基于密集多尺度自注意力变换网络的伪装对象分割方法

闫馨宇[1,2,3], 孙美君[1,2,3)*], 韩亚洪[1,2], 王征[1,2,3]

[1] (天津大学智能与计算学部　天津　300350)
[2] (天津机器学习重点实验室　天津　300350)
[3] (中国教育部城市智能化与数字化治理工程研究中心　天津　300350)
(meijunsun@tju.edu.cn)

摘　要: 为了充分地发挥深度自注意力变换网络在伪装对象分割任务中的潜力, 提出一种基于密集多尺度自注意力变换网络的伪装对象分割方法, 包含双分支可分离密集多尺度特征提取和快速注意力诱导的跨级交互融合 2 个模块. 首先以自注意力变换网络作为骨干特征提取器获取各级特征; 然后将提取的特征馈送到双分支可分离密集多尺度特征提取模块, 在局部分支和全局分支中, 利用密集递进相连的深度可分离卷积块提取丰富的多尺度上下文特征; 最后使用快速注意力诱导的跨级交互融合模块融合各级特征, 并利用每级融合特征预测伪装映射, 通过深度监督让各级特征在空间上保持高度一致性, 尽可能地集中注意力于伪装特征而避免背景噪声的干扰. 在 CHAMELEON, CAMO, COD10K 和 NC4K 这 4 个基准数据集上, 与其他 28 种主流模型进行定性可视化对比, 以及针对 PR 曲线、*S* 值、*F* 值、*E* 值及 MAE 这 5 种评价指标的定量对比实验结果表明, 所提出的基于密集多尺度自注意力变换网络是一种有效的伪装对象分割模型.

关键词: 伪装对象分割; 自注意力变换网络; 深度学习
中图法分类号: TP391.41　　**DOI:** 10.3724/SP.J.1089.2024.19950

Camouflaged Object Segmentation Based on Dense Multi-Scale Transformer

Yan Xinyu[1,2,3], Sun Meijun[1,2,3)*], Han Yahong[1,2], and Wang Zheng[1,2,3]

[1] (*College of Intelligence and Computing, Tianjin University, Tianjin*　300350)
[2] (*Tianjin Key Laboratory of Machine Learning, Tianjin*　300350)
[3] (*Engineering Research Center of City Intelligence and Digital Governance, Ministry of Education of the People's Republic of China, Tianjin* 300350)

Abstract: In order to make deep self-attentive transform networks reach their full potential in the task of camouflaged object segmentation, a camouflaged object segmentation method called dense multi-scale Transformer is proposed. It consists of two main modules: two-branch separable dense multi-scale feature extraction module and fast attention-induced cross-level interaction fusion module. First, the Transformer is used as the backbone feature extractor to acquire features at each level. Second, these features are fed to a two-branch separable dense multi-scale feature extraction module, and rich multi-scale contextual features are extracted using dense recursively connected deep separable convolutional blocks in local and global branches. Finally, the fast attention-inducing cross-level interaction fusion module is used to fuse features at all levels. Each level of fused features is used to predict the camouflaged maps, and the features at each level are made highly spatially consistent by deep supervision, focusing attention on the camouflaged features as

收稿日期: 2022-09-06; 修回日期: 2023-03-01. **基金项目**: 国家自然科学基金(62076180, 61876125). **闫馨宇**(1996—), 女, 博士研究生, 主要研究方向为显著性对象检测、伪装对象分割; **孙美君**(1978—), 女, 博士, 教授, 博士生导师, 论文通信作者, 主要研究方向为计算机图形学、高光谱成像、图像处理; **韩亚洪**(1977—), 男, 博士, 教授, 博士生导师, 主要研究方向为多媒体分析、机器学习、计算机视觉; **王征**(1980—), 男, 博士, 教授, 硕士生导师, 主要研究方向为视频分析、高光谱成像、计算机图形学.

much as possible while avoiding the interference of background noise. Qualitative visualization and quantitative comparison experiments (in terms of five evaluation metrics: PR curve, *S*-measure, *F*-measure, *E*-measure and MAE) with 28 existing mainstream methods on four benchmark datasets, namely CHAMELEON, CAMO, COD10K and NC4K, demonstrate that the proposed dense multi-scale Transformer is an effective model for camouflaged object segmentation.

Key words: camouflaged object segmentation; self-attentive transformer; deep learning

伪装对象指与背景的颜色及纹理高度一致、混淆人眼视觉机制从而难以被发现的对象[1], 可以分为天然伪装对象和人工伪装对象. 天然伪装对象一般指利用地形地貌、环境中的物体或自身的形态、颜色等天然因素躲避天敌的对象[2], 如沙漠中的蜥蜴、深海中的鱼类、庄稼地中的叶蝗虫等. 人工伪装对象则是人类通过特殊着色等技术降低可探测特征的对象, 如身着迷彩的士兵、特殊喷漆的战机等. 伪装对象分割(camouflaged object detection, COD)任务是将伪装于背景中的对象识别并分割出来, 在军事安全、工业探伤[3]、检测农业害虫[4]及医学成像[5-6]等现实场景中有着巨大的应用潜力.

近年来, 随着 Fan 等[7]构建了第 1 个大规模伪装对象分割数据集 COD10K, COD 吸引了越来越多研究人员的关注. 当前, 已有的伪装对象分割方法[7-20]以卷积神经网络为主, 通常采用编码器-解码器结构; 其中, 编码器利用在大型数据集上预训练得到的卷积神经网络将输入图像编码为多级特征, 解码器将提取的特征集成以预测最终的二值伪装映射图. 然而, 卷积运算虽然善于提取局部特征, 但对于全局依赖关系的建模能力较弱. 在 COD 中, 由于伪装对象与其周围背景的局部高度相似性, 全局依赖关系是非常重要的[8,21]. 单纯利用卷积运算在局部滑动窗口中提取特征不仅不利于识破伪装, 反而更加容易被局部高度相似的伪装对象和背景所迷惑. 尽管一些方法利用全局池化层[10-11]和非本地模块[8]来整合全局上下文, 但它们仅在某些层中这样做, 基于卷积操作的标准体系结构并没有改变.

在过去几年中, 诞生于自然语言处理领域的深度自注意力变换网络[22]由于其强大的全局关系建模能力, 在越来越多的计算机视觉任务中代替了卷积神经网络, 并得到飞跃式的发展, 其核心思想就是利用自注意机制捕获全局上下文信息, 从而建立长距离依赖关系. 为了更好地适应视觉任务中对图像的处理, 减少由密集注意力导致的过多的内存和计算成本, 多种层级式结构的视觉深度自注意力变换网络逐步涌现, 如 PVT[23], Swin Transformer[24], MPViT[25]等, 并已用于图像分类、对象检测、语义分割等任务中, 取得了比传统堆叠卷积操作的网络更好的性能. 因此, 将深度自注意力变换网络引入 COD, 利用全局线索提高 COD 的性能是非常有必要的.

鉴于任务本身的复杂性, 将视觉深度自注意力变换网络应用于 COD 并非是简单而直接的, 其存在 2 大问题: (1) 伪装对象的大小形状存在极高的可变性, 自注意力变换网络中有限的特征感受野[25]不可避免地限制了模型对伪装物体的识别能力. 层级式结构的视觉深度自注意力变换网络, 如 PVT[23], Swin Transformer[24], MPViT[25]等, 虽然更加适合于视觉任务中对图像的处理, 减少了由密集注意力导致的内存和计算成本问题, 但在 PVT[23]或 Swin Transformer[24]中, 更加专注于降低自注意力的二次复杂度, 采用的与数据无关的稀疏注意力可能会限制对远程关系的建模能力, 同时也较少关注构建有效的多尺度表示. MPViT[25]则通过对图像不同尺度分块及其构成的多路径结构, 在一定程度上丰富了网络中有效的多尺度特征表示. 然而, 与类别已知的语义分割任务不同, 在 COD 中, 伪装对象的类别、大小、形状高度可变, 仅仅利用其各个层级输出的特征远远无法满足其对丰富感受野的需求. (2) 伪装对象通常与背景之间存在高度一致性, 自注意力变换网络提取的各级特征很可能会受到超出伪装区域的无关部分的影响, 如何在整合各级特征时集中注意力于伪装特征, 尽可能地避免背景噪声的干扰, 仍有待研究.

针对以上 2 大问题, 本文提出一种基于密集多尺度自注意力变换网络的伪装对象分割(dense multi-scale Transformer for camouflaged object detection, DMT-COD)方法, 极大地改善了现有的 COD 性

能, 它包含双分支可分离密集多尺度特征提取模块(two-branch separable dense multi-scale feature extraction module, TSDMM)和快速注意力诱导的跨级交互融合模块(fast attention-induced cross-level interaction fusion module, FACIFM)这 2 个关键模块. 首先以层级式结构的视觉深度自注意力变换网络作为骨干特征提取器, 获取包含全局依赖关系的各级特征; 为了丰富各级特征有限的有效感受野, 将所得的特征馈送到 TSDMM, 在局部分支和全局分支中利用密集递进相连的深度可分离卷积块提取丰富的多尺度上下文特征; 最后使用FACIFM整合各级特征, 让各级特征最大限度地将注意力集中于伪装对象区域, 消除超出伪装区域的无关部分的干扰, 获得有关伪装目标对象的更纯净的表示. 此外, 利用每级特征表示预测伪装映射, 通过深度监督让各级特征在空间上保持高度一致性. 在 4 个广泛使用的 COD 数据集上进行的大量实验结果表明, 本文提出的基于密集多尺度自注意力变换网络是一种有效的 COD 模型, 并且明显优于 28 个已有的主流模型.

1 相关工作

1.1 伪装对象检测

1.1.1 基于传统方法的 COD

COD 是近 2 年逐渐兴起的研究方向. 最初的非深度学习的方法利用手工提取的低级特征区分前景和背景, 如纹理[26]、3D 凹凸性[27]、运动[28]等, 这些方法只适用于极少数简单的情况, 在复杂情况下往往会失败.

1.1.2 基于深度学习方法的 COD

随着深度学习方法广泛应用于视觉任务, Le 等[29]提出联合图像分类和 COD 的框架; Yan 等[15]引入对抗性攻击, 以提高 COD 精度; Fan 等[7]构建了当前最大的伪装对象数据集 COD10K, 正式开启了COD在深度学习时代的发展, 并开发了SINet用于逐步定位和搜索伪装物体; Mei 等[8]则模拟自然界动物的捕食过程, 提出仿生物学架构——定位聚焦网络, 并开发了一种分心挖掘策略, 用于分心地发现和移除伪装对象; 为了解决前景和背景难以分割的问题, Zhai 等[30]设计一种互图学习模型, 将传统的互图学习思想从规则网格推广到图域; Li 等[14]利用显著性对象与伪装对象的相反关系, 提出一种利用矛盾信息增强显著性对象检测和伪装目标分割能力的范式; Lv 等[9]拓宽了 COD 的概念, 提出一种基于排序的伪装对象检测网络, 可以同时对伪装目标进行定位、分割和排序; Chen 等[13]针对外观的多变性以及与环境的低对比度, 提出一种新颖的上下文感知跨级别融合网络(C2F-Net); Chen 等[20]则提出一种新颖的边界引导网络, 首先通过一个定位模块, 利用局部细节信息和全局上下文信息推断伪装对象的初始位置, 再利用边界引导融合模块探索伪装区域与其边界之间的互补关系. 上述基于深度学习的方法都利用在大型数据集上预训练得到的卷积神经网络作为底层特征提取器, 再利用解码器对这些特征进行处理, 进而识别伪装对象, 在一定程度上提高了 COD 的性能. 近 2 年出现并兴起的深度自注意力变换网络在多种计算机视觉任务中成为新范式, 本文将深度自注意力变换网络引入 COD, 进一步提高该任务的性能.

1.2 视觉深度自注意力变换网络

深度自注意力变换网络最早由 Vaswani 等[22]提出用于机器翻译, 已成为许多自然语言处理任务中的新范式. 为了适用于计算机视觉任务, 深度自注意力变换网络做出了一系列的改动, 并取得了出色的效果, 已应用在各种各样的视觉任务中, 如对象检测[24-25,31-32]、语义分割[23-25,32-33]等. 最先获得成功的 ViT[31]模拟自然语言处理任务中的输入形式, 将每幅图像分为一系列的 2D 补丁, 并送入深度自注意力变换网络中进行图像分类. 受此启发, 一些工作将这种分类模型扩展为密集预测任务的骨干特征提取网络. SETR[33]直接将 ViT 作为提取特征的编码器, 解码器则是几个简单的卷积层聚合来自编码器中特征, 并将特征映射上采样到原图大小进行语义分割; Wang 等[23]提出一种金字塔架构来适应 ViT 的密集预测任务, 并提出空间减少注意力层, 在一定程度上缓解了自注意力的二次复杂度; T2T-ViT[32]则通过 T2T 模块对局部结构进行建模, 并利用更高效的深度自注意力变换架构设计丰富的中间特征, 从而提升识别性能; Liu 等[24]提出 Swin Transformer 网络, 通过移动窗口计算自注意力, 大大降低了序列的长度, 在一定程度上减少了由密集注意力导致的内存和计算成本. 本文采用 MPViT[25]作为骨干特征提取网络, 并提出一种新颖的 TSDMM 和 FACIFM, 在一定程度上解决了 COD 对于丰富有效感受野的需求, 以及深度自注意力变换网络提取的各级特征受到超出伪装区域的无关部分影响的问题.

2　DMT-COD 方法

近年来, 越来越多的视觉任务利用更善于捕获全局关系的深度自注意力变换网络代替卷积神经网络, 并取得了一定的性能提升. 然而, 如何让深度自注意力变换网络在 COD 中充分发挥潜力仍是一个有待解决的问题. 受此启发, 本文提出一种 DMT-COD 方法, 它以自注意力变换网络为骨干特征提取器获取包含全局依赖关系的各级特征, 并通过 TSDMM 和 FACIFM, 极大程度地丰富了特征的有效感受野, 有效地解决了深度自注意力变换网络提取的各级特征受到超出伪装区域的无关部分影响的问题.

2.1　总体结构

图 1 所示为 DMT-COD 方法的整体结构. 给定单幅 RGB 图像, 首先将其输入 MPViT[25]这一基于自注意力变换网络的骨干特征提取器中, 该提取器包含 4 个层级模块(详见文献[25]), 依次得到多级特征 $\boldsymbol{f}_i\left(i\in[1,2,3,4]\right)$; 然后将 TSDMM_i, $i\in[1,2,3,4]$ 分别应用于骨干网络提取的每 1 级特征 $\boldsymbol{f}_i$ ($i\in[1,2,3,4]$)上, 在局部分支和全局分支中, 利用密集递进相连的深度可分离卷积块提取丰富的多尺度上下文特征 $\boldsymbol{f}_i^{\text{M}}\left(i\in[1,2,3,4]\right)$; 最后使用 $\text{FACIFM}_i, i\in[1,2,3]$ 融合各级特征, 让网络尽可能地集中注意力于伪装特征而避免背景噪声的干扰, 其输入为当前层级的多尺度上下文特征 $\boldsymbol{f}_i^{\text{M}}$、上一层级的 FACIFM_{i+1} 的输出 $\boldsymbol{f}_{i+1}^{\text{F}}$ ($i=3$ 时为 $\boldsymbol{f}_4^{\text{M}}$), 以及当前层级的骨干特征提取模块得到的 $\boldsymbol{f}_i$, 得到融合后的特征 $\boldsymbol{f}_i^{\text{F}}$. 为了让各级特征在空间上保持高度一致性, 将每级融合特征 $\boldsymbol{f}_i^{\text{F}}$ 和 $\boldsymbol{f}_4^{\text{M}}$ 送入通道数为 1 的卷积层预测伪装映射 $\boldsymbol{O}_i\left(i\in[1,2,3,4]\right)$, 并对多个输出进行深度监督, 取最顶层的输出 $\boldsymbol{O}_1$ 作为最终的输出结果. 下面分别对 TSDMM 和 FACIFM 进行详细介绍.

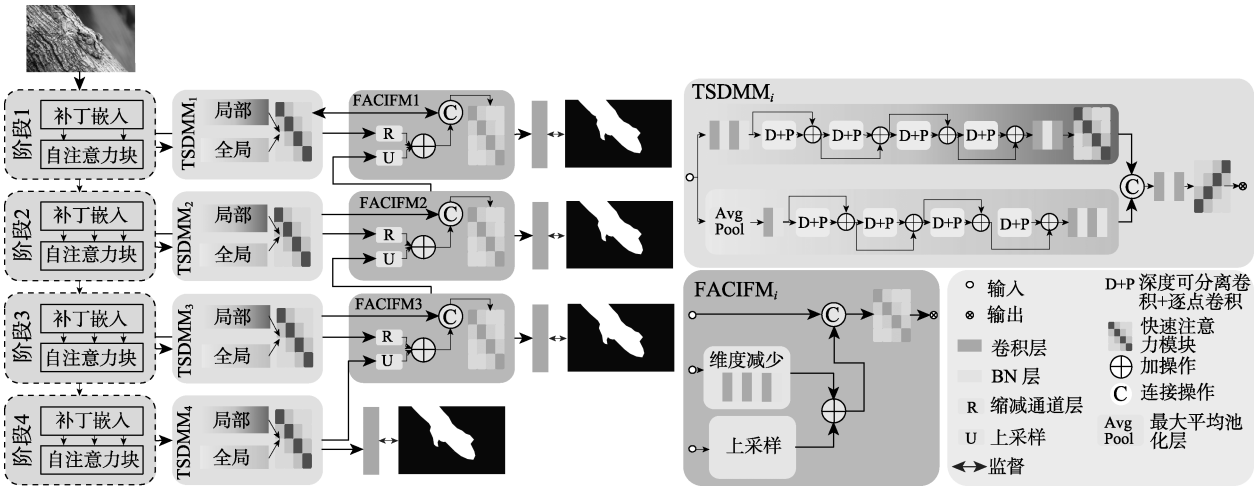


图 1　DMT-COD 方法整体结构

2.2　TSDMM

如图 1 右上部分所示, TSDMM 包括局部分支和全局分支, 利用密集递进相连的深度可分离卷积块分别以局部和全局视角提取特征, 并将其进行充分的融合. 以 TSDMM_i 为例, 依次介绍局部分支、全局分支和双分支融合部分的结构.

在局部分支中, 以骨干网络提取的特征 $\boldsymbol{f}_i$ 为输入, 首先通过 2 个核大小为(3, 3)的卷积层进行通道数的调整, 每个卷积层后都设置了一个 BN(batch normalization)层, 该过程为

$$\boldsymbol{f}'_{i\text{-local}}=\text{BN}\Big(\text{Conv}\Big(\text{BN}\big(\text{Conv}\big(\boldsymbol{f}_i\big)\big)\Big)\Big).$$

为了丰富特征的有效感受野, 希望通过不同核大小的卷积层捕获不同尺度的局部特征. 由于卷积核的增大会带来参数和计算量的剧增, 因此采用深度可分离卷积块控制模型的参数和计算开销, 每个深度可分离卷积块由对应核大小的深度卷积和 1×1 的逐点卷积组成; 为了避免信息的丢失, 将输入的特征以残差的形式加入到每个深度可分离卷积块的输出. 本文共设置 4 个深度可分离卷积块, 其核大小分别是 3, 5, 7 和 9. 该过程为

$$\boldsymbol{f}_{i,j}^{\text{local}}=\begin{cases}\text{pwConv}\Big(\text{dwConv}\big(\boldsymbol{f}'_{i\text{-local}}\big)\Big)+\boldsymbol{f}'_i, & \text{if}\quad j=1\\ \text{pwConv}\Big(\text{dwConv}\Big(\boldsymbol{f}_{i,j-1}^{\text{local}}\Big)\Big)+\boldsymbol{f}_{i,j-1}^{\text{local}}, & \text{otherwise}\end{cases}.$$

通过 2 个连续的 1×1 的卷积层进行融合

$$\boldsymbol{f}_i^{\text{local-multi}}=\text{Conv}\left(\text{BN}\left(\text{Conv}\left(\boldsymbol{f}_{i,4}^{\text{local}}\right)\right)\right).$$

对 4 个不同核大小的深度可分离卷积块进行处理后，$\boldsymbol{f}_i^{\text{local-multi}}$ 中聚合了不同大小感受野下捕获的特征. 为了对这些特征给予不同的关注度，受到文献[34]启发，本文设置了快速注意力模块，该过程包括 FA(fast attention)操作，LN(layer normalization)层和 MLP(multi-layer perceptron)层，具体过程为

$$\boldsymbol{f}_i^{\text{local-attn}'}=\text{FA}\left(\text{LN}\left(\boldsymbol{f}_i^{\text{local-multi}}\right)\right)+\boldsymbol{f}_i^{\text{local-multi}};$$

$$\boldsymbol{f}_i^{\text{local-attn}}=\text{MLP}\left(\text{LN}\left(\boldsymbol{f}_i^{\text{local-attn}'}\right)\right)+\boldsymbol{f}_i^{\text{local-attn}'}.$$

其中，FA 的计算方法有别于传统的注意力模块. 传统注意力模块在输入序列的长度较长时效果不佳，需要的计算时间及存储的空间都与输入长度呈平方级数增长. 如图 2 所示，本文使用的快速注意力模块中，将其等价为一个线性注意力框架，有效地降低了计算成本和计算复杂性. 其中，左边为标准的单向注意力机制，需要对注意矩阵进行掩码操作得到下三角矩阵；右边为标准注意力机制的无偏近似，将 $\boldsymbol{K}$ 向量和 $\boldsymbol{V}$ 向量的随机特征映射进行外积，得到的前缀和，且这个过程是动态构建的，最后将随机特征向量与 $\boldsymbol{Q}$ 向量进行乘积，得到最终矩阵中的新行.

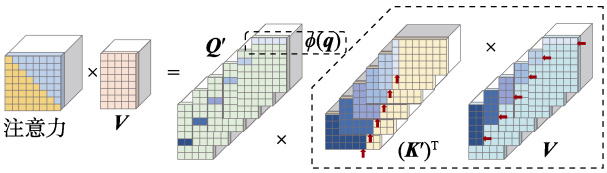


图 2 快速注意力模块示意图

在全局分支中，首先对输入 $\boldsymbol{f}_i$ 执行 AMP (adaptive max pooling)层，并通过 1×1 卷积及 BN 层进行调整，该过程为

$$\boldsymbol{f}'_{i\text{-global}}=\text{BN}\left(\text{Conv}\left(\text{AMP}\left(\boldsymbol{f}_i\right)\right)\right).$$

其中，AMP 层的输出大小设置为了 9，对应于深度可分离卷积块的最大核大小，保证了对特征全局感受野的覆盖. 与局部分支类似，本文设置了 4 组深度可分离卷积块，其核大小分别是 3，5，7 和 9. 该过程可表示为

$$\boldsymbol{f}_{i,j}^{\text{global}}=\begin{cases}\text{pwConv}\left(\text{dwConv}\left(\boldsymbol{f}'_{i\text{-global}}\right)\right)+\boldsymbol{f}'_i, & j=1\\ \text{pwConv}\left(\text{dwConv}\left(\boldsymbol{f}_{i,j-1}^{\text{global}}\right)\right)+\boldsymbol{f}_{i,j-1}^{\text{global}}, & j=2,3,4\end{cases}.$$

通过 2 个连续的 1×1 的卷积层进行融合，并通过双线性插值法将特征扩大到与输入 $\boldsymbol{f}_i$ 同样的分辨率. 该过程为

$$\boldsymbol{f}_i^{\text{global-multi}}=\text{Upsample}\left(\text{Conv}\left(\text{BN}\left(\text{Conv}\left(\boldsymbol{f}_{i,4}^{\text{global}}\right)\right)\right)\right).$$

在双分支融合部分中，将 2 个分支得到的结果在通道维度上进行连接，并通过 2 个 1×1 的卷积层进行融合. 同样地，为了对 2 个分支的特征给予不同的关注度，在其后设置了快速注意力模块，该过程为

$$\boldsymbol{f}_i^{\text{con}}=\text{Concat}\left(\boldsymbol{f}_i^{\text{global-multi}},\boldsymbol{f}_i^{\text{local-attn}}\right);$$

$$\boldsymbol{f}_i^{\text{fuse-g-l}}=\text{Conv}\left(\text{BN}\left(\text{Conv}\left(\boldsymbol{f}_i^{\text{con}}\right)\right)\right);$$

$$\boldsymbol{f}_i^{\text{M}'}=\text{FA}\left(\text{LN}\left(\boldsymbol{f}_i^{\text{fuse-g-l}}\right)\right)+\boldsymbol{f}_i^{\text{fuse-g-l}};$$

$$\boldsymbol{f}_i^{\text{M}}=\text{MLP}\left(\text{LN}\left(\boldsymbol{f}_i^{\text{M}'}\right)\right)+\boldsymbol{f}_i^{\text{M}'}.$$

为了证明 TSDMM 能够丰富各级特征的有效感受野，对每层级中通过 TSDMM 之前骨干网络提取的每级特征 $\boldsymbol{f}_i$ 的有效感受野，以及经过 TSDMM 处理得到的 $\boldsymbol{f}_i^{\text{M}}$ 的有效感受野进行可视化，结果如图 3 所示. 可以看出，与卷积神经网络相比，虽然自注意力变换网络提取的多级特征的有效感受野已经有所增大，但还有很大的扩展空间. 经过 TSDMM 处理后，特征的有效感受野可以明显增大，在一定程度上缓解了自注意力变换网络中有限的特征感受野限制了模型对伪装物体的识别能力的问题.

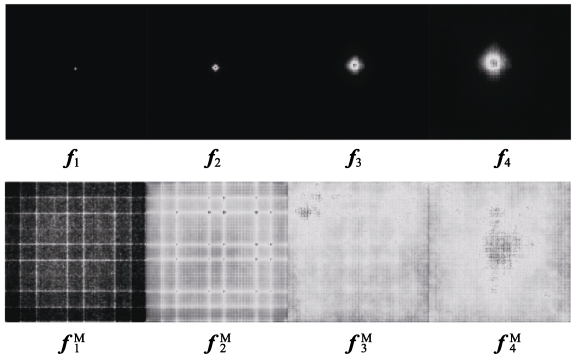


图 3 TSDMM 处理前后的特征有效感受野的可视化

2.3 FACIFM

FACIFM 的结构如图 1 右下部分所示. 该模块的输入包含 3 个部分：当前层级的多尺度上下文特征 $\boldsymbol{f}_i^{\text{M}}$，上一层级的 FACIFM_{i+1} 的输出 $\boldsymbol{f}_{i+1}^{\text{F}}$ ($i=3$ 时为 $\boldsymbol{f}_4^{\text{M}}$)，以及当前层级的骨干特征提取模块得到的 $\boldsymbol{f}_i$.

首先, 上一层级的 FACIFM$_{i+1}$ 的输出 $\boldsymbol{f}_{i+1}^{\mathrm{F}}$ ($i=3$时为 $\boldsymbol{f}_4^{\mathrm{M}}$)先通过双线性插值法, 上采样到与当前层级特征同样大小的分辨率; 然后将当前层级的骨干特征提取模块得到的 $\boldsymbol{f}_i$ 通过 3 个 1×1 的卷积层进行通道数的调整, 与上采样后的 $\boldsymbol{f}_{i+1}^{\mathrm{F}}{}'$ 进行逐像素相加, 并与当前层级的多尺度上下文特征 $\boldsymbol{f}_i^{\mathrm{M}}$ 在通道上进行连接, 该过程为

$$\boldsymbol{f}_{i+1}^{\mathrm{F}}{}' = \mathrm{Concat}\Big(\Big(\mathrm{UpSample}\big(\boldsymbol{f}_{i+1}^{\mathrm{F}}\big)\Big)+ \mathrm{Conv}\Big(\mathrm{BN}\Big(\mathrm{Conv}\Big(\mathrm{BN}\big(\mathrm{Conv}\big(\boldsymbol{f}_i\big)\big)\Big)\Big)\Big), \boldsymbol{f}_i^{\mathrm{M}}\Big).$$

类似地, 为了在整合各级特征时, 将注意力集中于伪装特征, 尽可能地避免背景噪声的干扰, 对来自 2 个层级的特征给予不同的关注度, 在其后设置了快速注意力模块, 该过程为

$$\boldsymbol{f}_i^{\mathrm{F}\prime\prime} = \mathrm{FA}\Big(\mathrm{LN}\big(\boldsymbol{f}_i^{\mathrm{F}'}\big)\Big)+ \boldsymbol{f}_i^{\mathrm{F}\prime};$$

$$\boldsymbol{f}_i^{\mathrm{F}} = \mathrm{MLP}\Big(\mathrm{LN}\big(\boldsymbol{f}_i^{\mathrm{F}\prime\prime}\big)\Big)+ \boldsymbol{f}_i^{\mathrm{F}\prime\prime}.$$

为了证明 FACIFM 能够让各级特征最大限度地将注意力集中于伪装对象区域, 消除超出伪装区域的无关部分的干扰, 获得有关伪装目标对象更纯净的表示, 对每级 FACIFM 的输入和输出特征进行可视化, 结果如图 4 所示. 可以看出, 经过每级 FACIFM 处理后, 特征分辨率不断扩大, 细节信息也不断被补充, 且输出的特征最大程度地将注意力集中于伪装对象区域, 未受到超出伪装区域的无关部分的干扰.

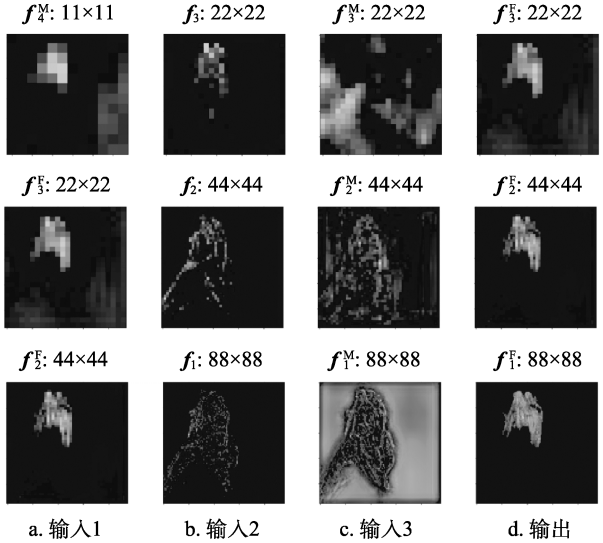


图 4　每级 FACIFM 的输入和输出特征的可视化

2.4　损失函数

二进制交叉熵(binary cross entropy, BCE)损失和交并比(intersection over union, IoU)损失广泛应用于 COD 任务[7-8]. 本文利用它们的总和监督模型的学习过程, 即 $L^{\mathrm{s}} = L_{\mathrm{BCE}}^{\mathrm{s}} + L_{\mathrm{IoU}}^{\mathrm{s}}$. 其中, L^{s} 表示模型其中一个输出的损失; $L_{\mathrm{BCE}}^{\mathrm{s}}$ 和 $L_{\mathrm{IoU}}^{\mathrm{s}}$ 分别表示 BCE 损失和 IoU 损失, 定义为

$$L_{\mathrm{BCE}}^{\mathrm{s}} = -\frac{1}{H\times W}\cdot \sum_{m=1}^{H}\sum_{n=1}^{W}\Big[G_{m,n}\log\big(P_{m,n}\big)+\big(1-G_{m,n}\big)\log\big(1-P_{m,n}\big)\Big].$$

$$L_{\mathrm{IoU}}^{\mathrm{s}} = 1-\frac{\sum_{m=1}^{H}\sum_{n=1}^{W}\big(P_{m,n}\times G_{m,n}\big)}{\sum_{m=1}^{H}\sum_{n=1}^{W}\big(P_{m,n}+G_{m,n}-P_{m,n}\times G_{m,n}\big)}.$$

其中, $G_{m,n}$ 和 $P_{m,n}$ 分别表示图像中 (m,n) 处的像素所对应的人工标注值和预测值.

如第 2.1 节所述, 本文将每级融合特征 $\boldsymbol{f}_i^{\mathrm{F}}$ 和最深一层的多尺度特征 $\boldsymbol{f}_4^{\mathrm{M}}$ 送入通道数为 1 的卷积层预测伪装映射 $\boldsymbol{O}_i\,(i\in[1,2,3,4])$, 并对其进行深度监督. 由于用于预测的特征的分辨率和重要性不同, 因此对输出赋予了不同的权重, 整个网络的损失为 $L^{\mathrm{all}} = L_{\boldsymbol{O}_1}^{\mathrm{s}} + \sum_{i=1}^{3}\frac{1}{2^i}L_{\boldsymbol{O}_{i+1}}^{\mathrm{s}}$.

3　实验及结果分析

3.1　数据集

为了评估DMT-COD方法, 在CHAMELEON[35], CAMO[29], COD10K[7]和 NC4K[9]这 4 个基准数据集进行实验.

CHAMELEON 数据集[35]由 76 幅高分辨率图像组成, 这些图像从互联网上通过 Google 搜索引擎, 使用“伪装动物”作为关键字收集, 并手动注释了对应的像素级标签. CAMO 数据集[29]包含 1 000 幅训练图像和 250 幅测试图像, 以及自然伪装对象和人工伪装对. COD10K 数据集[7]是目前最大的 COD 基准数据集, 由 5 066 幅伪装图像(其中, 3 040 幅图像用于训练, 2 026 幅图像用于测试)组成, 涵盖 5 个超级类和 69 个子类, 所有图像均从各种摄影网站下载. NC4K 数据集[9]提供了 4 121 幅测试图像, 以进行有效的模型评估.

本文遵循与现有工作[4,7]相同的训练集配置, 包括来自 CAMO 数据集[29]的 1 000 个样本和来自 COD10K 数据集[7]的 3 040 个样本; 然后, 在整个 CHAMELEON 数据集[35], NC4K 数据集[9], 以及 CAMO 数据集[29]和 COD10K 数据集[7]的测试集上

进行模型的性能实验.

3.2 评价指标

使用 5 个广泛采用的评估指标定量评估本文提出的模型的有效性: 准确率-召回率(precision-recall, PR)曲线, S 值(S-measure)[36], F 值(F-measure)[37], E 值(E-measure)[38]和平均绝对误差(mean absolute error, MAE)[39].

PR 曲线中, P 指准确率, R 指召回率, 它们是根据预测的伪装二值映射与真值之间的重叠面积而得到的, 计算公式分别为

$$\begin{cases} P = \dfrac{|M \cap G|}{M} \\ R = \dfrac{|M \cap G|}{G} \end{cases}.$$

PR 曲线反映了模型对正例的识别准确度和对正例的覆盖能力之间的权衡. 通常, 一个好的模型要求其准确率和召回率尽可能地高, 也就意味着 PR 曲线更加靠近右上角.

S 值[36]侧重于评估对象感知和区域感知的结构相似性, 定义为 $S_m = \alpha \times S_o(1-\alpha) \times S_r$; 其中, α 参照文献[36]设置为 0.5, 以平衡对象感知结构相似性 S_o 和区域感知结构相似性 S_r.

F 值[37]是一种综合评价标准, 同时考虑了 P 和 R, 公式为

$$F_\beta = \frac{\left(1+\beta^2\right) \times P \times R}{\beta^2 \times P + R}.$$

其中, 参照文献[37], 设置 $\beta^2 = 0.3$.

E 值[38]将局部像素值与图像级平均值相结合, 评估预测与真值之间的相似性, 同时评估了像素级匹配和图像级统计信息. 本文通过该指标评估 COD 结果的整体和局部准确性.

与 PR 曲线和 F 值等不同, MAE[39]更加平衡地比较预测的伪装映射和真值图的差别, 定义为

$$\mathrm{MAE} = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} \left| \boldsymbol{S}(x,y) - \boldsymbol{G}(x,y) \right|.$$

其中, W 和 H 表示输入图像的宽度与高度, $\boldsymbol{S}(x,y)$ 和 $\boldsymbol{G}(x,y)$ 分别表示预测的伪装对象映射中与真值图中点 (x,y) 的像素值. MAE 值越小, 表示该模型的性能越好.

为保证对比的公平性, 本文统一使用文献[7]中的工具(https://github.com/DengPingFan/CODToolbox)进行评估.

3.3 实验细节

本文模型是基于 Pytorch[40]库实现的. 使用配置 Intel 酷睿 i9-10900X 3.70 GHz CPU(32 GB RAM)和 NVIDIA GeForce RTX 3090 GPU(24 GB 内存)的 10 核 PC 机训练和测试模型. 在训练阶段, 将随机水平翻转、随机裁剪和多尺度图像输入作为数据增强技术, 避免过拟合问题. 本文分别使用 MPViT-B 和 MPViT-S[17]作为骨干特征提取网络, 其参数使用在 ImageNet 数据集[41]上预先训练的相应模型进行初始化, 其余的参数通过 PyTorch[40]进行初始化. 本文通过 AdamW 优化器[42]对整个网络进行端到端训练, 权重衰减为 0.01, 训练 60 000 轮迭代次数, 批大小设置为 9, 初始学习率为 10^{-4}, 并使用预热和线性衰减调整学习率. 在测试阶段, 将每幅图像的大小调整为 384×384 像素, 然后将其馈送到本文提出的网络预测结果, 而无需进行任何后处理, 如条件随机场[43]等. 最后, 将预测图的大小调整回输入图像的原始大小, 2 次调整大小的过程使用双线性插值法.

3.4 与主流方法的比较

将本文模型与 28 种主流的方法进行对比实验, 证明其有效性. 由于 COD 是一个新兴方向, 因此本文与包含显著性对象检测、医学图像分割和伪装对象分割的基线方法进行对比, 包括 PiCANet[44], SCRN[45], NLDF[46], CSNet[47], SSAL[48], UCNet[49], MINet[50], ITSD[51], PoolNet[52], BASNet[53], CPD[54], PFANet[55], EGNet[56], Anet-SRM[29], F^3Net[57], Pra-Net[5], SINet[7], PFNet[8], LSR[9], C2FNet[13], UJNet[14], MGL-R[30], MirrorNet[15], UGTR[16], ERRNet[17], OCENet[18], CubeNet[19]和 BgNet[20]. 为了保证比较的公平性, 本文使用作者公开的结果或者原文献作者发布的源码在各个数据集上运行得到的结果进行对比.

3.4.1 定量比较

表 1 所示为 29 种方法在 4 个数据集上的 S 值, E 值, F 值和 MAE 值的定量比较结果. 其中, 本文-B 是使用 MPViT-B 作为骨干特征提取网络的方法, 本文-S 是使用 MPViT-S 作为骨干特征提取网络的方法. 可以看出: (1) 在无须进行任何后处理的情况下, DMT-COD 方法在 4 个数据集上的评价指标下都优于对比方法. (2) 在 2 个大规模数据集 COD10K 和 NC4K 上, 与第 2 名 BgNet 方法[20]相比, DMT-COD 方法在所有指标上都实现了显著的性能提升; 在 COD10K 数据集上, F 值提升 7.19%, MAE 降低 32.35%; 在 NC4K 数据集上, F 值提升 4.08%, MAE 降低 28.89%. 12 种方法在 4 个数据集上的 PR 曲线对比如图 5 所示. 可以看出, DMT-COD 方法得出的 PR 曲线(红色曲线)始终位于其他

表 1　29 种方法在 4 个数据集上的对比

方法	CAMO-Test				CHAMELEON				COD10K-Test				NC4K			
	S 值↑	*E* 值↑	*F* 值↑	MAE↓	*S* 值↑	*E* 值↑	*F* 值↑	MAE↓	*S* 值↑	*E* 值↑	*F* 值↑	MAE↓	*S* 值↑	*E* 值↑	*F* 值↑	MAE↓
PiCANet[44]	0.609	0.753	0.529	0.156	0.769	0.836	0.698	0.085	0.649	0.678	0.485	0.090	0.758	0.808	0.685	0.888
SCRN[45]	0.779	0.848	0.738	0.090	0.876	0.894	0.836	0.042	0.789	0.789	0.699	0.047	0.830	0.864	0.793	0.059
NLDF[46]	0.665	0.796	0.648	0.123	0.798	0.872	0.769	0.063	0.701	0.786	0.599	0.059	0.738	0.834	0.713	0.083
CSNet[47]	0.771	0.847	0.740	0.092	0.856	0.888	0.810	0.047	0.778	0.791	0.679	0.047	0.750	0.812	0.669	0.088
SSAL[48]	0.644	0.765	0.601	0.126	0.757	0.849	0.709	0.071	0.668	0.782	0.535	0.066	0.699	0.805	0.654	0.093
UCNet[49]	0.739	0.811	0.708	0.094	0.880	0.929	0.851	0.036	0.776	0.867	0.691	0.042	0.811	0.883	0.782	0.055
MINet[50]	0.748	0.835	0.706	0.090	0.855	0.929	0.821	0.036	0.770	0.848	0.671	0.042	0.812	0.882	0.775	0.056
ITSD[51]	0.750	0.830	0.694	0.102	0.814	0.863	0.746	0.057	0.767	0.787	0.658	0.051	0.811	0.855	0.762	0.064
PoolNet[52]	0.702	0.790	0.629	0.129	0.776	0.824	0.706	0.081	0.705	0.708	0.570	0.074	0.785	0.842	0.736	0.073
BASNet[53]	0.618	0.719	0.519	0.159	0.687	0.742	0.581	0.118	0.634	0.676	0.451	0.105	0.695	0.784	0.617	0.095
CPD[54]	0.716	0.807	0.658	0.113	0.857	0.898	0.813	0.048	0.750	0.792	0.640	0.053	0.787	0.850	0.742	0.072
PFANet[55]	0.659	0.735	0.564	0.172	0.679	0.732	0.552	0.144	0.636	0.619	0.466	0.128	0.647	0.747	0.521	0.127
EGNet[56]	0.662	0.780	0.625	0.125	0.750	0.854	0.693	0.075	0.733	0.799	0.620	0.055	0.777	0.826	0.731	0.075
F^3Net[57]	0.711	0.802	0.630	0.109	0.848	0.895	0.798	0.047	0.739	0.818	0.609	0.051	0.780	0.853	0.719	0.070
PraNet[5]	0.769	0.833	0.782	0.094	0.860	0.898	0.830	0.044	0.789	0.839	0.704	0.045	0.822	0.871	0.786	0.059
Anet-SRM[29]	0.682	0.725	0.566	0.126												
SINet[7]	0.751	0.834	0.706	0.100	0.869	0.899	0.832	0.044	0.771	0.797	0.676	0.051	0.808	0.882	0.775	0.058
PFNet[8]	0.782	0.852	0.758	0.085	0.882	0.942	0.853	0.033	0.814	0.868	0.738	0.035	0.829	0.892	0.799	0.053
LSR[9]	0.787	0.855	0.753	0.080	0.890	0.936	0.862	0.030	0.804	0.882	0.732	0.037	0.840	0.902	0.815	0.048
UJNet[14]	0.800	0.865	0.779	0.073	0.891	0.943	0.862	0.030	0.800	0.882	0.725	0.040	0.842	0.903	0.816	0.047
MGL-R[30]	0.775	0.847	0.740	0.088	0.893	0.923	0.860	0.030	0.814	0.865	0.738	0.035	0.833	0.889	0.800	0.052
MirrorNet[15]	0.785	0.849	0.756	0.077												
UGTR[16]	0.784	0.856	0.752	0.086	0.887	0.918	0.846	0.031	0.817	0.850	0.741	0.036	0.839	0.886	0.807	0.052
ERRNet[17]	0.779	0.851	0.742	0.085	0.868	0.917	0.840	0.039	0.786	0.845	0.702	0.043	0.827	0.890	0.794	0.054
C2FNet[13]	0.796	0.864	0.771	0.080	0.888	0.932	0.863	0.032	0.813	0.886	0.743	0.036	0.838	0.898	0.810	0.049
OCENet[18]	0.802	0.863	0.777	0.080	**0.897**	0.928	0.870	0.027	0.827	0.883	0.764	0.032	0.853	0.904	0.831	0.045
CubeNet[19]	0.788	0.849	0.750	0.085	0.873	0.931	0.849	0.037	0.795	0.862	0.715	0.041				
BgNet[20]	0.831	0.882	0.813	0.065	0.894	0.933	0.870	0.029	0.826	0.883	0.765	0.034	0.855	0.904	0.833	0.045
DMT-COD-S	0.841	0.894	0.834	0.056	0.895	**0.951**	**0.876**	0.027	0.847	0.910	0.795	0.027	0.875	0.919	0.856	0.034
DMT-COD-B	**0.863**	**0.915**	**0.860**	**0.047**	0.895	0.943	0.871	**0.026**	**0.864**	**0.922**	**0.820**	**0.023**	**0.883**	**0.923**	**0.867**	**0.032**

注. 加粗表示最优值.

曲线之上, 与表 1 中的评价指标相一致. 实验结果表明, 在 4 个数据集上, DMT-COD 方法的所有评价指标都取得了最好的结果.

为了对比DMT-COD方法的计算效率, 将该方法的浮点数计算量(floating point operations, FLOPs)和参数量与其他方法进行实验, 结果如表2所示. 为了保证比较的公平性, 使用同一幅输入图像进行测算. 可以看出, 在准确性达到最佳的情况下, DMT-COD 方法在 FLOPs 和模型参数量上仍保持较强的竞争力(尤其是 DMT-COD-S); 与 MGL-R 方法[30]相比, DMT-COD-S 方法在 *F* 值平均提升 5.58%的情况下, FLOPs 减小 389.58G, 模型参数量减小了 48.15M, DMT-COD-B 方法在 *F* 值平均提升 7.00%的情况下, FLOPs 减小 301.11G; 与 UGTR 方法[16]相比, DMT-COD-S 方法在 *F* 值平均提升 5.38%的情况下, FLOPs 减小 79.61G, 模型参数量减小 7.22M; 与 EGNet 方法[56]相比, DMT-COD-S 方法在 *F* 值平均提升 17.3%的情况下, FLOPs 减小

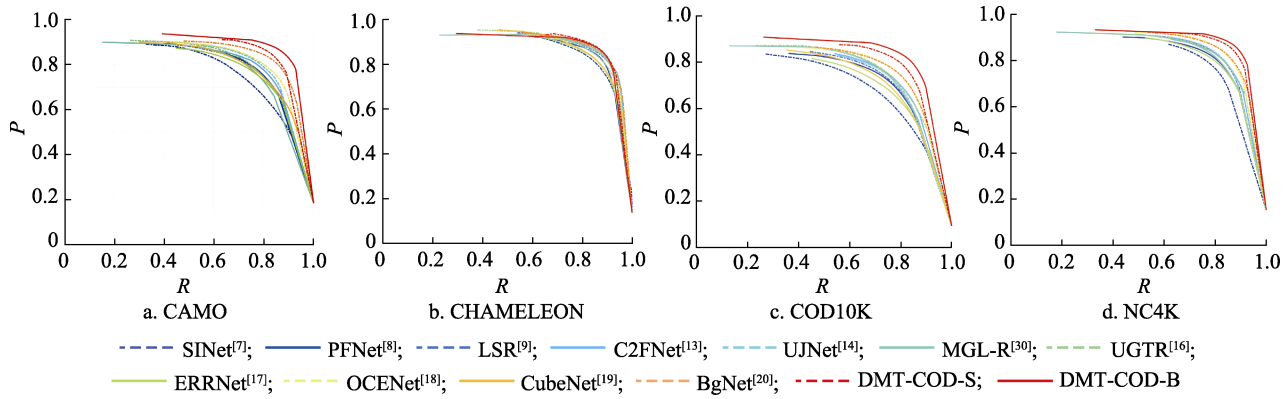


图 5 12 种方法在 4 个数据集上的 PR 曲线对比

79.08G, 模型参数量减小 81.34M, DMT-COD-B 在 F 值平均提升 18.73%的情况下, 模型参数量减小 17.58M; DMT-COD-B 方法虽然在准确性方面取得最佳结果, 但在 FLOPs 和参数量上只能优于少数方法; 而 DMT-COD-S 方法 FLOPs 和参数量上有着极大的优势, 可以超越表 2 中的所有方法, 且 S 与其他 28 种方法相比, 在 4 个数据集上可以取得除 DMT-COD-B 以外最好的结果. 当应用于不同的应用场景时, 可以根据任务精度要求和多媒体设备的负载能力选择对应版本的 DMT-COD 方法, 进一步证明了该方法具有较好的性能和灵活性.

表 2 不同方法 FLOPs 及模型参数量的对比

方法	FLOPs↓	参数量↓
PICANet[44]	59.80G	47.22M
PoolNet[52]	44.43G	69.56M
EGNet[56]	120.80G	111.66M
LSR[9]	68.50G	55.45M
UJNet[14]	56.21G	55.88M
MGL-R[30]	431.30G	78.47M
UGTR[16]	121.33G	37.54M
DMT-COD-S	**41.72G**	**30.32M**
DMT-COD-B	130.19G	94.08M

注. 加粗表示最优值.

3.4.2 定性比较

图 6 所示为本文模型与其他伪装对象分割方法的结果对比. 从图 6 中可以看出, DMT-COD 方法可以处理各种具有挑战性的场景, 包括较大的伪装目标(如第 1 行)、微小的伪装目标(如第 2 行)、多个伪装目标同时出现(如第 3 行)、前景干扰 (如第 4 行)、前景极难被发现(如第 5 行和第 6 行), 以及背景极易被误判(如第 7 行和第 8 行)等情况; 此外, 在其他大多数主流方法检测到正确的伪装对象的情况下, DMT-COD 方法可以获得更加完整且边缘细节更加精细的结果(如第 9 行和第 10 行). 原因如下: (1) 自注意力变换网络捕获了全局上下文信息从而建立长距离依赖关系, 对于伪装对象的识别起到至关重要的作用; (2) TSDMM 使特征具有了丰富的有效感受野, 让网络得以识别类别、大小、形状高度可变的伪装对象; (3) FACIFM 使得网络在整合各级特征时集中注意力于伪装特征, 尽可能避免背景噪声的干扰.

3.5 消融实验

对本文模型中自注意力变换网络, TSDMM 和 FACIFM 的有效性进行消融实验的, 结果如表 3 所示, 这些实验是独立进行的.

(1) 自注意力变换网络的有效性. 为了证明该网络对任务性能的提升, 将网络中的骨干特征提取网络 MPViT 替换为卷积神经网络 ResNet-50. 如表 3 所示, 第 1 行和第 9 行的对比中, 若将骨干特征提取网络替换为卷积神经网络, 性能将显著下降, 其中, S 值从 0.864 减小到 0.773, E 值从 0.922 减小到 0.856, F 值从 0.820 减小到 0.675, MAE 从 0.023 增加到 0.044, 表明自注意力变换网络在 COD 中应用的必要性.

(2) TSDMM 的有效性. 通过 5 种变体, 验证 TSDMM 的有效性, 包括 w/o TSDMM, w/o Local, w/o Global, w/o Local-attention, 和 w/o Fuse-attention. w/o TSDMM 指在没有 TSDMM 的情况下, 通过重新训练 DMT-COD 获得的结果. 可以看出, 由 COD10K-Test 数据集上 F 值从 0.798 增加到 0.820, MAE 从 0.031 减小到 0.023, 充分证明了 TSDMM 的整体有效性. w/o Local 和 w/o Global 分别表示去除局部或全局特征提取分支的情况. 可以看出, 无论删除哪个分支指标都会大幅下降. 本文还分别设计了去除局部分支中的注意力模块或者去除融合部分的注意力模块的实验, 结果表示,

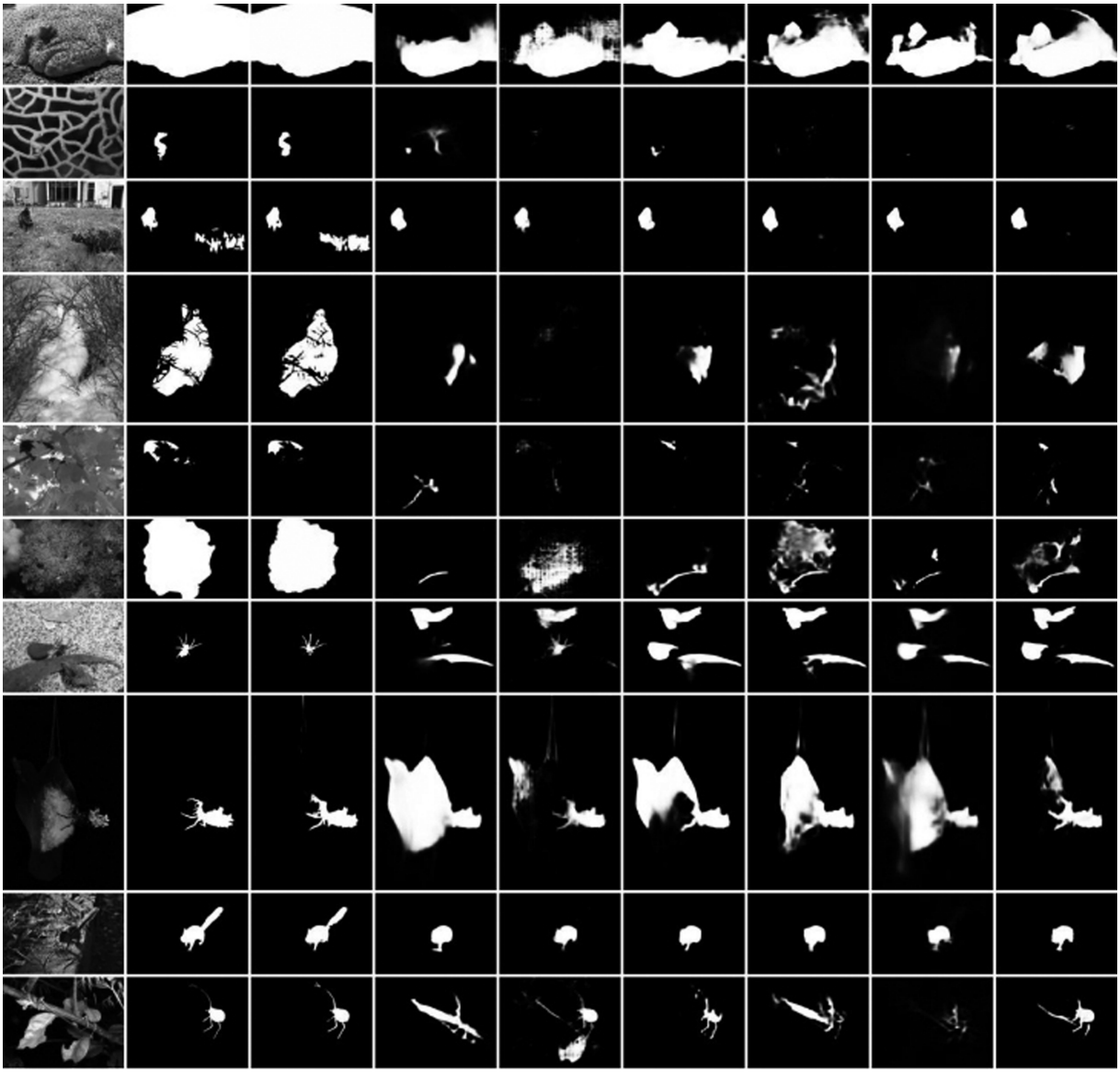


图 6　7 种方法的实验结果对比

注意力模块的去除将会导致性能指标的大幅下降. 为了证明 TSDMM 中局部分支的多尺度特征提取能力, 将局部分支修改为经典的多尺度特征提取模块 ASPP[58]结构, 实验结果如表 3 中第 7 行和第 9 行的对比所示, 可以看出, 本文设计的局部分支拥有比 ASPP 更强的多尺度特征提取能力. 以上消融实验都证实了 TSDMM 动机的合理性和结构的有效性.

(3) FACIFM 的有效性. 为了验证 FACIFM 的有效性, 用简单的串联层融合不同层级的特征. 如表 3 所示中第 8 行和第 9 行的对比所示, 若将 FACIFM 替换, 性能将显著下降, 其中, S 值为从 0.864 减小到 0.854, E 值为从 0.922 减小到 0.918, F 值从 0.820 减小到 0.807, MAE 为从 0.023 增加到 0.025, 表明 FACIFM 可以提高模型的性能.

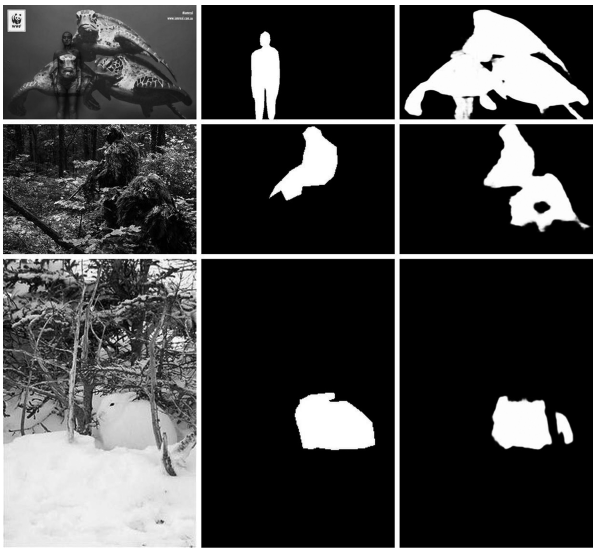
3.6　失败案例分析

图 7 所示为 DMT-COD 方法用于 COD 的 3 个典型失败案例. 例如, 在非常复杂的情况下模型有时会失败, 此时伪装对象与周围环境中的其他物体几乎融为一体, 导致模型将与其融为一体的整个区域分割出来, 如第 1 行所示; 第 2 行所示为伪装对象标注完整性的歧义, 人工标注中伪装对象被遮挡的部分被补全(遮挡了兔子的树枝被标记为伪装对象区域), 而本文模型则将树枝部分分割为背景区域; 第 3 行所示为由于标注者对伪装对象的误判而导致的失败情况, 其中包含多个独立的伪装对象, 但标注者只识别并标注出了其中的一个, 导致 DMT-COD 方法的分割结果与人工标注产生较大差距. 第 1 种情况属于当前 COD 方法要克服

表 3 在 COD10K-Test 上的消融实验结果

实验	方法	*S* 值↑	*E* 值↑	*F* 值↑	MAE↓
关于自注意力变换网络	w/o Transformer, w/ ResNet	0.773	0.856	0.675	0.044
关于 TSDMM	w/o TSDMM	0.832	0.894	0.798	0.031
	w/o Local	0.840	0.899	0.801	0.029
	w/o global	0.839	0.900	0.803	0.029
	w/o Local-attention	0.842	0.901	0.802	0.027
	w/o Fuse-attention	0.849	0.911	0.805	0.026
	局部分支→ASPP 结构	0.863	0.920	0.819	0.025
关于 FACIFM	w/o FACIFM, w/ Concat	0.854	0.918	0.807	0.025
本文网络	DMT-COD-B	**0.864**	**0.922**	**0.820**	**0.023**

注. 加粗表示最优值.



a. 原图 b. 人工标注 c. DMT-COD

图 7 典型失败案例展示

的主要难题, 后 2 种情况可以归因于数据集标注过程中对伪装对象定义的模糊性. 当其他方法遇到以上 3 种情况时, 会产生与 DMT-COD 方法相同甚至更差的结果.

由于数据集重标注的代价较大, 且标注歧义的情况在数据集中所占比例不大, 因此未来重点考虑对本文第 1 种情况进行改进, 可以考虑结合伪装对象的语义类别知识等对伪装对象整体的误判产生纠正作用. 此外, 本文仅对基于图像的 COD 进行研究, 未来可以考虑加入时序特征的提取, 对基于视频的 COD 方法进行探索.

4 结 语

本文提出一种有效的基于密集多尺度自注意力变换网络的 COD 方法, 将深度自注意力变换网络引入 COD, 利用全局线索提高 COD 的性能. 大量实验结果表明, 本文方法在多个 COD 数据集上优于其他 28 种主流方法. 尽管取得了显著的成果, 但本文方法也存在一些局限性. 例如, 在处理非常复杂的背景时仍有提升空间. 因此未来将考虑结合伪装对象的语义类别知识等对伪装对象整体的误判产生纠正作用. 此外, 本文仅对基于图像的 COD 进行研究, 未来将考虑对基于视频的 COD 方法进行探索.

参考文献(References):

[1] Owens A, Barnes C, Flint A, *et al*. Camouflaging an object from many viewpoints[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2014: 2782-2789

[2] Stevens M, Merilaita S. Animal camouflage: current issues and new perspectives[J]. Philosophical Transactions of the Royal Society B: Biological Sciences, 2009, 364(1516): 423-427

[3] Le X Y, Mei J H, Zhang H D, *et al*. A learning-based approach for surface defect detection using small image datasets[J]. Neurocomputing, 2020, 408: 112-120

[4] Pérez-de La Fuente R, Delclòs X, Peñalver E, *et al*. Early evolution and ecology of camouflage in insects[J]. Proceedings of the National Academy of Sciences of the United States of America, 2012, 109(52): 21414-21419

[5] Fan D P, Ji G P, Zhou T, *et al*. PraNet: parallel reverse attention network for polyp segmentation[C] //Proceedings of the 23rd International Conference on Medical Image Computing and Computer-Assisted Intervention. Heidelberg: Springer, 2020: 263-273

[6] Fan D P, Zhou T, Ji G P, *et al*. Inf-Net: automatic COVID-19 lung infection segmentation from CT images[J]. IEEE Transactions on Medical Imaging, 2020, 39(8): 2626-2637

[7] Fan D P, Ji G P, Sun G L, *et al*. Camouflaged object detection[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 2774-2784

[8] Mei H Y, Ji G P, Wei Z Q, *et al*. Camouflaged object segmentation with distraction mining[C] //Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 8768-8777

[9] Lv Y Q, Zhang J, Dai Y C, *et al*. Simultaneously localize, segment and rank the camouflaged objects[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 11586-11596

[10] Liu J W, Zhang J, Barnes N. Confidence-aware learning for camouflaged object detection[OL]. [2022-09-06]. https://arxiv.org/abs/2106.11641

[11] Zhu H W, Li P, Xie H R, *et al*. I can find you! Boundary-guided separated attention network for camouflaged object detection[C] //Proceedings of the 36th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2022: 3608-3616

[12] Liang Xinyu, Quan Jichuan, Yang Hui, *et al*. Camouflage target segmentation algorithm using multi-scale feature extraction and multi-level attention mechanism[J]. Journal of Computer-Aided Design & Computer Graphics, 2022, 34(5): 683-692(in Chinese)
(梁新宇, 权冀川, 杨辉, 等. 多尺度特征提取和多层次注意力机制的迷彩伪装目标分割算法[J]. 计算机辅助设计与图形学学报, 2022, 34(5): 683-692)

[13] Chen G, Liu S J, Sun Y J, *et al*. Camouflaged object detection via context-aware cross-level fusion[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(10): 6981-6993

[14] Li A X, Zhang J, Lv Y Q, *et al*. Uncertainty-aware joint salient object and camouflaged object detection[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 10066-10076

[15] Yan J N, Le T N, Nguyen K D, *et al*. MirrorNet: bio-inspired camouflaged object segmentation[J]. IEEE Access, 2021, 9: 43290-43300

[16] Yang F, Zhai Q, Li X, *et al*. Uncertainty-guided transformer reasoning for camouflaged object detection[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 4126-4135

[17] Ji G P, Zhu L, Zhuge M C, *et al*. Fast camouflaged object detection via edge-based reversible re-calibration network[J]. Pattern Recognition, 2022, 123: Article No.108414

[18] Liu J W, Zhang J, Barnes N. Modeling aleatoric uncertainty for camouflaged object detection[C] //Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Los Alamitos: IEEE Computer Society Press, 2022: 2613-2622

[19] Zhuge M C, Lu X K, Guo Y Y, *et al*. CubeNet: X-shape connection for camouflaged object detection[J]. Pattern Recognition, 2022, 127: Article No.108644

[20] Chen T Y, Xiao J, Hu X G, *et al*. Boundary-guided network for camouflaged object detection[J]. Knowledge-Based Systems, 2022, 248: Article No.108901

[21] Zhu J C, Zhang X Y, Zhang S, *et al*. Inferring camouflaged objects by texture-aware interactive guidance network[C] //Proceedings of the 35th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2021: 3599-3607

[22] Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need[C] //Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2017: 6000-6010

[23] Wang W H, Xie E Z, Li X, *et al*. Pyramid vision transformer: a versatile backbone for dense prediction without convolutions[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 548-558

[24] Liu Z, Lin Y T, Cao Y, *et al*. Swin transformer: hierarchical vision transformer using shifted windows[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 9992-10002

[25] Lee Y, Kim J, Willette J, *et al*. MPViT: multi-path vision transformer for dense prediction[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 7277-7286

[26] Sengottuvelan P, Wahi A, Shanmugam A. Performance of decamouflaging through exploratory image analysis[C] //Proceedings of the 1st International Conference on Emerging Trends in Engineering and Technology. Los Alamitos: IEEE Computer Society Press, 2008: 6-10

[27] Pan Y X, Chen Y W, Fu Q, *et al*. Study on the camouflaged target detection method based on 3D convexity[J]. Modern Applied Science, 2011, 5(4): Article No.152

[28] Yin J Q, Han Y B, Hou W D, *et al*. Detection of the mobile object with camouflage color under dynamic background based on optical flow[J]. Procedia Engineering, 2011, 15: 2201-2205

[29] Le T N, Nguyen T V, Nie Z L, *et al*. Anabranch network for camouflaged object segmentation[J]. Computer Vision and Image Understanding, 2019, 184: 45-56

[30] Zhai Q, Li X, Yang F, *et al*. Mutual graph learning for camouflaged object detection[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 12992-13002

[31] Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16x16 words: Transformers for image recognition at scale[OL]. [2022-09-06]. https://arxiv.org/abs/2010.11929

[32] Yuan L, Chen Y P, Wang T, *et al*. Tokens-to-token ViT: training vision transformers from scratch on ImageNet[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 538-547

[33] Zheng S X, Lu J C, Zhao H S, *et al*. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 6877-6886

[34] Choromanski K, Likhosherstov V, Dohan D, *et al*. Rethinking attention with performers[OL]. [2022-09-06]. https://arxiv.org/abs/2009.14794

[35] Skurowski P, Abdulameer H, Błaszczyk J, *et al*. Animal camouflage analysis: Chameleon database[J]. Unpublished manuscript, 2018, 2(6): 7

[36] Fan D P, Cheng M M, Liu Y, *et al*. Structure-measure: a new way to evaluate foreground maps[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 4558-4567

[37] Achanta R, Hemami S, Estrada F, *et al*. Frequency-tuned sali-

ent region detection[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2009: 1597-1604

[38] Fan D P, Gong C, Cao Y, *et al*. Enhanced-alignment measure for binary foreground map evaluation[C] //Proceedings of the 27th International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann, 2018: 698-704

[39] Perazzi F, Krähenbühl P, Pritch Y, *et al*. Saliency filters: contrast based filtering for salient region detection[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2012: 733-740

[40] Paszke A, Gross S, Chintala S, *et al*. Automatic differentiation in PyTorch[OL]. [2022-09-06]. https://openreview.net/pdf?id=BJJsrmfCZ

[41] Deng J, Dong W, Socher R, *et al*. ImageNet: a large-scale hierarchical image database[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2009: 248-255

[42] Loshchilov I, Hutter F. Decoupled weight decay regularization[OL]. [2022-09-06]. https://arxiv.org/abs/1711.05101

[43] Krähenbühl P, Koltun V. Efficient inference in fully connected CRFs with Gaussian edge potentials[C] //Proceedings of the 24th International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2011: 109-117

[44] Liu N, Han J W, Yang M H. PiCANet: learning pixel-wise contextual attention for saliency detection[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 3089-3098

[45] Wu Z, Su L, Huang Q M. Stacked cross refinement network for edge-aware salient object detection[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 7263-7272

[46] Luo Z M, Mishra A, Achkar A, *et al*. Non-local deep features for salient object detection[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 6593-6601

[47] Gao S H, Tan Y Q, Cheng M M, *et al*. Highly efficient salient object detection with 100k parameters[C] //Proceedings of the 16th European Conference on Computer Vision. Heidelberg: Springer, 2020: 702-721

[48] Zhang J, Yu X, Li A X, *et al*. Weakly-supervised salient object detection via scribble annotations[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 12543-12552

[49] Zhang J, Fan D P, Dai Y C, *et al*. UC-Net: uncertainty inspired RGB-D saliency detection via conditional variational autoencoders[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 8579-8588

[50] Pang Y W, Zhao X Q, Zhang L H, *et al*. Multi-scale interactive network for salient object detection[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 9410-9419

[51] Zhou H J, Xie X H, Lai J H, *et al*. Interactive two-stream decoder for accurate and fast saliency detection[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 9138-9147

[52] Liu J J, Hou Q B, 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. Los Alamitos: IEEE Computer Society Press, 2019: 3912-3921

[53] Qin X B, Zhang Z C, Huang C Y, *et al*. BASNet: boundary-aware salient object detection[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 7471-7481

[54] Wu Z, Su L, Huang Q M. Cascaded partial decoder for fast and accurate salient object detection[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 3902-3911

[55] Zhao T, Wu X Q. Pyramid feature attention network for saliency detection[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 3080-3089

[56] 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. Los Alamitos: IEEE Computer Society Press, 2019: 8778-8787

[57] Wei J, Wang S H, Huang Q M. F^3Net: fusion, feedback and focus for salient object detection[C] //Proceedings of the 34th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020: 12321-12328

[58] Chen L C, Papandreou G, Kokkinos I, *et al*. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848