计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(2):115−124 [doi: 10.15888/j.cnki.csa.009395]

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

基于 Transformer 的跨尺度交互学习伪装目标检测①

李建东, 王　岩, 曲海成

(辽宁工程技术大学 软件学院, 葫芦岛 125105)
通信作者: 王　岩, E-mail: 1434808301@qq.com

摘　要: 伪装目标检测 (COD) 旨在精确且高效地检测出与背景高度相似的伪装物体, 其方法可为物种保护、医学病患检测和军事监测等领域提供助力, 具有较高的实用价值. 近年来, 采用深度学习方法进行伪装目标检测成为一个比较新兴的研究方向. 但现有大多数 COD 算法都是以卷积神经网络 (CNN) 作为特征提取网络, 并且在结合多层次特征时, 忽略了特征表示和融合方法对检测性能的影响. 针对基于卷积神经网络的伪装目标检测模型对被检测目标的全局特征提取能力较弱问题, 提出一种基于 Transformer 的跨尺度交互学习伪装目标检测方法. 该模型首先提出了双分支特征融合模块, 将经过迭代注意力的特征进行融合, 更好地融合高低层特征; 其次引入了多尺度全局上下文信息模块, 充分联系上下文信息增强特征; 最后提出了多通道池化模块, 能够聚焦被检测物体的局部信息, 提高伪装目标检测准确率. 在 CHAMELEON、CAMO 以及 COD10K 数据集上的实验结果表明, 与当前主流的伪装物体检测算法相比较, 该方法生成的预测图更加清晰, 伪装目标检测模型能取得更高精度.

关键词: 深度学习; 伪装目标检测; 视觉特征金字塔; 卷积神经网络; 特征融合

引用格式: 李建东,王岩,曲海成.基于 Transformer 的跨尺度交互学习伪装目标检测.计算机系统应用,2024,33(2):115–124. http://www.c-s-a.org.cn/1003-3254/9395.html

Transformer-based Cross Scale Interactive Learning for Camouflage Object Detection

LI Jian-Dong, WANG Yan, QU Hai-Cheng

(School of Software, Liaoning Technical University, Huludao 125105, China)

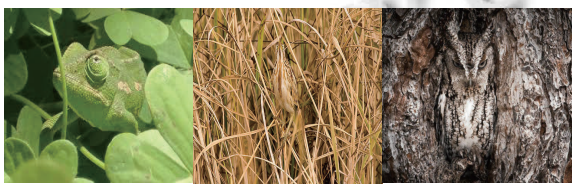
Abstract: Camouflage object detection (COD) aims to accurately and efficiently detect camouflaged objects that are highly similar to the background. Its method can assist in species protection, medical patient detection, and military monitoring, possessing high practical value. In recent years, using deep learning methods to detect camouflaged objects has become an emerging research direction. However, most existing COD algorithms apply a convolutional neural network (CNN) as the feature extraction network and ignore the influence of feature representation and fusion methods on detection performance when combining multi-level features. As the camouflage object detection model based on CNN has a weak ability to extract the global features of the detected object, this study proposes a cross scale interactive learning method for camouflage object detection based on Transformer. The model first puts forward a dual branch feature fusion module, which fuses features that have undergone iterative attention to better fuse high- and low-level features. Secondly, a multi-scale global context information module is introduced to fully integrate context information to enhance features. Finally, a multi-channel pooling module is proposed, which can focus on the local information of the detected object and improve the accuracy of camouflage target detection. The experimental results on the CHAMELEON, CAMO, and COD10K datasets show that this method generates clearer prediction maps and can achieve higher accuracy in

① 基金项目: 国家自然科学基金面上项目 (42271409); 辽宁省高等学校基本科研项目 (LIKMZ20220699)
收稿时间: 2023-08-06; 修改时间: 2023-09-09; 采用时间: 2023-09-26; csa 在线出版时间: 2023-12-18
CNKI 网络首发时间: 2023-12-19

camouflage object detection models than current mainstream camouflage object detection algorithms.

Key words: deep learning; camouflage object detection (COD); visual characteristic pyramid; convolutional neural network (CNN); feature fusion

伪装是自然界中一种普遍存在的防御机制, 使得某些物种可以隐藏在周围环境中保护自己以免受捕食者的伤害. 近年来, 从目标与背景高度相似的样本中检测目标, 即伪装目标检测 (COD) 引起了计算机视觉领域的广泛关注. 如图 1 所示, 由于伪装后的目标与背景之间在颜色以及纹理上高度相似, 难以检测到目标边缘信息. 以至于人类依赖于视觉难以发现伪装后的目标, 但在将彩色图片转换为张量后, 计算机可以检测到目标边界, 因此 COD 有良好的发展前景.



(a) 隐藏在叶子中的青蛙 (b) 与背景相似的鸟 (c) 隐藏在树干中的猫头鹰

图 1 伪装物体示例

早期研究者提出了多种基于传统特征提取的伪装目标检测方法[1−6]. 传统特征包括颜色与亮度、运动产生的特征、纹理特征等. 这些方法更多依赖于被检测物体的一些先验信息, 对伪装目标检测效果较差.

近年来, 多种基于深度学习的伪装目标检测方法以强大的特征提取和建模能力对伪装目标进行检测, 显示出巨大的潜力. 如 Sun 等人[7]提出的 C2FNet 通过聚合中高层特征联系上下文信息. 但因考虑到加入低层特征会造成计算量增加的问题, 而忽略了低层特征中所包含的边界信息. Mei 等人[8]提出的 PFNet 将更高层预测图和取反后的预测图分别与当前层特征相乘并输入到上下文探索块中以发现假阳性和假阴性预测, 接着分别使用逐元素加法和减法来抑制这两种干扰, 由于使用简单的加减法, 使得模型有着较快的推理速度, 但是直接将当前特征与预测先验直接相乘可能会导致特征混淆的问题. Fan 等人提出的 SINetV2[9]使用组反向注意力很好地解决了上述问题, 但该算法很难获得较大感受野, 从而导致无法得到充足的上下文信息, 并且不能将语义特征充分融合而造成特征损失. 以上方法存在的问题使得现有的伪装目标检测模型难以有效、完整地检测伪装目标的轮廓, 从而导致伪装目标检测效果不理想.

针对上述问题, 本文提出了一种新的伪装物体检测模型, 主要通过增强语义特征融合能力和增强全局上下文特征的联系来提高伪装目标检测的性能. 本文所作贡献如下.

(1) 设计一个基于迭代注意力[10]的跨层双分支特征融合模块 (dual branch feature fusion module, DBFF), 该模块跨层融合 PVT-V2 提取到的特征信息, 特征信息进入该模块与经过另一个双分支特征融合模块的特征信息交替集成进行初始特征的融合来优化伪装物体检测模型的检测过程.

(2) 设计全局上下文信息模块 (global communication context module, GCCM), 使用非对称卷积以及空洞卷积的混合卷积方式, 既增强了融合特征的表示, 还有效地融合了多层语义信息, 增强上下文信息的联系.

(3) 设计一个并行多通道池化模块 (parallel multi-channel pooling module, MCPM) 增强模型检测伪装物体的能力, 通过并行的池化操作, 聚合多尺度特征, 解决视觉 Transformer 主干网络因重点关注全局信息而导致局部信息的缺失的问题.

本文所提出的伪装目标检测模型实现了更强的目标定位能力和更精确的目标轮廓检测.

1 相关工作

1.1 视觉 Transformer 网络

在 COD 领域, 常用的主干网络是残差系列, 但视觉 Transformer 网络相比于卷积神经网络可以建模长程依赖, 具有更好的全局特征. ViT 是 Dosovitskiy 等人[11]提出的 Visual Transformer 模型, 但是 ViT 只能提取到单一尺度特征, 在语义分割任务上, 多尺度的特征非常重要. 因此 Wang 等人[12]提出了一种能够提取到多尺度特征的金字塔视觉 Transformer 主干网络——PVT. 模型总体上由 4 个 stage 组成, 用于生成不同尺度的特征. 与传统的 CNN 骨干网局部接受域随着网络深度的增加而增加, PVT 总是产生一个全局的接受域,

更适合检测和分割. PVT 继承了 CNN 和 Transformer 的优点, 具有更强大的特征提取能力, 可以作为 CNN 骨干的直接替代, 本文采用 PVT-V2[13]作为特征提取主干网络, 实现伪装目标检测精度的提升.

1.2 感受野增强

感受野模块具有增强特征可判别性和鲁棒性的作用, 可以进一步扩大视觉 Transformer 主干网络抽取特征的感受野范围, 使得模型取得更好的性能. RFB 模块[14]是模拟人类视觉的感受野从而加强网络的特征提取能力, 主要是在 Inception 的基础上加入了空洞卷积, 从而有效增大了感受野. 在该伪装目标检测模型中, 中高层特征在进入网络之前采用了该模块, 有利于提高图片全局特征提取的建模能力. 另一个增强感受野模块采用的是 PPM[15]. PPM 中的全局平均池化能够有效地融合全局上下文信息. 在该伪装目标检测模型中, 低层特征经过了该模块进行感受野增强, 有利于在网络最后将具有高分辨率的低层特征中的细节信息与中高层特征中的语义信息进行更好融合.

1.3 多尺度特征融合方法

经过感受野增强的特征采用有效聚合策略也是提升网络性能的重要因素. 为了实现特征的有效聚合, Fan 等人[16]使用搜索注意力和部分解码器组件 (partial decoder component, PDC)[17]对粗糙区域进行多尺度特征的细化. 但这种密集连接的特征融合方式可能会导致计算冗余. Wang 等人[18]提出 D2CNet, 在 PDC 的基础上引入整体注意、残差注意机制来增强特征, 并采用优化后的 U-Net[19]结构融合对等层的局部信息进行细化. 但该模型在产生最终预测时直接引入第 3 层特征而未做任何处理, 可能会引入误导信息. Chen 等人[20]使用空洞空间金字塔池化 (atrous spatial pyramid pooling, ASPP) 模块的不同扩张卷积结构扩大特征图的感受野来进行多尺度特征的融合.

2 基于 Transformer 的跨尺度交互学习伪装目标检测方法

2.1 模型基本架构

本文提出一个新的伪装目标检测方法分别从图像的特征增强和增强上下文信息联系两方面入手, 共同优化伪装目标检测模型. 提出的方法的实现过程如图 2 所示. 该模型共包括 3 个模块, 分别是迭代注意力双分支特征融合模块、多尺度全局上下文信息模块、并行多通道池化模块. 首先利用视觉 Transformer 主干网络对输入图像进行特征提取, 高层次特征经过基于双层注意力的多尺度特征融合模块之后输入到多尺度全局上下文信息模块中. 经过该模块的混合卷积后, 中高层特征的上下文信息被增强, 增强后的图像再送入多通道池化模块聚焦局部信息. 最后与低层次特征进行联合叠加, 得到预测结果. 经过信息增强的特征不仅具有正确的概率分布, 能够被分类器正确分类, 特征更显著. 通过实验证明, 该方法能够准确实现对隐藏在背景中的物体进行检测. 下面将详细介绍伪装目标检测模型的设计.

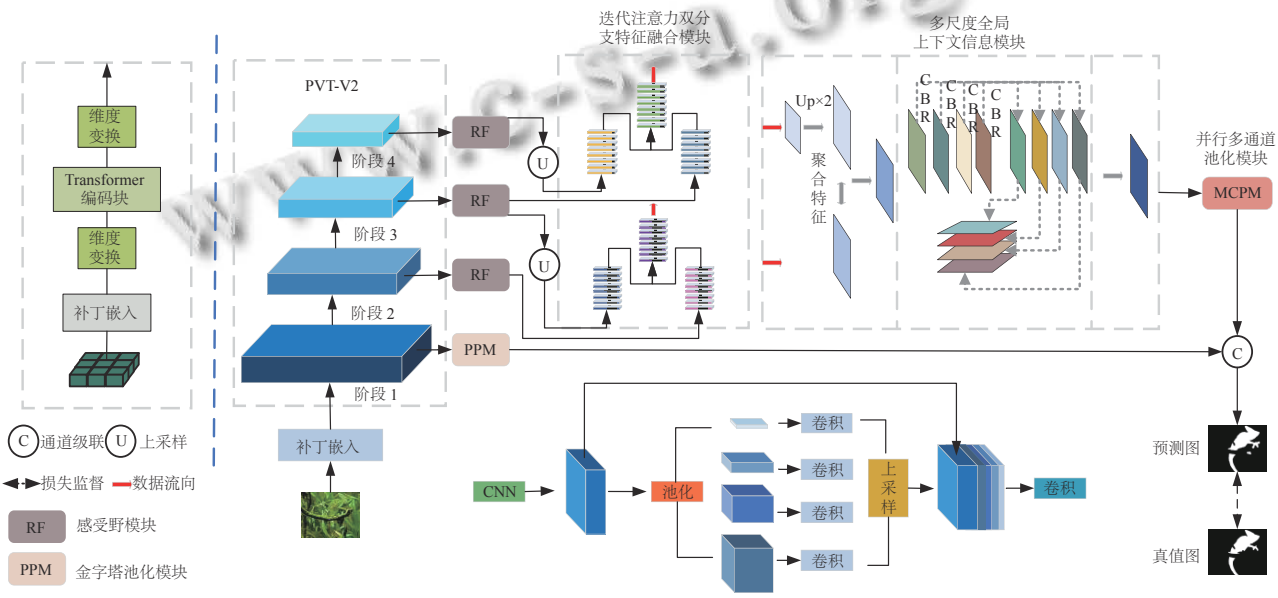


图 2 网络结构框架图

2.2 特征提取主干网络

由于现有的大部分伪装目标检测模型在提取特征时采用的都是深度卷积神经网络, 而 CNN 对图片的全局特征提取建模能力较差, 所以本文选用视觉 Transformer 主干网络进行特征提取. PVT 就是其中一种优秀的视觉 Transformer 编码器. 它采用渐进式收缩策略产生类似 CNN 的多尺度特征图, 利用空间裁剪注意力层降低计算代价. 经过改进的 PVT-V2 利用重叠切块, 保证局部连续性; 引入零填充位置编码, 以深度感知卷积适应不同分辨率输入; 改进空间裁剪注意力层为线性版本, 进一步降低多头自注意力的计算成本. 这些改进都保证了主干编码器抽取的特征兼具全局和局部特性, 表达能力更强. 本文采用 PVT-V2-b5 作为特征提取主干网络.

2.3 迭代注意力双分支特征融合模块

图片经过特征提取之后, 低层特征由于经过较少的卷积, 导致噪声较多, 但其分辨率更高, 且包含丰富的位置和细节信息. 而高级特征具有更强的语义信息, 但是分辨率较低, 对细节的感知能力较差. 为了能够将两者有效融合, 充分利用低级特征中的细节信息以及高级特征中的语义信息, 融合多尺度的跨级特征以提升网络性能是有必要的. 然而, 在融合跨级特征过程中, 由于尺度变化的问题, 往往会导致关键信息缺失的情况. 为了解决这一问题, 本文提出了迭代注意力双分支特征融合模块, 如图 3 所示, 通过引入 DBFF 模块更加细腻融合上文经过注意力引导融合的特征, 缓解由于多尺度变化而带来的不良影响, 提高伪装物检测的精确率.

迭代注意力双分支特征融合模块具体流程为: 经特征提取得到具有高分辨率 ($C1\times H1\times W1$) 特征 f_1 与具有低分辨率 ($C2\times H2\times W2$) 特征 f_2, f_2 经上采样与 f_1 相加传入 MSCA 模块, 以捕获隐藏在低级特征中的边缘特征, 再将经过 MSCA 模块的特征再次传入 MSCA 模块, 经过迭代注意力融合, 最后将特征融合输出得到特征 f'. 公式表示为:

$$f' = M(f_1 \oplus f_2)\otimes f_1 + (1 - M(f_1 \oplus f_2))\otimes f_2 \quad (1)$$

$$f_1 \oplus f_2 = M(f_1 + f_2)\otimes f_1 + (1 - M(f_1 + f_2))\otimes f_2 \quad (2)$$

其中, $f' \in R^{C\times H\times W}$ 是融合的特征, $\oplus$为初始特征集成, 图 3 中虚线表示为$1 - M(f_1 + f_2)$, 融合权重为$M(f_1 + f_2)$由 0 到 1 之间的实数组成.

迭代注意力模块充分融合多尺度信息后进入 DBFF 模块, 将经过迭代注意力的特征进行更加细腻的融合. DBFF 模块用不同的卷积核构造了一个双分支网络, 其中一个分支利用全局平均池化以获取全局上下文信息, 另一个分支保持了输入特征的大小, 以获取局部上下文信息. 在这种情况下, 双流网络之间的信息可以相互共享, 以捕捉不同尺度的特征. 将经过注意力引导融合的特征 f' 分为两个分支送入该模块中, 每个模块都经过 3×3 卷积层和 5×5 卷积层, 公式表示为:

$$f_{3\times3} = \zeta(Conv_3(\zeta(Conv_3(f_1'))\oplus\zeta(Conv_5(f_2')))) \quad (3)$$

$$f_{5\times5} = \zeta(Conv_5(\zeta(Conv_3(f_1'))\oplus\zeta(Conv_5(f_2')))) \quad (4)$$

其中, $\zeta(\cdot)$ 为 ReLU 激活函数.

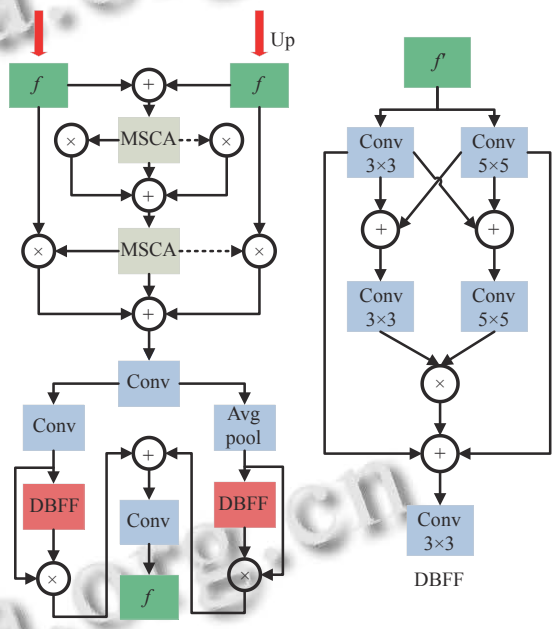


图 3 迭代注意力双分支特征融合模块

卷积过程中对不同层次的特征进行交互, 将经过各自卷积的特征相加再进行卷积, 最后相乘, 最终将全局和局部信息进行聚合, 完成细腻特征融合过程, 提升特征融合性能.

2.4 全局上下文信息模块

经过双分支特征融合模块融合过的特征包含丰富的上下文信息, 为了将上下文信息更好联系在一起, 全局上下文信息, 本文提出了 GCCM 模块. 如图 4 所示, 该模块将上文输出的两个特征通过 1×1 卷积进行跨通道的特征整合, 再均分为 4 个通道, 每个通道都进行 3×3、1×3、3×1 卷积和卷积核为 3×3 空洞率为 2 的空洞卷积的堆叠, 每个通道增强特征表示. 因为仅使用一个 3×3 卷积难以在一个阶段获得多尺度特征, 不利于

模型理解图像和分割, 所以在 3×3 卷积之后又采用了非对称卷积, 在提高检测精度的同时轻量化模型, 整体过程公式可表示为:

$$f'^{k}_{m} = F^{n_k}_{conv}\left(f'^{k-1}_{m} \oplus f'^{k}_{m} \oplus f'^{k+1}_{m}\right) \tag{5}$$

其中, $F^{n_k}_{conv}$ 代表每个通道进行的卷积, 在本实验中, 设置 $n_k = \{1,2,3,4\}$, 代表进行了 4 个通道的卷积堆叠.

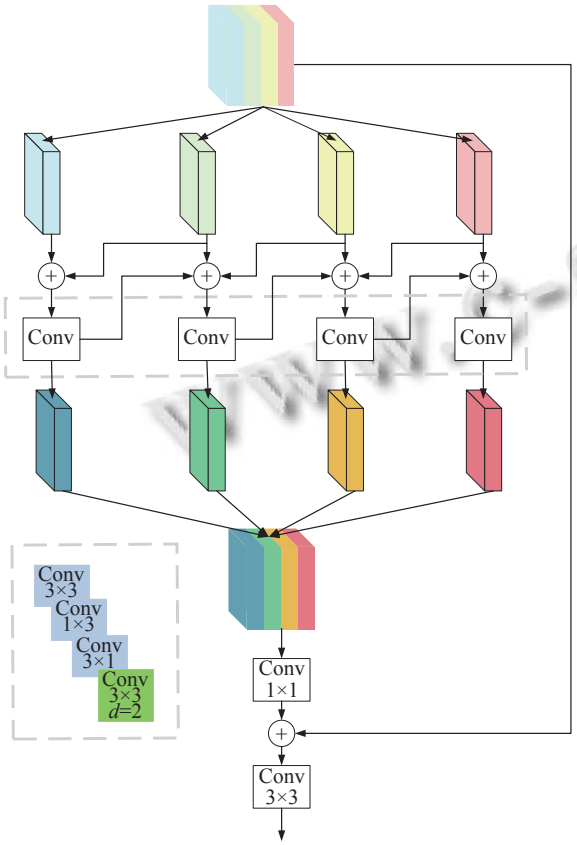


图 4 全局上下文信息模块

在对图像进行卷积提取特征时, 降低尺寸, 有助于减少计算量以及特征数量, 保留主要特征, 增大卷积核感受野, 防止过拟合, 所以将 3×3 卷积换成了不对称卷积 1×3 和 3×1. 为了捕获所有通道特征的依赖关系, 再将得到的信息通过 1×1 卷积融合在一起, 最后将初始输入特征与经过卷积处理的特征残差连接, 进行 3×3 的卷积整合信息:

$$\left[f'^{k}_{m}\right] = F_{conv1}\left(Cat^{4}_{k=1}\left(f'^{k}_{m}\right)\right) \tag{6}$$

$$f^{d}_{i} = F_{conv3}\left(\left(\left[f'^{k}_{m}\right]\right) \oplus f_{m}\right) \tag{7}$$

其中, $\left[f'^{k}_{m}\right]$表示经过 4 个堆叠卷积连接在一起的特征集合, f^{d}_{i} 代表经过该模块输出的最终特征, F_{conv1} 代表 1×1 卷积, F_{conv3} 代表 3×3 卷积.

2.5 多通道池化模块

由于视觉 Transformer 的局部特征提取信息建模能力较弱, 为了增强模型检测伪装物体的准确率, 本文提出了一个 MCPM 模块. 如图 5 所示, 该模块通过并行的池化操作, 可以聚焦局部特征, 弥补前面因重点关注全局信息而导致局部信息的缺失.

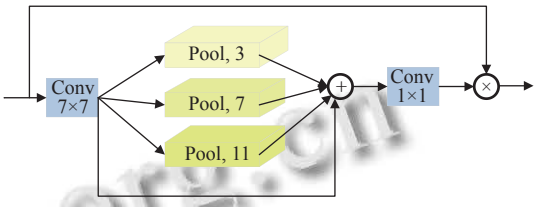


图 5 多通道池化模块

在网络最后几层, 最常见的是进入分类部分的全连接层前, 常常都使用平均池化. 这是因为最后几层都包含了比较丰富的语义信息, 使用平均池化会保留很多重要信息, 更好地聚焦局部信息. 因此在本文提出了 3 个并行的平均池化对上文已融合的高层特征进行操作, 进一步聚焦融合特征的局部信息.

对于来自 GCCM 的特征 $f^{d}_{i} \in C \times H \times W$, 首先应用 7×7 卷积层聚合局部信息.

$$f_{conv7} = F_{conv7}\left(f^{d}_{i}\right) \tag{8}$$

然后通过 3 个并行池化层用于捕获多尺度模态特征, 将这些特征与残差进行求和.

$$f_{\text{sum}} = \sum_{k\in\{3,\,7,\,11\}} Avgpool_{k\times k}\left(f_{conv7}\right) + f_{conv7} \tag{9}$$

并通过 1×1 卷积进行聚合信息.

$$f_{conv1} = F_{conv1}\left(f_{sum}\right) \tag{10}$$

利用 Sigmoid 函数计算注意力进行激活加权求和.

$$f_{M} = \sigma\left(f_{conv1}\right) \cdot f^{d}_{i} + f^{d}_{i} \tag{11}$$

其中, $\sigma(\cdot)$代表 Sigmoid 函数.

2.6 损失函数

本文采用了加权二值交叉熵损失L^{w}_{BCE}和加权交并比损失L^{w}_{IoU}作为该伪装目标检测模型的损失函数. 其中, 加权二值交叉熵损失函数可用于在神经网络中训练二元分类问题. 这个函数将输入的预测值与真实值进行比较, 并计算出预测值与真实值之间的差异, 以便更好地预测目标变量. 加权交并比损失函数在计算时, 模型预测结果通常是二值掩码, 因此需要计算两个掩码的交集和并集, 再计算交并比 (IoU) 来衡量相似度. 因此, 本文最终的总损失输出为:

$$L_{\text{total}} = L_{\text{BCE}}^{w} + L_{\text{IoU}}^{w} \tag{12}$$

3 实验评估

3.1 实验设置

实验均在 Cuda10.2, PyTorch 1.8, Python 3.6, GPU 为 NVIDIA GeForce RTX 3090 的环境配置下进行. 本文模型将所有输入图像大小调整为 352×352. 在训练阶段, batchsize 大小设置为 16, epoch 设置为 25, 并使用 Adax[21]优化算法, 通过将初始学习率设置为 1E–4 来优化整体参数.

3.2 数据集

实验在 3 个公开的伪装目标检测数据集上进行: CHAMELEON、CAMO 和 COD10K. CHAMELEON 数据集[19]是一个仅包含 76 张图像的公开数据集, 仅用于测试. CAMO 数据集包含伪装图像数据集和非伪装图像数据集两个子集. 本文采用伪装子集, 包含 1250 张图像, 其中 1000 张用于训练, 250 张用于测试. COD10K 数据集[17]是目前规模最大的伪装目标数据集, 包含 5066 张伪装图像, 其中 3040 张用于训练, 2026 张用于测试. 本文通过数据增强扩充数据集.

3.3 评估指标

分别针对 CHAMELEON、CAMO 和 COD10K 这 3 个数据集采用 4 个广泛使用的显著性监测评估指标结构度量 S-measure (S_α)、平均绝对误差 (MAE)、加权 F-measure (F_β^w) 、平均 E-measure (E_ϕ) 来判断本文方法的有效性.

1) 结构度量 S-measure (S_α)[22]用来评估预测图和真值图之间的结构相似性, 可以描述为:

$$S_\alpha = \alpha S_\beta + (1-\alpha) S_r \tag{13}$$

其中, S_β 计算目标感知, S_r 获取区域观测特征, α 和 β 是权重.

2) 平均 E-measure (E_ϕ)[23]通过比较预测图和真实图之间的差异来评估伪装目标检测结果的整体和局部精度, 可以描述为:

$$E_\phi = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} \phi(C(x,y) - G(x,y)) \tag{14}$$

其中, ϕ 是增强一致性矩阵, W 和 H 分别代表输入图像的宽度和高度, C 和 G 分别表示预测图和真值图.

3) 加权 F-measure (F_β^w) 是一种综合考虑加权查准率和加权查全率的整体度量, 可以表示为:

$$F_\beta^w = \frac{(\beta^2 + 1) PR}{\beta^2 P + R} \tag{15}$$

其中, P 代表精确率, R 代表召回率.

4) 平均绝对误差 (MAE) 评估预测图和真值图之间的像素级平均绝对误差, 可以表示为:

$$MAE = \frac{1}{H \times W} \sum_{x=1}^{H} \sum_{y=1}^{W} |(C(x,y) - G(x,y))| \tag{16}$$

3.4 实验结果与分析

在 CHAMELEON、CAMO、COD10K 这 3 个数据集下, 用本文提出的方法和当前先进的伪装目标检测算法进行对比. 通过观察模型的结构度量 S-measure (S_α)、平均绝对误差 (MAE)、加权 F-measure (F_β^w) 、平均 E-measure (E_ϕ) 来判断本文方法的有效性. 其中, SINet、PFNet、C2FNet 都是使用范围广泛, 比较经典的检测方法. 将本文的方法同这几个方法进行比较, 使得实验更具说服力. 本文方法在各个数据集上实验分析结果如表 1 和表 2 所示, 表 1 中 Ours-R 代表特征提取网络采取 Res2Net-50 模型, 表 2 中 Ours-P 代表特征提取网络采取 PVT-V2 模型. 表 2 中黑体数字表示最优值. 从表 2 数据可以看出: 本文方法检测精度明显优于其他模型检测精度, 随着模型性能的提升, 前 3 个指标的数值变大, MAE 的数值变小, 代表预测图与真值图更相似.

表 1 与基于卷积神经网络的不同方法在 3 个数据集上的指标结果比较

方法	会议年份	CHAMELEON				CAMO				COD10K			
		S_α	E_ϕ	F_β^w	MAE	S_α	E_ϕ	F_β^w	MAE	S_α	E_ϕ	F_β^w	MAE
EGNet[24]	ICCV2019	0.848	0.870	0.702	0.050	0.732	0.768	0.583	0.104	0.737	0.779	0.509	0.056
PraNet[25]	MICCAI2020	0.860	0.898	0.763	0.044	0.769	0.833	0.663	0.094	0.789	0.839	0.629	0.045
SINet[16]	CVPR2020	0.869	0.891	0.740	0.044	0.751	0.771	0.606	0.100	0.771	0.806	0.551	0.051
PFNet[8]	CVPR2021	0.882	0.930	0.810	0.033	0.782	0.840	0.695	0.085	0.800	0.868	0.660	0.040
C2FNet[7]	IJCAI2021	0.888	0.935	0.828	0.032	0.796	0.854	0.719	0.080	0.813	0.890	0.686	0.036
SINetV2[9]	TPAMI2022	0.888	0.930	0.816	0.030	0.820	0.875	0.743	0.070	0.815	0.863	0.680	0.037
FDNet[26]	CVPR2022	0.894	0.948	0.819	0.030	0.844	0.903	0.778	0.062	0.837	0.897	0.731	0.030
Ours-R	—	0.885	0.942	0.826	0.031	0.795	0.854	0.716	0.078	0.812	0.890	0.686	0.035

表 2 与基于 Transformer 的不同方法在 3 个数据集上的指标结果比较

方法	会议年份	CHAMELEON				CAMO				COD10K			
		S_α	E_ϕ	F_β^w	MAE	S_α	E_ϕ	F_β^w	MAE	S_α	E_ϕ	F_β^w	MAE
UGTR[27]	ICCV2021	0.888	0.918	0.796	0.031	0.785	0.859	0.686	0.086	0.818	0.850	0.667	0.035
COS-T[28]	ICCSIP2021	0.885	0.948	0.854	0.025	0.813	0.896	0.776	0.060	0.790	0.901	0.693	0.035
VST[29]	ICCV2021	0.888	0.936	0.820	0.033	0.805	0.863	0.780	0.069	0.810	0.866	0.680	0.035
ICON[30]	TPAMI2022	0.854	0.920	0.763	0.037	0.840	0.902	0.769	0.058	0.818	0.882	0.688	0.033
TPRNet[31]	TVCJ2022	0.891	0.930	0.816	0.031	0.814	0.870	0.781	0.076	0.829	0.892	0.725	0.034
DTINet[32]	ICPR2022	0.883	0.928	0.813	0.033	**0.857**	0.912	0.796	**0.050**	0.824	0.893	0.695	0.034
Ours-P	—	**0.906**	**0.953**	**0.860**	**0.024**	0.852	**0.914**	**0.802**	0.053	**0.845**	**0.919**	**0.746**	**0.027**

从图 6 分别展示出不同算法生成的图像, 可以看出本文训练生成的图像轮廓明显, 清晰度较高. 这主要得益于提出的基于迭代注意力的双分支特征融合模块, 通过该模块与另一个特征融合模块交替集成进行初始特征融合, 既集成了低层次的局部边缘信息和高层次的全局位置信息, 又可以将提取出来的特征充分融合. 本文模型生成的预测图在细节方面处理的更好, 这是因为全局上下文信息模块将提取出来的中高层特征通过多种卷积方式的堆叠进行增强及融合. 该模块使网络更加关注被检测对象的结构和细节, 通过一系列卷积运算来挖掘和聚合多尺度上下文语义, 从而生成更显著的特征. 然后, 从上到下逐步聚合多层融合特征, 预测伪装目标. 多通道池化模块可以很好地注意到局部特征信息使预测图的边缘细节信息更完整, 生成图片更加清晰.

图 6 算法生成图像对比图

本文通过对图像水平翻转、垂直翻转、垂直水平翻转以及 mix up 等数据增强方法对训练数据集进行扩充, 以达到 Transformer 训练时所需要的较大训练量. 图 7 是选取若干张图片进行数据增强的图片效果图.

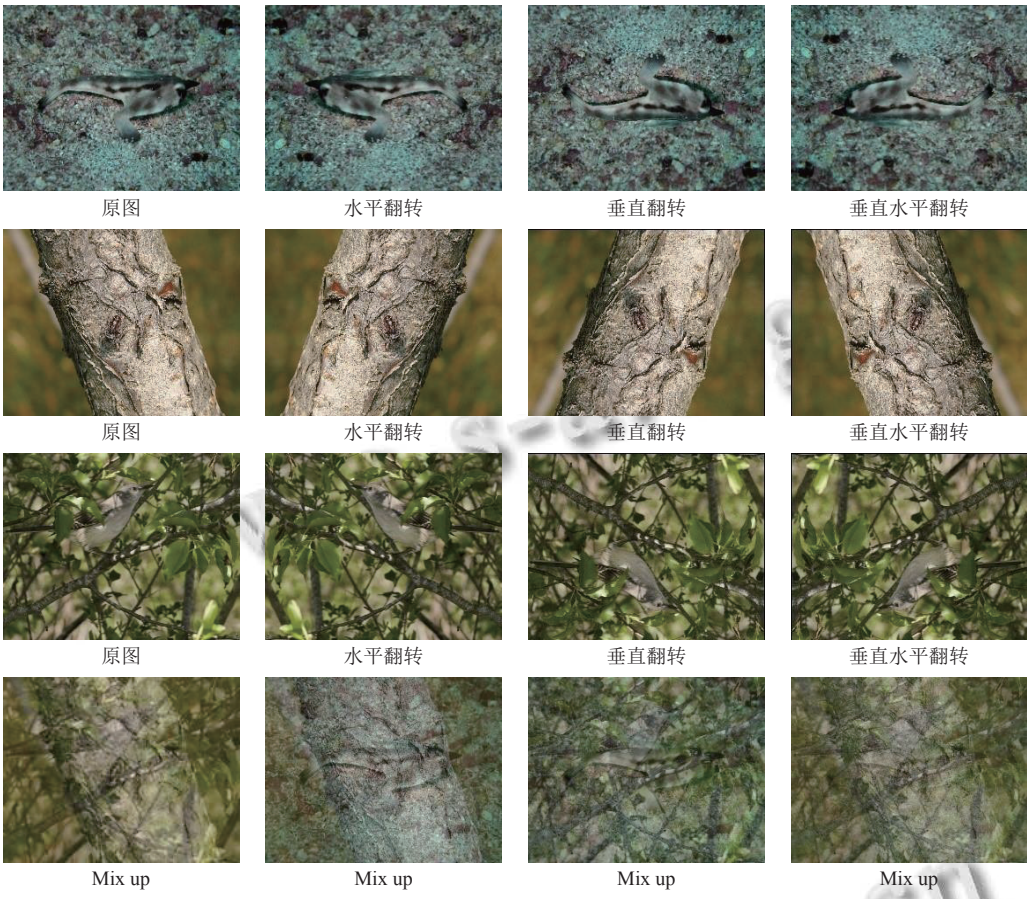


图 7 数据增强图片效果图

3.5 消融实验

为了验证本文提出的 3 个关键模块的有效性, 本节针对各个子模块进行了详细的消融实验分析. 消融实验分析结果如表 3 所示. 消融实验可视化结果如图 8. 方法 1 只采用了双分支特征融合模块中基于迭代注意力进行特征融合的部分, 特征融合之后直接得出预测图. 方法 2 在 Basic 基础上加入 GCCM 模块, 与方法 1 进行对比, 该模块能够全局上下文信息, 提高检测精度. 方法 3 在方法 2 基础上加入 DBFF 模块. 考虑到如果能细腻融合高级特征中丰富的语义信息和低级特征中丰富的细节信息能够帮助模型获取到更多的伪装目标位置及细节特征, 加入了 DBFF 模块的性能相比于方法 2 有所提升, 能够获取到更丰富的细节信息. 方法 4 在方法 3 基础上加入 MCPM 模块, 以验证所提出模块的有效性, MCPM 模块聚焦局部信息, 使生成图像边界更清晰, 与方法 3 对比可知, MCPM 模块能够通过对局部特征信息的处理提升算法性能, 说明 MCPM 模块的有效性.

表 3 本文方法在 3 个测试数据集上的消融实验

方法	CHAMELEON				CAMO				COD10K			
	S_α	E_ϕ	F_β^w	MAE	S_α	E_ϕ	F_β^w	MAE	S_α	E_ϕ	F_β^w	MAE
1	0.893	0.950	0.850	0.024	0.848	0.907	0.791	0.054	0.840	0.915	0.739	0.028
2	0.899	0.951	0.855	0.024	0.851	0.911	0.799	0.053	0.844	0.917	0.744	0.028
3	0.903	0.948	0.852	0.024	**0.858**	**0.919**	**0.812**	**0.050**	0.845	0.919	0.745	0.027
4	**0.906**	**0.953**	**0.860**	**0.024**	0.852	0.914	0.802	0.053	**0.845**	**0.919**	**0.746**	**0.027**

注: 1代表Basic, 2代表Basic+GCCM, 3代表Basic+DBFF+GCCM, 4代表Basic+DBFF+GCCM+MCPM

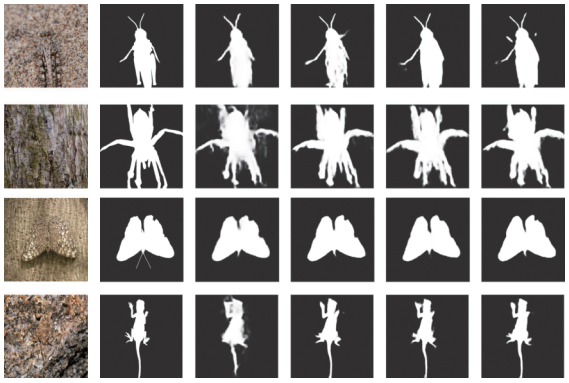


图 8 消融实验可视化结果图

此外, 为了验证多通道池化模块中池化核大小的不同给实验带来的影响, 本文还对该模块设置了不同的池化核组合在 3 个数据集上进行了实验, 表 4 中的结果表示, 在最具挑战性、场景最复杂的 COD10K 数据集上, 本文池化核的组合在 4 个指标表现得最好. 不是池化核设置得越大越好, 通过实验发现将池化核组合设置为{3, 7, 11}效果最好.

表 4 设置不同池化核的对比实验

数据集	池化核	S_α	E_ϕ	F_β^w	MAE
CHAMELEON	{3, 7, 11}	**0.906**	**0.953**	**0.860**	**0.024**
	{5, 9, 13}	0.895	0.947	0.843	0.026
	{7, 11, 21}	0.898	0.943	0.847	0.025
CAMO	{3, 7, 11}	0.852	**0.914**	**0.802**	**0.053**
	{5, 9, 13}	0.846	0.906	0.790	0.056
	{7, 11, 21}	**0.853**	0.912	0.801	0.053
COD10K	{3, 7, 11}	**0.845**	**0.919**	**0.746**	**0.027**
	{5, 9, 13}	0.839	0.916	0.733	0.029
	{7, 11, 21}	0.843	0.915	0.738	0.028

4 结论与展望

本文提出了新的伪装目标检测方法, 用来准确快速地检测出与背景融为一体的物体. 从伪装目标检测准确度的结果分析, 本方法明显优于当下的深度学习方法, 面对不同的图片, 本文的方法比目前的深度学习方法都使得目标图片中伪装的物体得到较高准确度的检测, 可以确保提取出来的伪装物体与真实值之间高度相似, 弥补了检测准确率不高的问题. 本文的相关实验结果证明了该伪装目标检测的有效性. 深度学习仍然是未来发展的必然趋势, 保证它能够安全地应用到各个领域是最关键的环节. 因此, 简单、高效的伪装物体检测是该领域未来的重点研究内容.

参考文献

1 Zhang X, Zhu C, Wang S, *et al*. A Bayesian approach to camouflaged moving object detection. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 27(9): 2001–2013. [doi: 10.1109/TCSVT.2016.2555719]

2 Beiderman Y, Teicher M, Garcia J, *et al*. Optical technique for classification, recognition and identification of obscured objects. Optics Communications, 2010, 283(21): 4274–4282. [doi: 10.1016/j.optcom.2010.06.059]

3 Galun M, Sharon E, Basri R, *et al*. Texture segmentation by multiscale aggregation of filter responses and shape elements. Proceedings of the 9th IEEE International Conference on Computer Vision. Nice: IEEE, 2003. 716–723.

4 Guo HX, Dou YL, Tian T, *et al*. A robust foreground segmentation method by temporal averaging multiple video frames. Proceedings of the 2008 International Conference on Audio, Language and Image Processing. Shanghai: IEEE, 2008. 878–882.

5 Chaduvula K, Rao BP, Govardhan A. An efficient content based image retrieval using color and texture of image sub-blocks. International Journal of Engineering Science and Technology, 2020, 3(2): 1060–1068.

6 Hall JR, Cuthill IC, Baddeley R, *et al*. Camouflage, detection and identification of moving targets. Proceedings of the Royal Society B: Biological Sciences, 2013, 280(1758): 20130064. [doi: 10.1098/rspb.2013.0064]

7 Sun YJ, Chen G, Zhou T, *et al*. Context-aware cross-level fusion network for camouflaged object detection. Proceedings of the 30th International Joint Conference on Artificial Intelligence. Montreal: IJCAI.org, 2021. 1025–1031.

8 Mei HY, Ji GP, Wei ZQ, *et al*. Camouflaged object segmentation with distraction mining. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 8772–8781.

9 Fan DP, Ji GP, Cheng MM, *et al*. Concealed object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(10): 6024–6042. [doi: 10.1109/TPAMI.2021.3085766]

10 Dai YM, Gieseke F, Oehmcke S, *et al*. Attentional feature fusion. Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2021. 3560–3569.

11 Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16×16 words: Transformers for image recognition at

scale. Proceedings of the 9th International Conference on Learning Representations. OpenReview.net, 2021.
12 Wang WH, Xie EZ, Li X, *et al*. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 548–558.
13 Wang WH, Xie EZ, Li X, *et al*. PVT v2: Improved baselines with pyramid vision transformer. Computational Visual Media, 2022, 8(3): 415–424. [doi: 10.1007/s41095-022-0274-8]
14 Liu ST, Huang D, Wang YH. Receptive field block net for accurate and fast object detection. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 385–400.
15 He KM, Zhang XY, Ren SQ, *et al*. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904–1916. [doi: 10.1109/TPAMI.2015.2389824]
16 Fan DP, Ji GP, Sun GL, *et al*. Camouflaged object detection. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 2774–2784.
17 Wu Z, Su L, Huang QM. Cascaded partial decoder for fast and accurate salient object detection. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3907–3916.
18 Wang K, Bi HB, Zhang Y, *et al*. D^{2}C-Net: A dual-branch, dual-guidance and cross-refine network for camouflaged object detection. IEEE Transactions on Industrial Electronics, 2022, 69(5): 5364–5374. [doi: 10.1109/TIE.2021.3078379]
19 Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. Proceedings of the 18th International Conference on Medical Image Computing and Computer-assisted Intervention. Munich: Springer, 2015. 234–241.
20 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]
21 Li WJ, Zhang ZY, Wang XJ, *et al*. AdaX: Adaptive gradient descent with exponential long term memory. arXiv:2004.09740, 2020.
22 Fan DP, Cheng MM, Liu Y, *et al*. Structure-measure: A new way to evaluate foreground maps. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 4548–4557.
23 Fan DP, Gong C, Gao Y, *et al*. Enhanced-alignment measure for binary foreground map evaluation. Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm: IJCAI.org, 2018. 698–704.
24 Zhao JX, Liu JJ, Fan DP, *et al*. EGNet: Edge guidance network for salient object detection. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 8779–8788.
25 Fan DP, Ji GP, Zhou T, *et al*. PraNet: Parallel reverse attention network for polyp segmentation. Proceedings of the 23rd International Conference on Medical Image Computing and Computer-Assisted Intervention. Lima: Springer, 2020. 263–273.
26 Zhong YJ, Li B, Tang L, *et al*. Detecting camouflaged object in frequency domain. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 4504–4513.
27 Yang F, Zhai Q, Li X, *et al*. Uncertainty-guided transformer reasoning for camouflaged object detection. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 4146–4155.
28 Wang HW, Wang XZ, Sun FC, *et al*. Camouflaged object segmentation with transformer. Proceedings of the 6th International Conference on Cognitive Systems and Signal Processing. Suzhou: Springer, 2022. 225–237.
29 Liu N, Zhang N, Wan KY, *et al*. Visual saliency transformer. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 4722–4732.
30 Zhuge MC, Fan DP, Liu N, *et al*. Salient object detection via integrity learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(3): 3738–3752.
31 Zhang Q, Ge YL, Zhang C, *et al*. TPRNet: Camouflaged object detection via transformer-induced progressive refinement network. The Visual Computer, 2023, 39(10): 4593–4607. [doi: 10.1007/s00371-022-02611-1]
32 Liu ZY, Zhang ZL, Tan YC, *et al*. Boosting camouflaged object detection with dual-task interactive transformer. Proceedings of the 26th International Conference on Pattern Recognition. Montreal: IEEE, 2022. 140–146.

(校对责编: 孙君艳)