CamoFormer: 用于伪装目标检测的掩码可分离注意力机制

Bowen Yin, Xuying Zhang, Deng-Ping Fan, *Senior Member, IEEE*, Shaohui Jiao, Ming-Ming Cheng, *Senior Member, IEEE*, Luc Van Gool, Qibin Hou

Abstract—如何从背景中识别并分割伪装目标是一项具有挑战性的任务。受Transformer 多头自注意力机制的启发，我们提出了一种简单的掩码可分离注意力（Masked Separable Attention, MSA）用于伪装目标检测。首先，我们将多头自注意力划分为三个部分，并通过不同的掩码策略来区分伪装目标和背景。此外，我们基于一个简单的自顶向下解码器，结合所提出的MSA，逐步捕获高分辨率的语义表示，以获得精确的分割结果。该结构与主干编码器相结合，形成了一种新模型，称为CamoFormer。大量实验表明，CamoFormer 在三个广泛使用的伪装目标检测基准上达到了新的最先进性能。为了更好地评估CamoFormer 在边界区域的检测性能，我们提出了两个新指标，即BR-M 和BR-F。 在S-measure 和加权F-measure 方面，相较于现有方法，CamoFormer 平均提升约**~5%**。我们的代码可在https://github.com/HVision-NKU/CamoFormer 获取。

Index Terms—伪装目标检测，自注意力，掩码可分离注意力，自顶向下解码器

1 引入

Foreground 和Background（前景/背景，FG/BG）分割技术 [5], [22], [39] 在计算机视觉领域中发挥着至关重要的作用，旨在准确区分并分离主要目标（前景）与其周围环境（背景），推动了目标识别 [3]、场景理解 [4], [33] 等领域的发展。伪装目标检测（Camouflaged Object Detection, COD）是一项新兴且具有挑战性的FG/BG 分割任务 [13]，近年来受到了广泛关注 [13], [14], [16], [26], [51]。生物学研究表明，人类视觉感知系统很容易被各种伪装策略所欺骗 [57]，因为伪装物体与其周围环境高度相似，或者尺寸极小。伪装物体与其周围环境的高度相似性使得伪装物体检测（COD） [14] 相较于传统目标检测任务 [38], [50] 更具挑战性。研究表明，该任务在不同领域的应用中具有重要价值，如艺术领域（例如，照片级真实感融合 [18] 和娱乐艺术 [8]）以及医学诊断（例如，息肉分割 [16]）。

越来越多的研究采用复杂的深度学习技术 [51], [52], [79] 来解决这一任务，尤其是在大规模数据集被提出后 [13]。然而，即使是最先进的方法，在处理某些复杂场景时，仍然难以准确分割出形状精细的伪装目标，这主要是由于伪装物体本身的特性。一些示例展示在 Fig. 1 中。一个更有前景的方法是分别编码前景和背景线索，并突出对比信息，而不是对前景和背景信息一视同仁地处理，以更好地识别伪装物体。

作为一种成功的尝试，PFNet [51] 表明，假阳性和假阴性预测通常自然地出现在分割结果中。为了解决这个问

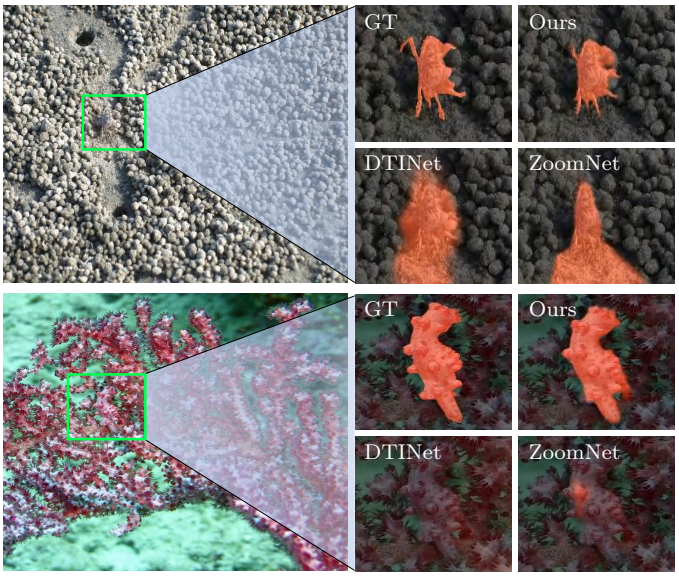


Fig. 1. 我们的方法CamoFormer 与近期先进方法（如DTINet [44] 和ZoomNet [52]）在伪装物体检测任务中的视觉比较。尽管近期的先进方法仍然难以捕捉伪装物体并区分其周围相似的背景，而我们的CamoFormer 能够更准确地捕捉和分割目标。最佳效果请在彩色模式下查看。

- 本研究得到了国家自然科学基金（*NSFC*，编号*62225604*，编号*62276145*）、中央高校基本科研业务费（南开大学，*070-63223049*）的资助。计算资源由南开大学超级计算中心（*NKSC*）提供支持。
- *B. Yin*、*X. Zhang*、*D.P. Fan*、*M.M. Cheng* 和*Q. Hou* 隶属于南开大学计算机学院视觉计算与智能感知实验室（*VCIP*），天津，中国（*bowenyin@mail.nankai.edu.cn, zhangxuying1004@gmail.com*）。*D.P. Fan*、*M.M. Cheng* 和*Q. Hou* 亦隶属于南开国际先进研究院（深圳福田）。
- *S. Jiao* 就职于字节跳动（*Bytedance Inc.*）。
- *L. V. Gool* 隶属于瑞士苏黎世联邦理工学院（*ETH Zurich*）计算机视觉实验室（*CVL*）。
- *Q. Hou* 为通讯作者（*andrewhoux@gmail.com*）。
- 前两位作者贡献相同。

Manuscript received March 1, 2022; revised August 26, 2022.

题，[51] 中提出了一种干扰挖掘策略，通过分别处理目标和背景的特征，从而去除这些错误预测。尽管PFNet 表现良好，但它仅关注于局部特征进行点级细化，而忽视了前景和背景特征之间相互作用的重要性。

在本文中，我们提出了*Masked Separable Attention (MSA)*，该方法从一个新的角度考虑了伪装物体和背景特征的编码方式。我们的MSA 构建在多头自注意力机制之上，但不同于传统方法仅仅利用多个注意力头来增强特征表示，我们提出利用不同的注意力头计算不同区域的像素相关性。具体来说，我们将自注意力头分成三组。我们首先使用两组头独立计算前景和背景区域的像素相关性。我们的目标是利用预测头内生成的前景的注意力得分，从完整的表示中索引伪装物体，背景部分也类似。此外，我们保留了一组正常的注意力头，用于计算完整图像的像素相关性，这可以帮助从全局视角区分伪装物体。因此，这三组头是互为补充的。

给定提出的MSA，我们将其应用于编码器-解码器架构 [36], [37], [40]，以逐步优化分割图，如 Fig. 2 所示。在解码器的每个特征层级，都会预测一个分割图，并将其送入MSA 块以提高预测质量。这一逐步优化过程使得随着特征分辨率的增加，我们能够获得高质量的伪装物体预测。如 Fig. 1 所示，我们的CamoFormer 能够比其他先进方法更准确地识别伪装物体并生成具有更精细边界的分割图。

为了验证CamoFormer的有效性，我们在三个流行的COD基准测试上进行了广泛的实验（NC4K [47]，COD10K [14]和CAMO [34]）。在所有这些基准测试中，我们的CamoFormer 相较于近期的先进方法，达到了新的最先进记录。特别地，我们的方法在COD10K 测试集上取得了0.793 的加权F 值和0.022 的MAE，而第二最佳模型FDNet [79] 的相应结果为0.731 和0.030。此外，我们还进行了全面的可视化实验，结果也显示了我们的CamoFormer在现有COD 方法中的优越性。

除了传统的评估指标外，考虑到伪装物体与其周围区域非常相似且尺度通常较小，我们还测量了伪装物体边界区域的分割质量。具体来说，我们提出了计算伪装物体边界区域的加权F 值和平均绝对误差得分，从而得出了两个新的评估指标，即BR-M 和BR-F。实验表明，在这两个边界区域的评估中，我们的CamoFormer 的表现甚至超过了其他最先进的方法。

我们的主要贡献可以总结如下：

- 我们提出了掩蔽可分离注意力（MSA），这是一种新颖的方法，通过使用不同的自注意力头来计算不同区域的视觉相似度，并同时显式建模前景和背景之间的全局依赖关系。
- 我们提出了一种新的网络架构，命名为CamoFormer，其中构建了一个自上而下的路径来充分挖掘我们的MSA 的潜力。实验结果表明，我们的方法比以往的工作表现更好。
- 我们提出了两个新的简单评估指标，用于评估伪装物体检测模型在边界区域的表现，并证明我们的方法在处理边界区域时表现更好。

2 相关工作

2.1 伪装物体检测

传统的COD 方法 [1], [17], [20], [21], [31], [77] 提取伪装物体与背景之间的各种手工特征来分割伪装目标。这些方法能够处理简单场景，但在复杂条件下准确性会急剧下降。通过深度学习方法 [6], [15], [45], [46], [76] 发展这一领域已成为当前的趋势。

最近，COD 主流方法是基于CNN 的方法 [14], [27], [28], [51], [52], [74], [79]，这些方法可以分为三种策略：i) 多尺度特征聚合：CubeNet [82] 结合注意力融合和X 形连接，充分整合来自多个层的特征。ZoomNet [52] 在三个尺度下处理输入图像，并统一不同尺度的特定外观特征。ii) 多阶段策略：由于伪装物体的隐蔽性，SINet [14] 提出了先定位后区分的方法，以提高性能。PreyNet [74] 模拟捕食过程，将伪装目标的检测过程分为初步检测和捕食者学习。SINetV2 [13] 采用周围连接解码器和组反转注意力来提高性能。SegMaR [28] 是一个多阶段训练和推理框架，先定位目标，然后放大物体区域，逐步检测伪装物体。iii) 联合训练策略：UJSC [35] 利用矛盾信息增强显著物体检测和伪装物体检测的检测能力。SLSR [47] 将伪装排名和伪装物体检测结合起来，构建了联合训练框架。然而，这些方法对于形状复杂的伪装物体仍然力不从心。

我们的工作与PFNet [51] 的工作也有关。然而，PFNet 仅通过干扰挖掘策略将特征图进行分割，伪装物体和背景的特征被分别处理以去除错误预测。与这种方法不同，我们的CamoFormer 利用自注意力机制，并通过不同的自注意力头计算不同区域的视觉相似性，从整个特征图中索引目标。

2.2 计算机视觉中的Transformers

与传统的卷积神经网络 [23], [25], [55], [56], [60]相比，Transformer能够有效地编码全局上下文信息，因此在各种视觉任务中得到了广泛应用，包括图像分类 [10], [61], [64], [70]、语义分割 [30], [66], [69], [78]、目标检测 [2]、超分辨率 [80]和显著物体检测 [19], [41], [81]。

基于Transformer 的模型也正在成为COD 中的一个新趋势。UGTR [68] 明确利用概率表示模型学习伪装物体在Transformer 框架下的不确定性。DTINet [44] 设计了一个双任务交互Transformer 来分割伪装物体及其详细边界。TPRNet [75] 提出了一个基于Transformer 的渐进式优化网络，利用高层特征的语义信息引导伪装目标的检测。此外，在基于Transformer 的框架下，SLTNet [7] 利用短期动态和长期时间一致性捕捉视频中的动态伪装物体。

我们的CamoFormer 也是基于流行的Transformer 框架构建的。我们并不专注于新的架构设计，而是旨在研究更高效地利用自注意力进行COD 的方法，并取得比其他方法更好

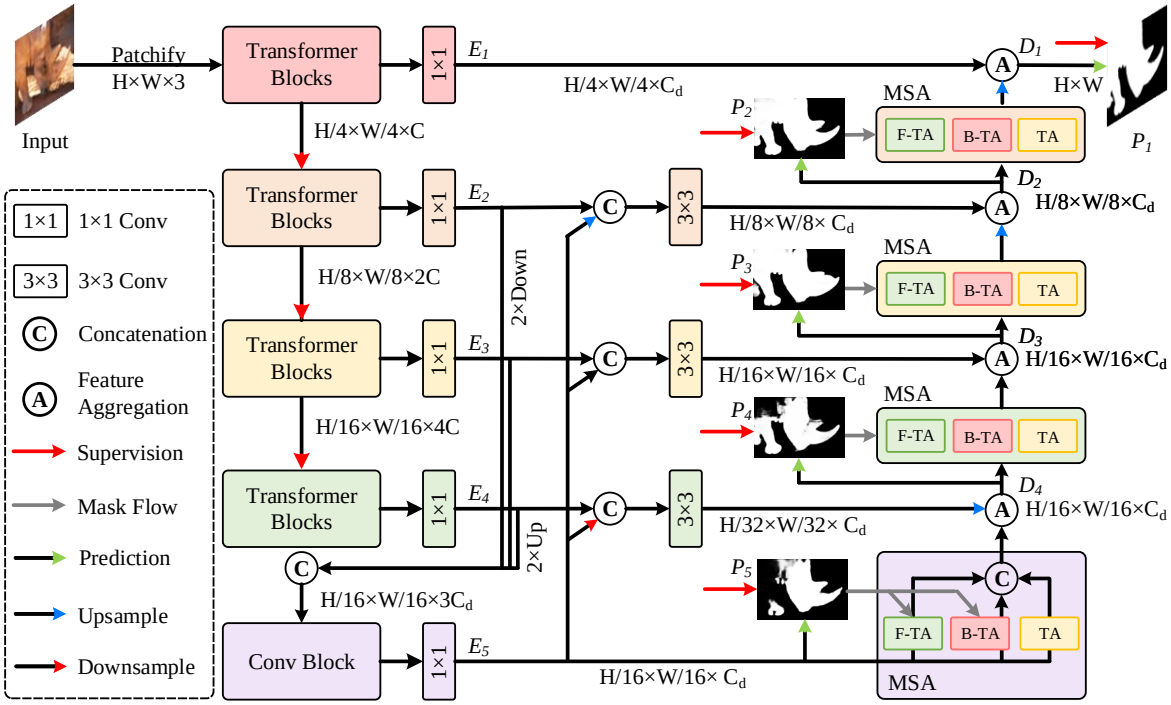


Fig. 2. 我们CamoFormer 模型的整体架构。首先，利用一个预训练的基于Transformer 的主干网络提取输入图像的多尺度特征。然后，将最后三个阶段的特征聚合生成粗略预测。接下来，配备掩蔽可分离注意力（MSA）的渐进式优化解码器被应用于逐步完善预测结果。F-TA、B-TA 和TA 头分别计算预测的前景、背景和整个图像中的注意力分数，MSA 利用这些分数更好地识别伪装物体。所有生成的预测都由真实标签进行监督。

的性能。我们为不同的注意力头分配不同的功能，分别处理前景和背景区域，这使得我们的工作与其他基于Transformer的COD 方法有所不同。

3 CamoFormer的提出

3.1 整体架构

与大多数先前的工作 [14], [16], [29], [52], [79]类似，我们采用了一个编码器-解码器架构来构建我们的CamoFormer，如Fig. 2 所示。

编码器。 默认情况下，我们采用PVTv2 [65] 作为我们的编码器，因为视觉Transformer 在二值分割任务中表现出了优异的性能 [41], [73]。给定输入图像$I \in \mathbb{R}^{H\times W\times 3}$，我们将其输入编码器，生成来自四个阶段的多尺度特征图，记作$\{E_i\}_{i=1}^{4}$。因此，E_1 的空间大小为$\frac{H}{4}\times\frac{W}{4}$，而$E_4$ 的空间大小为$\frac{H}{32}\times\frac{W}{32}$。然后，我们将来自编码器最后三个阶段的特征聚合，并将其送入卷积块，得到具有更高级语义的表示E_5。

解码器。 解码器建立在编码器基础上。编码器中的多层语义特征$\{E_i\}_{i=1}^{5}$ 被送入解码器。为了在效率和性能之间取得更好的平衡，我们首先在每个层次的特征图上连接一个1×1 卷积层，通道数为$C_d = 128$。如Fig. 2 所示，我们采用渐进式方式来细化来自编码器顶部的特征。在每个特征层次上，使用掩码可分离注意力（MSA）以更好地区分伪装目标和背景。在渐进融合的初始层次上，聚合的特征D_4 可以表示为：

$$D_4 = \mathrm{MSA}(E_5)\cdot\mathcal{F}_{\mathrm{up}}(E_4) + \mathcal{F}_{\mathrm{up}}(E_4), \tag{1}$$

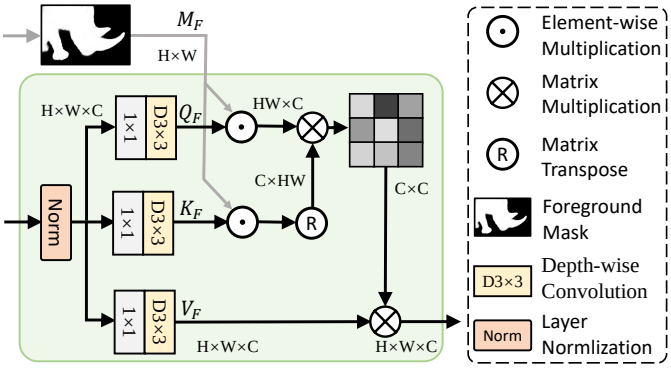


Fig. 3. 我们MSA 中提出的F-TA 的示意图。我们的B-TA 具有相似的结构，区别在于掩码部分。

其中，$\mathcal{F}_{\mathrm{up}}(\cdot)$ 是形状匹配的双线性上采样操作。在接下来的层次中，聚合的特征$\{D_i\}_{i=1}^{3}$ 可以表示为：

$$D_i = \mathcal{F}_{\mathrm{up}}(\mathrm{MSA}(D_{i+1}))\cdot E_i + E_i. \tag{2}$$

与先前的工作 [14], [28], [52] 主要使用加法操作或拼接操作来融合不同特征层次的特征不同，我们首先计算它们之间的逐元素乘积，然后再使用加法操作。我们通过实验发现，这样的简单修改在NC4K [47]、COD10K-test [13]和CAMO-test [34] 上，S-measure 和加权F-measure 平均提升了0.2%+。

损失函数。 遵循 [24], [67]，我们在每个特征层次添加了侧面监督。我们将CamoFormer 解码器生成的预测表示为$\{P_i\}_{i=1}^{5}$。除了最终的预测图P_1 外，其他所有预测图P_i都用于上述的MSA 进行渐进式细化。在训练过程中，每个P_i

会被重新缩放到与输入图像相同的大小，所有预测图都使用BCE 损失 [9] 和IoU 损失 [49]进行监督。遵循 [13]，整体损失是多阶段损失的总和。我们CamoFormer 的总损失可以表示为：

$$\mathcal{L}(P,G)=\sum_{i=1}^{5}\mathcal{L}_{bce}(P_i,G)+\mathcal{L}_{iou}(P_i,G), \tag{3}$$

其中G 是真实标注。

3.2 掩码可分离注意力

伪装物体在尺度上多样化，并且与背景高度相似，这使得它们难以完全分割。如何准确地从背景中识别伪装物体是至关重要的。我们通过提出掩码可分离注意力（MSA）来解决这个问题，其中不同的注意力头负责不同的功能。我们打算使用部分注意力头分别计算在预测的前景和背景区域中的注意力得分，并利用这些得分更好地识别伪装物体。

我们的MSA基于修改版的自注意力机制，以节省计算量，具体来说是多通道深度卷积头转置注意力（Multi-Dconv Head Transposed Attention）[71]，我们简称为TA。给定输入$\mathbf{X}\in\mathbb{R}^{HW\times C}$，其中$H$和$W$分别是高度和宽度，$C$是通道数，TA可以表示为：

$$\text{TA}(\mathbf{Q},\mathbf{K},\mathbf{V})=\mathbf{V}\cdot\text{Softmax}(\frac{\mathbf{Q}^\top\mathbf{K}}{\alpha}), \tag{4}$$

其中$\mathbf{Q}$、$\mathbf{K}$和$\mathbf{V}$分别是查询、键和值矩阵，可以通过三个分别的1×1卷积层后接3×3深度卷积生成，α是一个可学习的缩放参数。在实际应用中，Eqn. 4也可以扩展为多头版本，正如原始自注意力机制 [62]所做的那样，以增强特征表示。

掩码可分离注意力。 上述TA中的注意力头在编码空间信息时被均等使用。与此不同，在我们的MSA中，我们提出引入一个预测掩码，该掩码可以在每个特征层生成，并作为前景-背景对比先验，以更好地识别伪装物体。为此，我们将所有注意力头分为三组：前景头TA（F-TA）、背景头TA（B-TA）和普通TA。我们MSA的结构细节如Fig. 3所示。

具体而言，给定一个预测的前景掩码M_F，F-TA的公式可以写为：

$$\text{F-TA}(\mathbf{Q}_F,\mathbf{K}_F,\mathbf{V}_F)=\mathbf{V}_F\cdot\text{Softmax}(\frac{\mathbf{Q}_F^\top\mathbf{K}_F}{\alpha_F}), \tag{5}$$

其中$\mathbf{Q}_F$、$\mathbf{K}_F$是掩码后的查询和键矩阵，可以通过与M_F相乘来生成，$\mathbf{V}_F$是未经掩码处理的值矩阵。通过这种方式，特征可以通过在前景区域内建立成对关系来进行细化，从而避免背景中可能包含的污染信息的影响。类似地，给定背景掩码通过广播减法$M_B=1-M_F$，我们也可以对背景进行此过程。因此，B-TA的公式可以写为：

$$\text{B-TA}(\mathbf{Q}_B,\mathbf{K}_B,\mathbf{V}_B)=\mathbf{V}_B\cdot\text{Softmax}(\frac{\mathbf{Q}_B^\top\mathbf{K}_B}{\alpha_B}). \tag{6}$$

除了F-TA头和B-TA头外，第三组头部保持与Eqn. 4中的结构不变，用于建立前景与背景之间的关系。然后，所有头部的输出被拼接，并送入一个3×3卷积层进行特征聚合，并将通道数映射到C_d：

$$\mathbf{Z}=\text{Conv}_{3\times3}([\text{F-TA},\text{B-TA},\text{TA}]), \tag{7}$$

其中$[\cdots]$表示拼接操作。

Fig. 4. 边界区域的示意图。‘αBR@15’ 和‘αBR@30’ 是通过分别使用r和C=15 的膨胀核、r 和C=30 的膨胀核对GT 的边界进行膨胀生成的自适应边界区域。

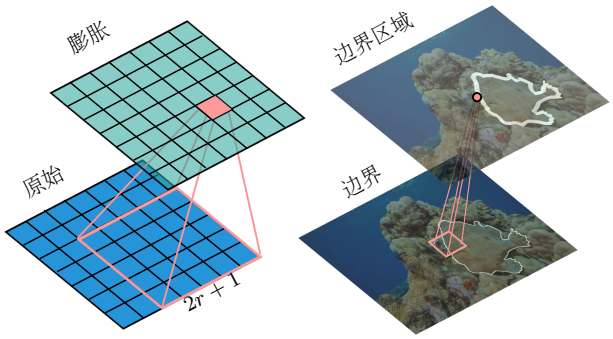


Fig. 5. 膨胀操作的可视化。左图：(i,j) 处的膨胀操作，膨胀率为5。右图：膨胀操作在图像上的可视化。

掩码生成。 在每个特征层级，通过3×3 卷积后跟Sigmoid 函数生成一个掩码，并在我们的MSA 中使用该掩码。由于在每个特征层级都添加了监督，我们直接使用预测值$\{P_i\}_{i=2}^{5}$ 作为掩码，并将每个掩码发送到相应的MSA。注意，我们并未将预测图像二值化，而是保留它们作为从0 到1 的连续图像，这在我们的实验中表现得更好。

4 边界区域测量

如在 Sec. 1中提到，伪装物体通常与其周围环境共享相似的模式，使得物体边界周围的区域难以识别。在本节中，我们提出使用加权F 值（wF）和平均绝对误差（M）来评估伪装

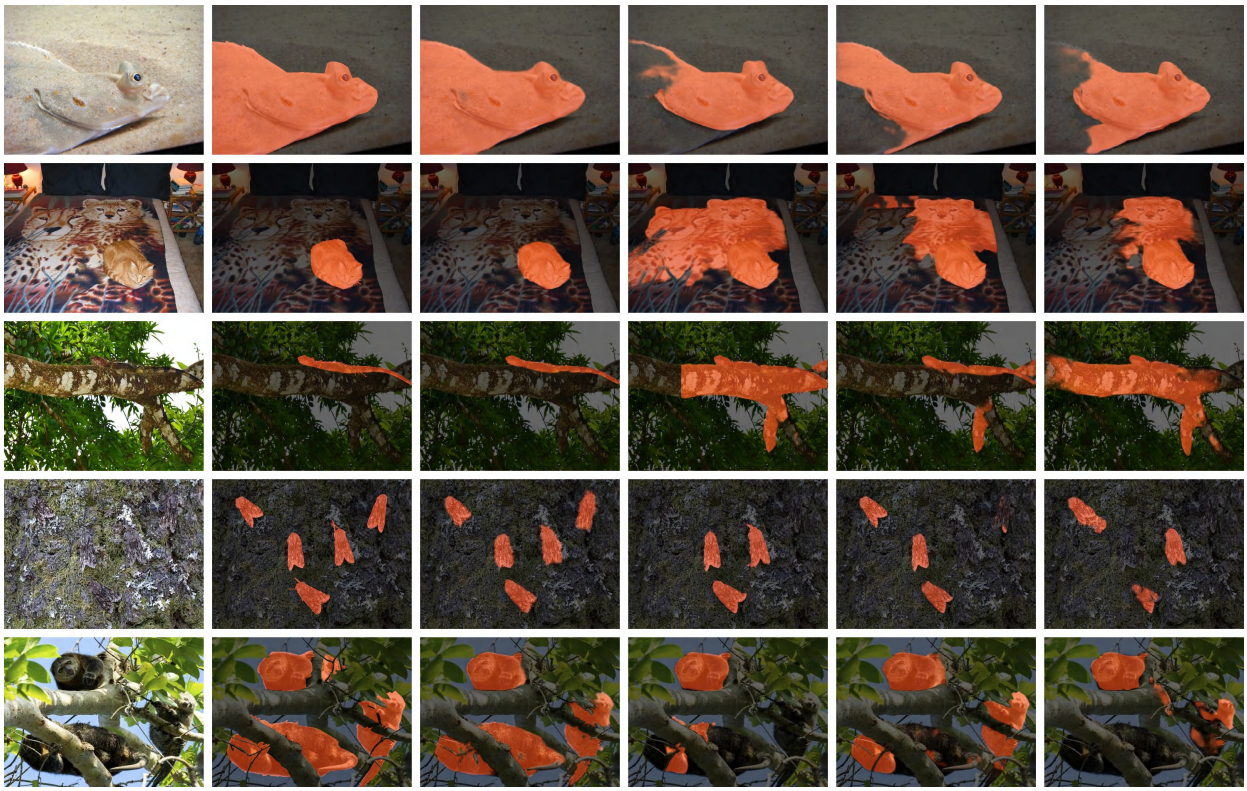


Fig. 6. 我们CamoFormer与其他SOTA方法的可视化对比。分割结果以橙色显示。

区域边界周围的分割结果，从而得出两个新指标：自适应边界区域约束加权F 值（BR-F）和平均绝对误差（BR-M）。这两个指标的详细计算过程如下所述。

边界区域生成。 给定地面真值（GT），边界$B \in \{0,1\}^{H\times W}$ 通过二值图像边界搜索方法 [59]获得，如 Fig. 4所示。我们尝试在相对较大的区域上进行测量，这有助于观察边界的纹理偏差。边界区域是通过膨胀GT 边界生成的，如 Fig. 5所示。具体来说，膨胀后位置(i,j) 的值可以表示为：

$$BR(i,j) = \text{Max}(B((i,j),r)), \tag{8}$$

其中$B((i,j),r)$ 是边界图中以位置(i,j) 为中心，边长为$2r+1$ 的正方形区域。考虑到目标大小的多样性，固定宽度的边界区域并不适用。因此，我们根据物体的大小计算膨胀率r：

$$r = \sum G/(\alpha \times H \times W), \tag{9}$$

其中$G \in [0,1]^{H\times W}$ 是图像的地面真值，α 是一个缩放因子，用于将r 缩放到接近1。我们设置α 为0.05。然后，$C \cdot r$ 就是边界区域的膨胀宽度，其中C 是基础宽度，r 根据样本调整。

整个边界区域掩码$BR \in \{0,1\}^{H\times W}$ 可以通过膨胀边界并使用每个伪装区域边界位置对应的膨胀宽度$C \cdot r$ 来生成。我们将基础膨胀宽度为C 的自适应边界区域表示为‘αBR@C’，并在 Fig. 4中可视化了边界区域。

指标计算。 对于指标BR-M，我们在选定的边界区域计算M [54] 值。给定边界区域掩码BR，BR-M 可以定义为：

$$\text{BR-M} = \frac{\sum |P-G| \cdot BR}{\sum BR}, \tag{10}$$

其中P 和G 分别是预测图像和相应的GT，它们的形状与BR 相同。

与BR-M相似，指标BR-F 通过计算选定边界区域的加权F 值 [48] 实现。根据 [48]，加权误差图E^w 根据P 和G 计算。我们在BR-F 中的四个基本量，即真阳性（TP）、真阴性（TN）、假阳性（FP）和假阴性（FN），可以定义如下：

$$\begin{aligned} TP^{BR} &= BR \cdot (1-E^w) \cdot G, \\ TN^{BR} &= BR \cdot (1-E^w) \cdot (1-G), \\ FP^{BR} &= BR \cdot E^w \cdot (1-G), \\ FN^{BR} &= BR \cdot E^w \cdot G. \end{aligned} \tag{11}$$

注意，我们的量与原始量的区别在于我们的计算过程仅限于边界区域。边界区域约束的加权精度和召回率可以表示为：

$$\begin{aligned} \text{Precision}^{BR} &= \frac{TP^{BR}}{TP^{BR}+FP^{BR}}, \\ \text{Recall}^{BR} &= \frac{TP^{BR}}{TP^{BR}+FN^{BR}}. \end{aligned} \tag{12}$$

最后，BR-F 衡量指标定义为：

$$\text{BR-F} = (1+\beta^2)\frac{\text{Precision}^{BR} \cdot \text{Recall}^{BR}}{\beta^2 \cdot \text{Precision}^{BR} + \text{Recall}^{BR}}. \tag{13}$$

5 实验结果

5.1 实验设置

实现细节。 我们使用Pytorch 库 [53] 实现了我们的CamoFormer。一个在ImageNet 数据集 [32] 上预训练的PVTv2 [65] 被用作我们网络的编码器。除非另有说明，我们采用PVTv2 [65] 作为主干网络。此外，我们还报告了使用其他主干网络的结果，例如基于Transformer 的Swin Transformer [42]、基于CNN 的ResNet [23] 和ConvNeXt [43]。我们使用带动量0.9 和权重衰减2e-4 的SGD 作为优化器。学习率初始设置为5e-3，并采用余弦学习率衰减策略。在训练过程中，所有输入图像都调整为384×384。整个模型在NVIDIA V100 GPU 上训练60 个epoch，总共耗时约7 小时，批量大小为6。

数据集。 我们在三个流行的COD 基准上评估我们的方法，包括CAMO [34]、COD10K [13] 和NC4k [47]。CAMO 包含2,500 张图像，其中一半含有伪装物体，另一半不含。COD10K 包括5,066 张伪装图像、3,000 张背景图像和1,934 张非伪装图像。NC4K 是一个大规模的COD 数据集，包含4,121 张测试图像。按照以前的工作 [14], [28], [52]，我们使用CAMO 数据集中的1,000 张图像和COD10K 数据集中的3,040 张图像用于训练，其余的用于测试。

评估指标。 按照 [28], [52], [79]，我们使用四个标准指标进行评估，包括结构测量（S_m） [11]、均值绝对误差（M） [54]、加权F 测量（wF） [48] 和自适应E 测量（αE） [12]。M 是预测图像和GT 之间的绝对差异。S_m 同时评估预测图像与GT 之间的区域感知和物体感知结构相似性。wF 是对召回率和精确度的全面评估。αE 评估元素级相似性和图像级统计信息。此外，我们绘制了精度-召回率（PR）曲线、F_β 阈值（F_β）曲线（见Fig. 8）和假阴性比率（FNR）曲线（见Fig. 9）。

5.2 定性评估

预测结果的可视化。 Fig. 6 展示了我们的方法CamoFormer 与三种之前的SOTA 方法的可视化样本。为了更好地展示这些模型的表现，我们选择了包含不同复杂场景的几个典型样本，来自于伪装目标检测（COD）领域。如图所示，第一行的结果表明，其他方法仍然难以从相似的背景中准确感知伪装目标。有时，由于缺乏全局对比信息，模型也很难识别出伪装目标，如下两行所示。总的来说，在处理复杂条件时，这些方法因缺乏对前景和背景的全面理解，导致部分区域被误判或目标的某些部分被漏掉。相比之下，通过显式地感知前景-背景线索，我们的CamoFormer 即使在复杂条件下，也能生成高质量的伪装目标分割图。

物体边界质量比较。 伪装目标有时具有独特的形状，如Fig. 7所示。为了展示我们的CamoFormer 在处理这些类型的物体时的表现，我们展示了一些预测结果并在Fig. 7中用白色曲线标示了真实边界（GT）。我们方法预测的边界更接近真实物体的边界，而其他方法的预测则存在明显偏差。这些可视化结果表明，我们的模型能够分割出更精确的伪装目标。

5.3 定量评估

我们将CamoFormer与12个基于CNN的最新COD模型进行比较，包括ZoomNet [52]、FDNet [79]、SegMaR [28]、DGNet [26]、SINetV2 [13]、C^2FNet [58]、UJSC [35]、PFNet [51]、MGL-R [72]、SLSR [47]、SINet [14] 和PraNet [16]，以及6 个基于Transformer 的方法，包括COST [63]、TPRNet [75]、VST [41]、DTINet [44]、UGTR [68] 和ICON [81]。为了公平比较，所有预测结果均由各方法作者提供或通过其经过良好训练的模型生成。

在目标区域上的表现。 如 Tab. 1所示，我们提出的CamoFormer 在所有三个基准测试中始终显著超越了以前的方法，且在训练过程中未使用任何后处理技巧或额外的数据。与最近的基于CNN 的COD 方法（如ZoomNet [52]、FDNet [79] 和SegMaR [28]）相比，尽管它们采用了多阶段训练和推理等策略，这些方法在计算开销上较大，但我们的CamoFormer 仍在所有基准测试中以较大优势超越它们。与此同时，与基于Transformer 的模型（如TPRNet [75] 和DTINet [44]）相比，我们的方法同样表现得更好。

边界区域的表现。 由于伪装目标的边界形状不规则且与周围环境高度相似，因此其边界难以检测。为了量化边界区域的分割表现，我们分别计算了在 Sec. 4中描述的BR-F 和BR-M 分数。边界区域的大小取决于膨胀操作的卷积核大小。Tab. 2 展示了在边界区域内计算的表现，分别为‘αBR@15’ 和‘αBR@30’。值得注意的是，CamoFormer 在边界区域表现上显著优于其他方法，表明我们的方法在GT 对象边界的预测效果更好，而其他方法的预测存在显著偏差。

COD 方法的PR 和F_β 曲线。 我们展示了在NC4K、CAMO 和COD10K 数据集上，我们的CamoFormer 和之前方法的PR 和F_β 曲线，如 Fig. 8 所示。请注意，曲线越高，模型的表现越好。可以明显看出，我们的CamoFormer（红色曲线）超越了所有其他方法。

计算成本比较。 Tab. 3 展示了我们方法CamoFormer 和最近SOTA 方法的参数和MACs 成本。可以看出，与其他方法相比，我们的CamoFormer 在计算成本上是可接受的，但获得了更好的结果。

假阴性比例(FNRs)。 我们还采用了FNR [81] 来直观展示我们的CamoFormer 相较于其他前沿COD 模型的优越性。请注意，假阴性度量指的是目标物体上错误预测的像素，FNR [81] 旨在计算在目标物体上的错误预测像素占该物体所有像素的比例。FNR 越低，模型越好。具体而言，FNR

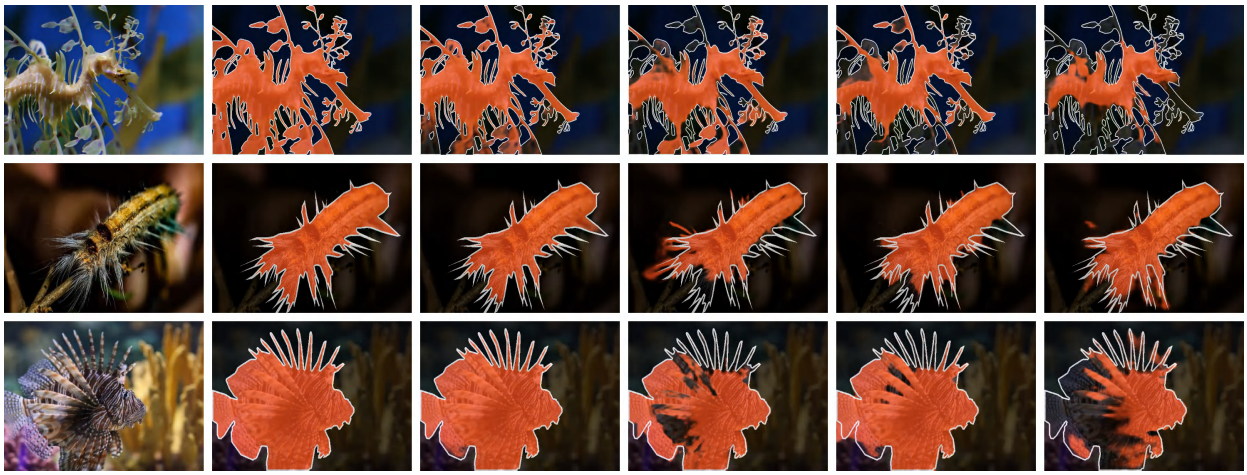


Fig. 7. 我们的方法CamoFormer 与其他SOTA 方法在分割边界上的比较。GT 的边界以白色标记，预测的边界以橙色标记。

TABLE 1

我们的方法CamoFormer 与最近的SOTA 方法的比较。'-R'：ResNet50 [23]，'-C'：ConvNext-base [43]，'-S'：Swin Transformer-base [42]，'-P'：PVTv2-b4 [65]。从表中可以看出，我们的方法CamoFormer-P 在使用CNN 或Transformer 模型时，比之前的方法表现更好。'↑'：越高越好，'↓'：越低越好。

方法	NC4K (4,121)				COD10K-Test (2,026)				CAMO-Test (250)			
	$S_m\uparrow$	$\alpha E\uparrow$	$wF\uparrow$	M↓	$S_m\uparrow$	$\alpha E\uparrow$	$wF\uparrow$	M↓	$S_m\uparrow$	$\alpha E\uparrow$	$wF\uparrow$	M↓
基于*CNN* 的方法												
PraNet$_{2020}$ [16]	0.822	0.871	0.724	0.059	0.789	0.839	0.629	0.045	0.769	0.833	0.663	0.094
SINet$_{2020}$ [14]	0.808	0.883	0.723	0.058	0.776	0.867	0.631	0.043	0.745	0.825	0.644	0.092
SLSR$_{2021}$ [47]	0.840	0.902	0.766	0.048	0.804	0.882	0.673	0.037	0.787	0.855	0.696	0.080
MGL-R$_{2021}$ [72]	0.833	0.893	0.739	0.053	0.814	0.865	0.666	0.035	0.782	0.847	0.695	0.085
PFNet$_{2021}$ [51]	0.829	0.892	0.745	0.053	0.800	0.868	0.660	0.040	0.782	0.852	0.695	0.085
UJSC$_{2021}$ [35]	0.842	0.907	0.771	0.047	0.809	0.891	0.684	0.035	0.800	0.853	0.728	0.073
C^{2}FNet$_{2021}$ [58]	0.838	0.898	0.762	0.049	0.813	0.886	0.686	0.036	0.796	0.864	0.719	0.080
SINetV2$_{2022}$ [13]	0.847	0.898	0.770	0.048	0.815	0.863	0.680	0.037	0.820	0.875	0.743	0.070
SegMaR$_{2022}$ [28]	0.841	0.905	0.781	0.046	0.833	0.895	0.724	0.033	0.815	0.872	0.742	0.071
ZoomNet$_{2022}$ [52]	0.853	0.907	0.784	0.043	0.838	0.893	0.729	0.029	0.820	0.883	0.752	0.066
FDNet$_{2022}$ [79]	0.834	0.895	0.750	0.052	0.837	0.897	0.731	0.030	0.844	0.903	0.778	0.062
DGNet$_{2023}$ [26]	0.857	0.907	0.784	0.042	0.822	0.877	0.693	0.033	0.839	0.901	0.769	0.057
CamoFormer-R (Ours)	0.857	0.915	0.793	0.041	0.838	0.898	0.730	0.029	0.817	0.884	0.756	0.066
CamoFormer-C (Ours)	0.884	0.936	0.833	0.033	0.860	0.923	0.767	0.024	0.860	0.920	0.811	0.051
基于*Transformer*的方法												
COS-T$_{2021}$ [63]	0.825	0.881	0.730	0.055	0.790	0.901	0.693	0.035	0.813	0.896	0.776	0.060
VST$_{2021}$ [41]	0.830	0.887	0.740	0.053	0.810	0.866	0.680	0.035	0.805	0.863	0.780	0.069
UGTR$_{2021}$ [68]	0.839	0.886	0.746	0.052	0.817	0.850	0.666	0.036	0.784	0.859	0.794	0.086
ICON$_{2022}$ [81]	0.858	0.914	0.782	0.041	0.818	0.882	0.688	0.033	0.840	0.902	0.769	0.058
TPRNet$_{2022}$ [75]	0.854	0.903	0.790	0.047	0.829	0.892	0.725	0.034	0.814	0.870	0.781	0.076
DTINet$_{2022}$ [44]	0.863	0.915	0.792	0.041	0.824	0.893	0.695	0.034	0.857	0.912	0.796	0.050
CamoFormer-S (Ours)	0.888	**0.941**	0.840	0.031	0.862	0.932	0.772	0.024	0.876	**0.935**	0.832	**0.043**
CamoFormer-P (Ours)	**0.893**	0.940	**0.850**	**0.030**	**0.872**	**0.934**	**0.793**	**0.022**	**0.878**	0.934	**0.839**	0.044

的公式定义如下：

$$FN(x,y)=\begin{cases}1, & G(x,y)=1\&P(x,y)=0\\ 0, & \text{其他情况}\end{cases},$$
$$\text{FNR}=\frac{\sum_{x=1}^{W}\sum_{y=1}^{H}FN(x,y)}{\sum_{x=1}^{W}\sum_{y=1}^{H}G(x,y)}, \tag{14}$$

其中FN 是一个像素级指标，用于确定哪些像素属于假阴性，G 是地面真值。

如图 9 所示，我们的CamoFormer 在所有数据集上都达到了最低的FNR 分数，这证明了我们模型在捕捉完整目标方面的高效性。

TABLE 2
我们的方法CamoFormer 与近期SOTA 方法在‘αBR@15’ 和‘αBR@30’ 上的比较。‘-R’：ResNet [23]，‘-C’：ConvNext [43]，‘-P’：PVTv2 [65]，‘-S’：Swin Transformer [42]，‘↑’：越高越好，‘↓’：越低越好。

方法	NC4K (4,121)				COD10K-Test (2,026)				CAMO-Test (250)			
	αBR@15		αBR@30		αBR@15		αBR@30		αBR@15		αBR@30	
	wF ↑	M↓	wF ↑	M↓	wF ↑	BR-M↓	wF ↑	M↓	wF ↑	M↓	wF ↑	M↓
基于*CNN*的方法												
PraNet$_{20}$ [16]	0.727	0.069	0.756	0.052	0.635	0.093	0.667	0.074	0.669	0.081	0.688	0.063
SINet$_{20}$ [14]	0.706	0.071	0.741	0.053	0.610	0.093	0.651	0.074	0.632	0.083	0.654	0.065
SLSR$_{21}$ [47]	0.746	0.062	0.777	0.046	0.653	0.084	0.689	0.066	0.678	0.076	0.703	0.059
MGL-R$_{21}$ [72]	0.721	0.067	0.755	0.050	0.637	0.085	0.676	0.067	0.660	0.079	0.685	0.061
PFNet$_{21}$ [51]	0.736	0.066	0.768	0.049	0.653	0.087	0.689	0.068	0.694	0.075	0.717	0.057
UJSC$_{21}$ [35]	0.755	0.061	0.785	0.045	0.670	0.081	0.705	0.063	0.712	0.070	0.737	0.053
C^{2}FNet$_{21}$ [58]	0.751	0.063	0.781	0.047	0.667	0.084	0.704	0.065	0.709	0.071	0.731	0.055
DGNet$_{23}$ [26]	0.762	0.061	0.795	0.045	0.669	0.086	0.708	0.067	0.745	0.065	0.776	0.047
SINetV2$_{22}$ [13]	0.754	0.063	0.788	0.046	0.664	0.087	0.703	0.068	0.734	0.069	0.759	0.051
SegMaR$_{22}$ [28]	0.748	0.060	0.783	0.044	0.723	0.070	0.755	0.054	0.737	0.065	0.762	0.049
ZoomNet$_{22}$ [52]	0.767	0.058	0.797	0.043	0.698	0.072	0.735	0.056	0.730	0.066	0.758	0.049
FDNet$_{22}$ [79]	0.726	0.071	0.765	0.051	0.668	0.088	0.715	0.066	0.745	0.067	0.777	0.048
CamoFormer-R (Ours)	0.768	0.057	0.800	0.042	0.696	0.071	0.740	0.055	0.725	0.066	0.755	0.048
CamoFormer-C (Ours)	0.806	0.050	0.838	0.035	0.725	0.070	0.765	0.052	0.789	0.055	0.817	0.039
基于*Transformer*的方法												
VST$_{21}$ [41]	0.754	0.061	0.786	0.046	0.677	0.080	0.714	0.062	0.714	0.071	0.736	0.054
UGTR$_{21}$ [68]	0.725	0.068	0.763	0.050	0.629	0.088	0.673	0.069	0.670	0.079	0.698	0.061
ICON$_{22}$ [81]	0.742	0.068	0.785	0.048	0.639	0.097	0.682	0.075	0.732	0.071	0.769	0.051
TPRNet$_{22}$ [75]	0.758	0.061	0.789	0.045	0.677	0.081	0.712	0.064	0.718	0.070	0.741	0.053
DTINet$_{22}$ [44]	0.776	0.058	0.808	0.042	0.674	0.084	0.714	0.065	0.776	0.060	0.806	0.043
CamoFormer-S (Ours)	0.803	0.051	0.839	0.036	0.710	0.074	0.754	0.055	0.797	0.054	0.831	0.037
CamoFormer-P (Ours)	**0.819**	**0.047**	**0.850**	**0.033**	**0.741**	**0.066**	**0.779**	**0.050**	**0.805**	**0.051**	**0.835**	**0.036**

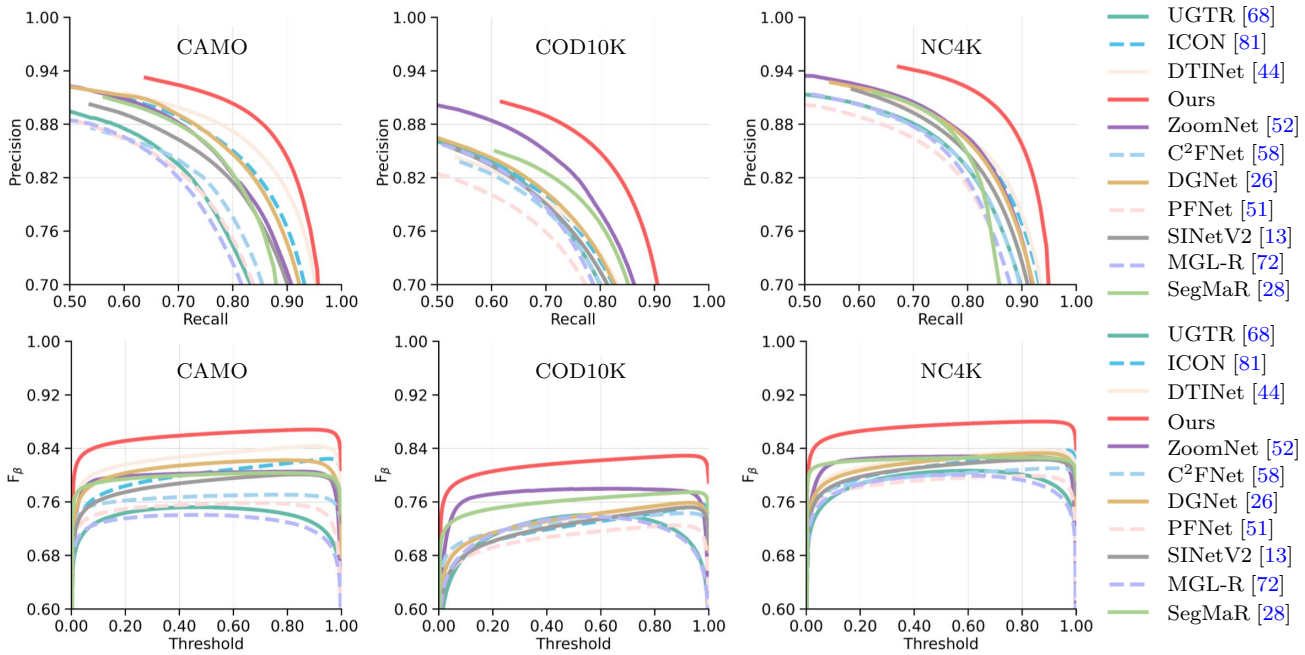


Fig. 8. 我们的CamoFormer 与最近的最先进的算法在所有基准测试上的PR 曲线和F_β 曲线比较。

TABLE 3
我们的方法CamoFormer 与其他SOTA 方法在参数数量和MACs 上的比较。该计算基于相同的代码，并在保持对应论文中的推理设置的情况下进行评估。推理时间，即每秒帧数（FPS），是在单个NVIDIA 3090 GPU 上计算的。

方法	我们的方法	DTINet [44]	ZoomNet [52]	UGTR [68]	C^2F-Net [58]	UJSC [35]	PFNet [51]	MGL-R [72]	SINet [14]
参数	71.3M	266.3M	33.4M	48.9M	28.4M	218.0M	46.5M	63.6M	48.9M
MACs	47.1G	145.6G	101.8G	500.1G	13.1G	56.2G	26.7G	277.0G	19.4G
FPS	41.3	19.7	24.0	16.6	65.8	34.2	62.6	13.4	56.5

TABLE 4
我们方法CamoFormer 各种变体的消融实验。'Baseline'：变换器主干和若干卷积层；'+MSA'：在Baseline 基础上加入MSA；'w/TA only'：仅在Baseline 中加入TA 和迭代优化方式。

设置	NC4K (4,121)				COD10K-Test (2,026)				CAMO-Test (250)			
	$S_m\uparrow$	$\alpha E\uparrow$	$wF\uparrow$	M↓	$S_m\uparrow$	$\alpha E\uparrow$	$wF\uparrow$	M↓	$S_m\uparrow$	$\alpha E\uparrow$	$wF\uparrow$	M↓
Baseline	0.859	0.916	0.801	0.043	0.830	0.904	0.719	0.032	0.838	0.891	0.780	0.058
Baseline+MSA	0.875	0.926	0.815	0.036	0.848	0.906	0.735	0.029	0.858	0.918	0.808	0.052
CamoFormer-P w/ TA only	0.871	0.920	0.808	0.039	0.844	0.908	0.741	0.029	0.859	0.910	0.810	0.055
CamoFormer-P	**0.893**	**0.940**	**0.850**	**0.030**	**0.872**	**0.934**	**0.793**	**0.022**	**0.878**	**0.934**	**0.839**	**0.044**

TABLE 5
我们提出的MSA 的消融实验。所有三个分支（'TA'、'F-TA' 和'B-TA'）都对整体性能有贡献。此外，去除'F-TA' 或'B-TA' 分支都会影响性能。

	解码器			NC4K (4,121)				COD10K-Test (2,026)				CAMO-Test (250)			
设置	TA	F-TA	B-TA	$S_m\uparrow$	$\alpha E\uparrow$	$wF\uparrow$	M↓	$S_m\uparrow$	$\alpha E\uparrow$	$wF\uparrow$	M↓	$S_m\uparrow$	$\alpha E\uparrow$	$wF\uparrow$	M↓
1				0.865	0.921	0.795	0.041	0.836	0.915	0.724	0.030	0.857	0.915	0.797	0.053
2	✓			0.871	0.920	0.808	0.039	0.844	0.908	0.741	0.029	0.859	0.910	0.810	0.055
3		✓		0.881	0.927	0.824	0.034	0.855	0.918	0.747	0.026	0.868	0.920	0.821	0.053
4			✓	0.879	0.924	0.819	0.035	0.851	0.915	0.742	0.028	0.856	0.914	0.809	0.055
5		✓	✓	0.887	0.937	0.844	0.032	0.869	0.930	0.783	0.024	0.875	0.933	0.838	0.045
6	✓		✓	0.885	0.933	0.839	0.033	0.860	0.924	0.763	0.025	0.866	0.926	0.827	0.048
7	✓	✓		0.889	**0.940**	0.845	0.031	0.866	0.927	0.773	0.024	0.871	0.929	0.832	0.046
8	✓	✓	✓	**0.893**	**0.940**	**0.850**	**0.030**	**0.872**	**0.934**	**0.793**	**0.022**	**0.878**	**0.934**	**0.839**	**0.044**

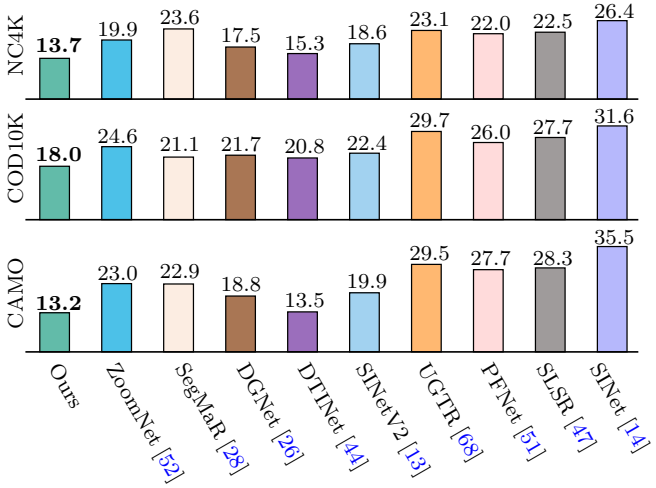


Fig. 9. 10 种方法在三个不同数据集上的FNR 统计结果。最佳结果以红色高亮显示。

5.4 方法分析

整体结果。 我们首先对CamoFormer 的网络架构进行消融实验。结果如Tab. 4 所示。'Baseline' 指的是仅使用编码器并通过卷积进行预测的模型。当应用我们的MSA 时，性能在所有评估指标上相比'Baseline' 明显提升。接着，我们尝试仅添加含TA 的解码器。这种逐步融合策略相较于'Baseline' 也有所帮助。最后，我们在解码器中加入我们的MSA，如 Fig. 2 所示。可以看到，性能进一步提升。上下两部分都表明MSA 对于COD 的重要性。

掩码可分离注意力。 接着，我们对MSA 中每个组件的作用进行消融实验。Tab. 5 展示了实验结果。第一行中的'Baseline' 可以看作是一个简单的特征金字塔网络，即在 Fig. 2 中没有加入MSA。我们可以观察到，每种注意力组件都有助于提升性能。尽管TA 相较于F-TA 和B-TA 提升了更多性能，但将F-TA 或B-TA 与TA 结合使用可以进一步提升结果。特别是，加入所有三个组件在所有三个数据集上都获得了最佳结果。这一系列实验表明，使用提出的MSA 分别处理前景和背景对于分割伪装物体是有益的。

特征可视化。 为了更深入地了解我们的MSA，我们还对其周围的特征进行可视化。注意，我们选择了第2阶段和第5阶段的MSA 进行可视化。如 Fig. 10 所示，F-TA 和B-TA 能够

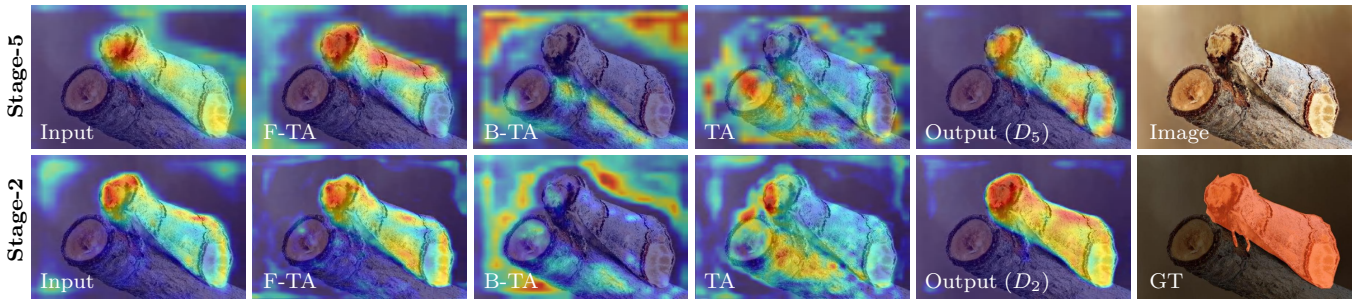


Fig. 10. MSA 周围特征图的可视化。选择了第2 阶段和第5 阶段的特征进行比较。

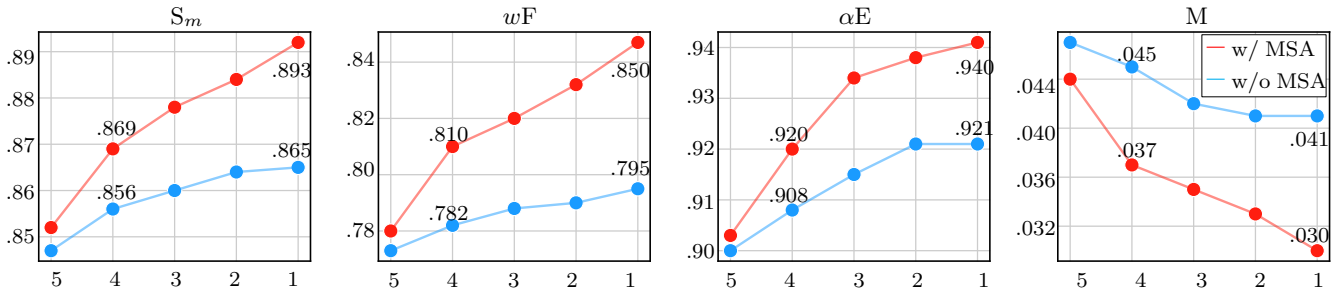


Fig. 11. 在NC4K 数据集上，我们逐步融合策略在不同特征层次的表现。

有效地获取前景和背景的线索。F-TA、B-TA 和TA 的特征是互补的，能够形成完整的伪装物体。因此，伪装物体的精确位置和详细信息能够轻松捕获。此外，与第5阶段的特征相比，我们可以根据第2阶段的特征更清晰地区分伪装目标。

Fig. 12. SINet [14] 和ZoomNet [52] 配备我们提出的MSA 的可视化结果。

MSA在渐进融合中的作用。 在我们的模型中，构建了一条自上而下的路径以渐进地优化解码器中的分割图。为了展示我们MSA在该策略中的影响，我们展示了在有无MSA设置下，模型在不同特征层次上的性能曲线。结果见于 Fig. 11。我们可以看到，从特征层级5到特征层级1，模型在没有MSA（蓝线）和有MSA的情况下，所有四项评估指标的性能差距逐渐增大。这表明，所提出的MSA与渐进融合解码器兼容。

MSA的泛化性。 为了说明我们的MSA模块的通用性，我们将其应用于其他一些COD方法，如SINet [14]和ZoomNet [52]。具体来说，对于其他方法，我们使用3×3卷积核生成中间预测，并将我们的MSA模块集成到其解码器的每个单元中。如 Tab. 6所示，MSA在所有伪装物体检测基准上为其他方法带来了持续且显著的改进。我们还展示了 Fig. 12中的可视化结果。这些结果证明了我们方法的泛化能力。

解码器宽度。 解码器的宽度（通道数）不仅影响模型的大小，还影响推理速度。Tab. 7展示了在通道数变化时，模型参数、计算成本和性能的变化。我们可以看到，当C_d从32增加到128时，模型性能有了明显提升。因此，C_d设置为128，以在效率和模型性能之间进行权衡。

6 结论

我们提出了CamoFormer用于伪装物体分割。CamoFormer的核心是掩蔽可分离注意力（MSA），它使用不同的注意力头分别处理前景和背景区域。为了更好地利用我们的MSA，我们采用了渐进式精细化解码器，以自上而下的方式逐渐提高不同特征层次上的分割质量。大量实验表明，CamoFormer在18个现有的最先进模型中取得了显著的性能提升。

References

[1] Yevgeny Beiderman, Mina Teicher, Javier Garcia, Vicente Mico, and Zeev Zalevsky. Optical technique for classification, recognition and identification of obscured objects. *Opt. Commun.*, 283(21):4274–4282, 2010.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Eur. Conf. Comput. Vis.*, 2020.

[3] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

TABLE 6
MSA 模块的泛化能力。我们将MSA 模块应用于SINet [14] 和ZoomNet [52] 并报告结果。

模型	计算量		NC4K (4,121)				COD10K-Test (2,026)				CAMO (250)			
	参数	MAC	$S_m \uparrow$	$\alpha E \uparrow$	$wF \uparrow$	$M \downarrow$	$S_m \uparrow$	$\alpha E \uparrow$	$wF \uparrow$	$M \downarrow$	$S_m \uparrow$	$\alpha E \uparrow$	$wF \uparrow$	$M \downarrow$
SINet [14]	49M	19G	0.808	0.883	0.723	0.058	0.776	0.867	0.631	0.043	0.745	0.825	0.644	0.092
+MSA	57M	25G	0.829	0.900	0.752	0.052	0.809	0.883	0.675	0.035	0.790	0.844	0.681	0.077
ZoomNet [52]	33M	102G	0.853	0.907	0.784	0.043	0.838	0.893	0.729	0.029	0.820	0.883	0.752	0.066
+MSA	39M	111G	0.870	0.920	0.800	0.038	0.850	0.904	0.745	0.026	0.835	0.899	0.770	0.061

TABLE 7
解码器中通道数的消融研究。所有变体均配备了我们的MSA。

设置	计算量		NC4K (4,121)				COD10K-Test (2,026)				CAMO (250)			
	参数	MAC	$S_m \uparrow$	αE	$wF \uparrow$	M	$S_m \uparrow$	αE	$wF \uparrow$	M	$S_m \uparrow$	αE	$wF \uparrow$	M
$C_d = 32$	63M	30G	0.890	0.939	0.844	0.031	0.867	0.929	0.782	0.024	0.873	0.928	0.831	0.046
$C_d = 64$	65M	34G	0.890	0.940	0.846	0.031	0.869	0.929	0.787	0.023	0.876	0.931	0.833	0.046
$C_d = 128$	71M	47G	**0.893**	0.940	0.850	**0.030**	0.872	**0.934**	**0.793**	**0.022**	0.878	**0.934**	**0.839**	**0.044**
$C_d = 192$	82M	69G	**0.893**	**0.942**	**0.851**	**0.030**	**0.873**	**0.934**	0.790	0.023	**0.879**	**0.934**	0.838	**0.044**
$C_d = 256$	97M	99G	0.892	**0.942**	0.846	**0.030**	0.870	0.932	0.790	0.023	0.876	0.931	0.837	0.045

[4] Gong Cheng, Jiabao Wang, Ke Li, Xingxing Xie, Chunbo Lang, Yanqing Yao, and Junwei Han. Anchor-free oriented proposal generator for object detection. *IEEE Trans. Geosci. Remote Sens.*, 60:1–11, 2022.

[5] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(3):569–582, 2014.

[6] Shupeng Cheng, Ge-Peng Ji, Pengda Qin, Deng-Ping Fan, Bowen Zhou, and Peng Xu. Large model based referring camouflaged object detection. *arXiv preprint arXiv:2311.17122*, 2023.

[7] Xuelian Cheng, Huan Xiong, Deng-Ping Fan, Yiran Zhong, Mehrtash Harandi, Tom Drummond, and Zongyuan Ge. Implicit motion handling for video camouflaged object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

[8] Hung-Kuo Chu, Wei-Hsin Hsu, Niloy J Mitra, Daniel Cohen-Or, Tien-Tsin Wong, and Tong-Yee Lee. Camouflage images. *ACM Trans. Graph.*, 29(4):51–1, 2010.

[9] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Ann. Oper. Res.*, 134(1):19–67, 2005.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Int. Conf. Learn. Represent.*, pages 1–12, 2021.

[11] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Int. Conf. Comput. Vis.*, 2017.

[12] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *Int. Joint Conf. on Artif. Intel.*, 2018.

[13] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10):6024–6042, 2022.

[14] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[15] Deng-Ping Fan, Ge-Peng Ji, Peng Xu, Ming-Ming Cheng, Christos Sakaridis, and Luc Van Gool. Advances in deep concealed scene understanding. *Visual Intelligence*, 1(1):16, 2023.

[16] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *Inter. Conf. on Med. Image Comp. and Computer-Assisted Interv.*, 2020.

[17] Meirav Galun, Eitan Sharon, Ronen Basri, and Achi Brandt. Texture segmentation by multiscale aggregation of filter responses and shape elements. In *Int. Conf. Comput. Vis.*, 2003.

[18] Shiming Ge, Xin Jin, Qiting Ye, Zhao Luo, and Qiang Li. Image editing by object-aware optimal boundary searching and mixed-domain composition. *Comp. Visual Media*, 4(1):71–82, 2018.

[19] Yuchao Gu, Lijuan Wang, Ziqin Wang, Yun Liu, Ming-Ming Cheng, and Shao-Ping Lu. Pyramid constrained self-attention network for fast video salient object detection. In *AAAI Conf. on Artif. Intel.*, 2020.

[20] Hongxing Guo, Yaling Dou, Ting Tian, Jingli Zhou, and Shengsheng Yu. A robust foreground segmentation method by temporal averaging multiple video frames. In *IEEE International conference on audio, language and image processing*, 2008.

[21] Joanna R Hall, Innes C Cuthill, Roland Baddeley, Adam J Shohet, and Nicholas E Scott-Samuel. Camouflage, detection and identification of moving targets. *Proc. Royal Soc. B*, 280(1758):20130064, 2013.

[22] Junwei Han, Dingwen Zhang, Gong Cheng, Nian Liu, and Dong Xu. Advanced deep-learning techniques for salient and category-specific object detection: a survey. *IEEE Signal Process. Mag.*, 35(1):84–100, 2018.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

[24] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip Torr. Deeply supervised salient object detection with short connections. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(4):815–828, 2019.

[25] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[26] Ge-Peng Ji, Deng-Ping Fan, Yu-Cheng Chou, Dengxin Dai, Alexander Liniger, and Luc Van Gool. Deep gradient learning for efficient camouflaged object detection. *Mach. Intell. Research*, 20(1):92–108, 2023.

[27] Ge-Peng Ji, Lei Zhu, Mingchen Zhuge, and Keren Fu. Fast camouflaged object detection via edge-based reversible re-calibration network. *Pattern Recognition*, 123:108414, 2022.

[28] Qi Jia, Shuilian Yao, Yu Liu, Xin Fan, Risheng Liu, and Zhongxuan Luo. Segment, magnify and reiterate: Detecting camouflaged objects the hard way. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

[29] Xinhao Jiang, Wei Cai, Zhili Zhang, Bo Jiang, Zhiyong Yang, and Xin Wang. Magnet: A camouflaged object detection network simulating the observation effect of a magnifier. *Entropy*, 24(12):1804, 2022.

[30] Zi-Hang Jiang, Qibin Hou, Li Yuan, Daquan Zhou, Yujun Shi, Xiaojie Jin, Anran Wang, and Jiashi Feng. All tokens matter: Token labeling for training better vision transformers. In *Adv. Neural Inform. Process. Syst.*, 2021.

[31] Ch Kavitha, B Prabhakara Rao, and A Govardhan. An efficient content based image retrieval using color and texture of image sub blocks. *Inter. Journal of Eng. Sci. and Tech.*, 3(2):1060–1068, 2011.

[32] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, 2017.

[33] Chunbo Lang, Gong Cheng, Binfei Tu, Chao Li, and Junwei Han. Base and meta: A new perspective on few-shot segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023.

[34] Trung-Nghia Le, Tam V Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *Comp. Vis. and Image Understanding*, 184:45–56, 2019.

[35] Aixuan Li, Jing Zhang, Yunqiu Lv, Bowen Liu, Tong Zhang, and Yuchao Dai. Uncertainty-aware joint salient object and camouflaged object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

[36] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

[37] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[38] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning for generic object detection: A survey. *Int. J. Comput. Vis.*, 128(2):261–318, 2020.

[39] Nian Liu and Junwei Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

[40] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[41] Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer. In *Int. Conf. Comput. Vis.*, 2021.

[42] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Int. Conf. Comput. Vis.*, 2021.

[43] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

[44] Zhengyi Liu, Zhili Zhang, and Wei Wu. Boosting camouflaged object detection with dual-task interactive transformer. *Int. Conf. Pattern Recog.*, 2022.

[45] Xue-Jing Luo, Shuo Wang, Zongwei Wu, Christos Sakaridis, Yun Cheng, Deng-Ping Fan, and Luc Van Gool. Camdiff: camouflage image augmentation via diffusion. *CAAI Artificial Intelligence Research*, 2, 2023.

[46] Ziyang Luo, Nian Liu, Wangbo Zhao, Xuguang Yang, Dingwen Zhang, Deng-Ping Fan, Fahad Khan, and Junwei Han. Vscode: General visual salient and camouflaged object detection with 2d prompt learning. *arXiv preprint arXiv:2311.15011*, 2023.

[47] Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. Simultaneously localize, segment and rank the camouflaged objects. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

[48] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014.

[49] Gellért Máttyus, Wenjie Luo, and Raquel Urtasun. Deeproadmapper: Extracting road topology from aerial images. In *Int. Conf. Comput. Vis.*, 2017.

[50] Gerard Medioni. Generic object recognition by inference of 3-d volumetric. *Object Categorization: Computer and Human Vision Perspectives*, 87, 2009.

[51] Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

[52] Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

[53] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Adv. Neural Inform. Process. Syst.*, 2019.

[54] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2012.

[55] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[56] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *Int. Conf. Learn. Represent.*, pages 1–14, 2015.

[57] Martin Stevens and Sami Merilaita. Animal camouflage: current issues and new perspectives. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, 364(1516):423–427, 2009.

[58] Yujia Sun, Geng Chen, Tao Zhou, Yi Zhang, and Nian Liu. Context-aware cross-level fusion network for camouflaged object detection. In *Int. Joint Conf. on Artif. Intel.*, pages 1025–1031, 2021.

[59] Satoshi Suzuki et al. Topological structural analysis of digitized binary images by border following. *Computer vision, graphics, and image processing*, 30(1):32–46, 1985.

[60] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

[61] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In

Inter. Conf. Mach. Learning, 2021.

[62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Adv. Neural Inform. Process. Syst.*, pages 5998–6008, 2017.

[63] Haiwen Wang, Xinzhou Wang, Fuchun Sun, and Yixu Song. Camouflaged object segmentation with transformer. In *Cog. Sys. and Inform. Process.*, 2021.

[64] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Int. Conf. Comput. Vis.*, 2021.

[65] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Comp. Visual Media Journal*, 8(3):415–424, 2022.

[66] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Adv. Neural Inform. Process. Syst.*, 2021.

[67] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Int. Conf. Comput. Vis.*, 2015.

[68] Fan Yang, Qiang Zhai, Xin Li, Rui Huang, Ao Luo, Hong Cheng, and Deng-Ping Fan. Uncertainty-guided transformer reasoning for camouflaged object detection. In *Int. Conf. Comput. Vis.*, 2021.

[69] Bowen Yin, Xuying Zhang, Zhong-Yu Li, Li Liu, Ming-Ming Cheng, and Qibin Hou. Dformer: Rethinking rgbd representation learning for semantic segmentation. In *Proc. Int. Conf. Learn. Represent.*, pages 1–14, 2024.

[70] Li Yuan, Qibin Hou, Zihang Jiang, Jiashi Feng, and Shuicheng Yan. Volo: Vision outlooker for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022. doi:https://doi.org/10.1109/TPAMI.2022.3206108.

[71] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

[72] Qiang Zhai, Xin Li, Fan Yang, Chenglizhao Chen, Hong Cheng, and Deng-Ping Fan. Mutual graph learning for camouflaged object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

[73] Jing Zhang, Jianwen Xie, Nick Barnes, and Ping Li. Learning generative vision transformer with energy-based latent space for saliency prediction. In *Adv. Neural Inform. Process. Syst.*, 2021.

[74] Miao Zhang, Shuang Xu, Yongri Piao, Dongxiang Shi, Shusen Lin, and Huchuan Lu. Preynet: Preying on camouflaged objects. In *ACM Int. Conf. Multimedia*, 2022.

[75] Qiao Zhang, Yanliang Ge, Cong Zhang, and Hongbo Bi. Tprnet: camouflaged object detection via transformer-induced progressive refinement network. *The Visual Computer*, 39(10):4593–4607, 2023.

[76] Xuying Zhang, Bowen Yin, Zheng Lin, Qibin Hou, Deng-Ping Fan, and Ming-Ming Cheng. Referring camouflaged object detection. *arXiv preprint arXiv:2306.07532*, 2023.

[77] Xiang Zhang, Ce Zhu, Shuai Wang, Yipeng Liu, and Mao Ye. A bayesian approach to camouflaged moving object detection. *IEEE Trans. Circuit Syst. Video Technol.*, 27(9):2001–2013, 2016.

[78] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

[79] Yijie Zhong, Bo Li, Lv Tang, Senyun Kuang, Shuang Wu, and Shouhong Ding. Detecting camouflaged object in frequency domain. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

[80] Yupeng Zhou, Zhen Li, Chun-Le Guo, Song Bai, Ming-Ming Cheng, and Qibin Hou. Srformer: Permuted self-attention for single image super-resolution. In *Int. Conf. Comput. Vis.*, 2023.

[81] Mingchen Zhuge, Deng-Ping Fan, Nian Liu, Dingwen Zhang, Dong Xu, and Ling Shao. Salient object detection via integrity learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022. doi: https://doi.org/10.1109/TPAMI.2022.3179526.

[82] Mingchen Zhuge, Xiankai Lu, Yiyou Guo, Zhihua Cai, and Shuhan Chen. Cubenet: X-shape connection for camouflaged object detection. *Pattern Recognition*, 127:108644, 2022.

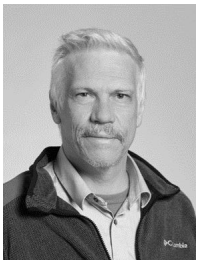
Bowen Yin is a Ph.D. student from the college of computer science, Nankai university. He is supervised by Prof. Qibin Hou. His research interests include computer vision and multimodal scene perception.

Xuying Zhang is a Ph.D. student from the college of computer science, Nankai university. He is supervised by Prof. Ming-Ming Cheng. His research interests include multimodal learning, camouflaged scene understanding, and 2D/3D visual perception.

Deng-Ping Fan (Senior Member, IEEE) is a Full Professor and deputy director of the Media Computing Lab (MC Lab) at the College of Computer Science, Nankai University, China. Before that, he was postdoctoral, working with Prof. Luc Van Gool in Computer Vision Lab @ ETH Zurich. He is one of the core technique members in TRACE-Zurich project on automated driving.

Shaohui Jiao received her PhD degree from Chinese Academy of Sciences in 2010. She is now a researcher in MultiMedia Lab, Bytedance Inc. Her research interests include computer graphics, computer vision, VR, and AIGC.

Ming-Ming Cheng (Senior Member, IEEE) received his PhD degree from Tsinghua University in 2012, and then worked with Prof. Philip Torr in Oxford for 2 years. Since 2016, he is a full professor at Nankai University, leading the Media Computing Lab. His research interests include computer vision and computer graphics. He received awards, including ACM China Rising Star Award, IBM Global SUR Award, *etc*. He is a senior member of the IEEE and on the editorial boards of IEEE TPAMI and IEEE TIP.



Luc Van Gool is a professor at ETH Zurich and the head of the Computer Vision Laboratory (CV Lab). His main research interests include 2D and 3D object recognition, texture analysis, distance acquisition, stereo vision, robot vision, and optical flow. He has served as a member of the procedural committee for multiple top international conferences, including ICCV, ECCV, and CVPR. Received the David Marr Prize in 1998.

Qibin Hou (Member, IEEE) received his Ph.D. degree from the School of Computer Science, Nankai University. Then, he spent two wonderful years working at the National University of Singapore as a research fellow. Now, he is an associate professor at School of Computer Science, Nankai University. He has published more than 40 papers on top conferences/journals, including IEEE TPAMI, CVPR, ICCV, NeurIPS, etc. His research interests include deep learning and computer vision.