置为 32，学习率从 $1e^{-4}$ 开始，每 50 个 epoch 除以 10。整个训练过程共有 100 个 epoch。GPU 为 Tesla V100。

3.3 性能比较

本节将所提算法与 13 个经典的深度学习模型进行性能比较，包括 FPN[17]、MaskRCNN[18]、PSPNet[19]、Unet + +[20]、PiCANet[21]、MSRCNN[22]、PFANet[8]、CPD[23]、HTC[24]、EGNet[11]、PraNet[25]、SINet[7] 和 SINet-V2[13]。由于伪装目标检测是一个新兴的领域，因此部分深度学习模型的原本目的是应用于显著性目标检测，其中 PraNet、SINet 和 SINet-V2 是直接针对伪装目标检测的模型。关于这些对比模型的性能指标，主要来自文献[13]。

3.3.1 定量分析

本节采用平均绝对误差（M）、E-measure 分数（E_φ）、S-measure 分数（S_α）和加权的 F-measure 分数（F_β^ω）四个数值性的评估指标，在 COD10K 测试集上与十三个经典的深度学习方法进行性能比较。其比较结果如表 1 所示，最好的结果以粗体显示。从表中可以看出，所提算法除了在 S-measure 分数上略低于 SINet-V2，其他指标相比较这 13 个经典模型均取得了最好的性能。

表 1 在 COD10K 测试集上的性能比较

基准模型	$S_\alpha \uparrow$	$E_\varphi \uparrow$	$F_\beta^\omega \uparrow$	$M \downarrow$
FPN	0.697	0.691	0.411	0.075
MaskRCNN	0.613	0.748	0.402	0.080
PSPNet	0.678	0.680	0.377	0.080
Unet++	0.623	0.672	0.350	0.086
PiCANet	0.649	0.643	0.322	0.090
MSRCNN	0.641	0.706	0.419	0.073
PFANet	0.636	0.618	0.286	0.128
CPD	0.747	0.770	0.508	0.059
HTC	0.548	0.520	0.221	0.088
EGNet	0.737	0.779	0.509	0.056
PraNet	0.789	0.861	0.629	0.045
SINet	0.771	0.806	0.551	0.051
SINet-V2	**0.815**	0.887	0.680	0.037
Ours	0.807	**0.897**	**0.704**	**0.032**

尤其是与 SINet-V2（目前最好的伪装目标检测模型）相比，E-measure、F-measure 分别增长了 0.010（1.1%）、0.024（3.5%）。MAE 下降了 0.005（13.5%）。

3.3.2 定性分析

为了更直观地比较所提模型和其他经典的深度学习方法，文中还进行了一系列和其他经典深度学习方法的视觉对比实验，并提供可视化展示。由于篇幅有限，本节只列举了在 COD10K 测试集上同 SINet、SINet

-V2(分别是 2020 年的 SOTA 方法和 2021 年的 SOTA 方法)两个经典模型的比较,这两个模型的预测图是根据开源的代码重新训练和测试生成的,如图 5 所示。

图 5　与两个经典模型的可视化比较

从整体来看,文中所提出的伪装目标检测算法相比这两个经典方法识别效果更佳,识别出的伪装目标更加准确,更加完整,更符合对应的真值图像。具体地说,从前三列图像中可以看出文中提出的模型输出的检测结果几乎与对应的真值图像完全一致,

而另外两个模型对于伪装目标的检测效果都有一定的缺失,对伪装区域均存在误判的情况。对于第四、五、六列的输入图像,尽管 SINet 和 SINet-V2 也能识别到大致的伪装区域,但伪装目标的完整性以及边界细节均不如文中所提出的算法。总的来说,无论是在整体的目标区域还是边界细节方面,文中所提出的算法相比较其他的模型均能更准确地识别到伪装目标。

另外,针对 Flying 子集和 Amphibian 子集中普遍存在的小目标场景(第七、八、九、十列),文中提出的算法能够精确地识别出伪装目标,而另外两个模型都存在被图像中的其他目标所干扰的情况,说明文中提出的混合尺度编码器能够有效地组合多尺度特征,提升模型对较小伪装目标的识别效果。

3.4　消融实验

本节进行消融实验,通过分离各个子模块,分别验证文中所提模型中各个模块的有效性。具体结果如表 2 所示。其中 Baseline 表示使用相邻连接解码器对特征编码器进行解码,+MSD 表示仅加入混合尺度解码器,+AG 表示仅加入注意力引导模块,MSD+AG 表示既加入混合尺度解码器,也加入注意力引导模块。实验结果表明,文中提出的混合尺度解码器和注意力引导模块可以有效提高模型的检测精度。尤其是注意力引导模块,通过引入反向注意力机制,可以引导网络有效地挖掘潜在的伪装区域,从而提升模型对伪装目标的检测效果。

另外,还探究了注意力引导模块中多阶段细化中不同阶段切片化大小对模型效果的影响,主要探究不同阶段基于统一策略和基于渐进式策略的方法。{ * , * , * }表示从第一个切片化处理到最后一个切片化处理时的不同切片大小,如{32,8,1}表示三个阶段的切片化处理分别将候选特征 p_i 沿通道维度分成 32 片、8 片和 1 片。如表 3 所示,相比较其他方案,基于渐进式策略的方法性能更优。因此,在文中的其他对比实验中,均采用渐进式策略的方法。

表 2　在 COD10K 测试集上的消融实验

模型	$S_\alpha\uparrow$	$E_\varphi\uparrow$	$F_\beta^\omega\uparrow$	$M\downarrow$
Baseline	0.787	0.862	0.646	0.043
+MSD	0.794	0.873	0.653	0.040
+AG	0.803	0.889	0.694	0.036
MSD+AG	**0.807**	**0.897**	**0.704**	**0.032**

表 3　注意力引导模块中不同切片化策略的影响

	$S_\alpha\uparrow$	$E_\varphi\uparrow$	$F_\beta^\omega\uparrow$	$M\downarrow$
{1,1,1}	0.806	0.896	0.695	0.034
{8,8,8}	0.806	0.896	0.699	0.034
{32,32,32}	0.806	**0.899**	0.701	0.032
{1,8,32}	0.807	**0.899**	0.701	0.033
{32,8,1}	**0.807**	0.897	**0.704**	**0.032**

4　结束语

针对伪装目标结构多样、尺度不一和目标边界与其背景具有高度相似性的情况,提出了一种基于注意力机制和多尺度特征的伪装目标检测算法,通过混合尺度解码器和反向注意力模块,提升了模型的检测性能。借助四个评估指标将文中算法与现有的十三种算法在 COD10K 数据集上进行测试,结果表明文中算法具有更好的性能,可获得识别位置更准确、边界细节更完善的伪装目标检测结果。

参考文献:

[1] ZHAO Z Q, ZHENG P, XU S, et al. Object detection with deep learning: a review[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(11): 3212-3232.

[2] GALUN M, SHARON E, BASRI R, et al. Texture segmentation by multiscale aggregation of filter responses and shape elements[C]//IEEE international conference on computer vision. Nice: IEEE, 2003: 716.

[3] 鲜晓东,李克文. 基于颜色和纹理特征的伪装色矿工目标检测[J]. 计算机应用,2013,33(2):539-542.

[4] 周　静,窦一民,李金屏. 基于光流场分割的伪装色运动目标检测[J]. 济南大学学报:自然科学版,2020,34(4):328-334.

[5] LE T N, NGUYEN T V, NIE Z, et al. Anabranch network for camouflaged object segmentation[J]. Computer Vision and Image Understanding, 2019, 184: 45-56.

[6] ZHENG Y,ZHANG X,WANG F,et al. Detection of people with camouflage pattern via dense deconvolution network [J]. IEEE Signal Processing Letters,2018,26(1):29-33.

[7] FAN D P,JI G P,SUN G,et al. Camouflaged object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [s. l.]: IEEE, 2020:2777-2787.

[8] ZHAO T,WU X. Pyramid feature attention network for saliency detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Long Beach:IEEE,2019:3085-3094.

[9] CHEN Z,XU Q,CONG R,et al. Global context-aware progressive aggregation network for salient object detection [C]//Proceedings of the AAAI conference on artificial intelligence. Menlo Park:AAAI,2020:10599-10606.

[10] SUN Y, CHEN G,ZHOU T,et al. Context-aware cross-level fusion network for camouflaged object detection[J]. arXiv:2105.12555,2021.

[11] ZHAO J X, LIU J J,FAN D P,et al. EGNet:edge guidance network for salient object detection[C]//Proceedings of the IEEE/CVF international conference on computer vision. Long Beach:IEEE,2019:8779-8788.

[12] JI G P,ZHU L, ZHUGE M,et al. Fast camouflaged object detection via edge-based reversible re-calibration network [J]. Pattern Recognition,2022,123:108414.

[13] FAN D P,JI G P, CHENG M M,et al. Concealed object detection[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2022,44(10):6024-6042.

[14] 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. Long Beach:IEEE,2019:7479-7489.

[15] XU N, PRICE B, COHEN S, et al. Deep image matting [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu: IEEE, 2017: 2970-2979.

[16] GAO S H,CHENG M M, ZHAO K,et al. Res2net:a new multi-scale backbone architecture[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2019,43(2):652-662.

[17] LIN T Y,DOLLáR P, GIRSHICK R,et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu:IEEE,2017:2117-2125.

[18] HE K,GKIOXARI G, DOLLáR P,et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. Menlo Park:AAAI,2017:2961-2969.

[19] ZHAO H,SHI J,QI X, et al. Pyramid scene parsing network [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu: IEEE, 2017: 2881-2890.

[20] ZHOU Z,RAHMAN SIDDIQUEE M M, TAJBAKHSH N, et al. Unet++:a nested u-net architecture for medical image segmentation[M]//Deep learning in medical image analysis and multimodal learning for clinical decision support. Berlin: Springer,2018:3-11.

[21] LIU N,HAN J, YANG M H. Picanet:learning pixel-wise contextual attention for saliency detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [s. l.]:IEEE,2018:3089-3098.

[22] HUANG Z, HUANG L,GONG Y,et al. Mask scoring r-cnn [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Long Beach: IEEE 2019: 6409-6418.

[23] WU Z,SU L, HUANG Q. Cascaded partial decoder for fast and accurate salient object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Long Beach:IEEE,2019:3907-3916.

[24] CHEN K, PANG J,WANG J,et al. Hybrid task cascade for instance segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Long Beach:IEEE,2019:4974-4983.

[25] FAN D P, JI G P,ZHOU T,et al. Pranet:parallel reverse attention network for polyp segmentation[C]//International conference on medical image computing and computer-assisted intervention. [s. l.]:Springer,2020:263-273.