指向性伪装目标检测

张旭迎, 尹博文, 林铮, 侯淇彬, 范登平, *Senior Member, IEEE*, 程明明, *Senior Member, IEEE*,

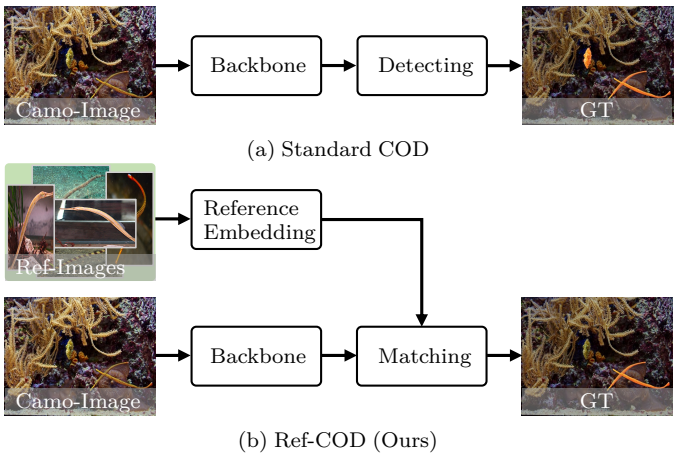
Abstract—我们关注指向性伪装目标检测这一新颖任务，旨在通过少量包含显著目标的参考图像，对指定的伪装目标进行分割。为推动该任务的研究，我们构建了一个大规模数据集R2C7K，包含7,000张来自真实场景的图像，涵盖64个目标类别。在此基础上，我们提出了一个简单而高效的双分支框架，称为R2CNet。该框架由两个分支组成：参考分支用于从参考图像中提取目标对象的通用表征，分割分支在该表征的引导下对伪装目标进行识别与分割。具体而言，我们设计了参考掩码生成模块，用于生成像素级的先验掩码；以及参考特征增强模块，以提升模型对指定伪装目标的识别能力。大量实验结果表明，相较于传统伪装目标检测方法，我们的方法在分割指定伪装目标及识别其主体结构方面表现出更优性能。代码和数据集已公开发布：https://github.com/zhangxuying1004/RefCOD。

Index Terms—伪装目标检测；通用表征；R2C7K数据集；R2CNet框架

1 引言

伪装目标检测（Camouflaged Object Detection, COD）旨在分割那些在视觉上与周围环境融为一体、难以察觉的目标，近年来已引起越来越多的关注 [11], [26], [57], [72]。该研究方向在诸多现实应用中发挥着重要作用，例如医学图像分割 [14], [16]、表面缺陷检测 [35], [73]、害虫识别 [77]等。值得注意的是，在真实场景中可能存在多个伪装目标，但在多数实际应用中，我们通常仅需检测特定的目标对象。以野外探险为例，探索者往往只关注某些特定物种，而这些目标可能混杂于大量外观相似的其他对象之中。若能借助相关参考信息，目标的定位过程将更具方向性与效率。因此，引入参考信息以辅助伪装目标检测，成为一条值得深入探索的新路径。本文将此类任务定义为指向性伪装目标检测。

指向性伪装目标检测借助参考信息引导指定伪装目标的检测过程，与人类在识别伪装物体时的视觉机制相契合 [76]。核心问题在于，采用哪种形式的参考信息最为有效。近年来已有相关研究在图像分割任务中探索多种参考形式，例如使用文本作为引导的表达式分割 [24]以及基于图像的少样本分割 [90]。然而，这些方法通常依赖于精细标注的伪装目标图像或详细的文本描述，获取成本高昂，难以直接迁移至伪装目标检测任务中。考虑到网络上易于获取包含显著目标的图像，一个自然的问题是：能否利用这些显著目标图像，辅助识别特定的伪装目标?



(a) Standard COD

(b) Ref-COD (Ours)

图 1. 标准伪装目标检测与指向性伪装目标检测的视觉比较：对于一幅包含多个伪装目标的图像，传统的伪装目标检测模型往往会无差别地检测出所有与背景融合的伪装目标；而指向性伪装目标检测模型则在给定一组参考图像的前提下，尝试有针对性地识别出指定的伪装目标。

为回答该问题，本文提出了一个新的指向性伪装目标检测任务基准。我们希望借助显著目标检测（Salient Object Detection, SOD）的研究进展，从参考图像中提取目标对象的通用表征，进而引导对指定伪装目标的分割。图1 展示了传统伪装目标检测与指向性伪装目标检测任务之间的差异。指向性伪装目标检测将伪装目标检测从“盲目识别伪装目标”的过程转变为“目标驱动的匹配检测”，使伪装目标的检测更具明确性与实用性。为支持对该新任务的深入研究，我们构建了一个大规模数据集R2C7K，涵盖丰富的真实场景且无版权争议。其基本信息如下：1）包含7000 张图像，涵盖64 个类别；2）包含两个子集：Camo子集包含伪装目标图像，Ref子集包含显著目标图像；3）Ref子集中每类图像数量固定，Camo子集的每类图像数量则不作限制。

在方法设计上，我们提出了一个双分支网络框架，命名为R2CNet，包括参考分支与分割分支。参考分支用于从显

- 本研究得到了国家自然科学基金（*NSFC*，编号分别为*62225604*和*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。
- 张旭迎、尹博文、侯淇彬、范登平和程明明隶属于南开大学计算机学院视觉计算与智能感知实验室，天津，中国。侯淇彬、范登平和程明明亦隶属于南开国际先进研究院（深圳福田）。
- 林铮隶属于清华大学计算机科学与技术系，北京，中国。
- 侯淇彬和林铮为通讯作者。
- 前两位作者贡献相同。

Manuscript received January 12, 2025; revised August 10, 2024.

著目标图像中提取目标对象的通用表征，作为检测伪装目标的引导信息。我们进一步设计了一个参考掩码生成模块，通过将通用表征与分割分支的视觉特征逐像素比对，生成像素级的参考先验掩码。然而，即使属于相同类别，显著目标与伪装目标在外观上可能存在较大差异，增加了识别的难度。为缓解该问题，我们引入了双源信息融合策略，以减少两个信息源之间的表征差异。此外，我们还设计了一个参考特征增强模块，在参考掩码的引导下，实现多尺度特征的交互与融合，进一步突出目标区域。

我们在R2C7K数据集上进行了广泛实验。具体而言，我们以基于特征金字塔网络 [40]的多尺度融合模型作为伪装目标检测基线，对比评估了R2CNet的性能，并采用伪装目标检测常用指标 [9], [10], [55], [60]进行量化分析。实验结果显示，R2CNet在各项指标上均显著优于基线模型。进一步地，我们将指向性伪装目标检测架构扩展至当前7 种最先进的伪装目标检测方法，改进后的方法在不引入复杂设计的前提下，普遍超越原始伪装目标检测版本。此外，可视化结果也表明指向性伪装目标检测方法（如R2CNet）在指定目标分割与目标主体识别方面具有更优表现。

综上，本文的主要贡献如下：

- 提出了一项新任务基准指向性伪装目标检测，首次尝试结合显著目标检测与伪装目标检测，通过显著目标图像辅助检测指定伪装目标；
- 构建了一个大规模数据集R2C7K，为指向性伪装目标检测研究提供了基础数据资源与深入探索的可能；
- 提出了一个新颖的指向性伪装目标检测框架，实验证明其为该新任务提供了有效且通用的解决方案。

2 相关工作

在本节中，我们首先回顾伪装目标检测任务的研究进展与现存问题。随后，介绍显著目标检测领域的发展历程。最后，我们综述使用不同形式参考信息的目标分割相关研究。

2.1 伪装目标检测

从复杂背景中识别伪装目标是一项极具挑战性的任务，其主要难点在于目标与周围环境的高度相似性、尺度的多样性以及外观的模糊性等。为解决这一问题，近年来大量相关研究不断涌现。其中，较早的一项系统性研究由文献 [12]提出，发布了高质量标注的大规模数据集COD10K，并构建了一个精心设计的“搜索-识别”框架，为该研究领域奠定了基础。

随后，多种策略相继被提出以提升伪装目标检测模型的性能，包括多尺度特征融合策略 [2], [8], [57], [71], [104]、多阶段细化方法 [11], [28], [79], [93]、图学习 [88]、弱监督方法 [21]、不确定性建模 [30], [36], [43]、前景与背景分离 [56], [87], [89]以及注意力机制 [7], [91]等。更多相关工作可参考最新的综述论文 [13]。

近年来，也有研究尝试引入额外信息以进一步提高分割精度，如边界 [27], [62], [72], [99], [101]、纹理 [26], [64], [102]、频域特征 [39], [97]以及深度信息 [84], [85], [92]等。然而，从现有伪装图像中获取高质量的此类辅助信息往往成本高昂，且这些信息难以直接明确地告诉模型应当分割哪些目标。

为此，本文提出了一个新颖的研究基准，旨在从易获取的参考图像中提取目标对象的通用表征，作为显式的语义引导，用于准确定位和分割指定的伪装目标。该方法与传统伪装目标检测任务在任务定义与研究思路上均存在显著差异。

2.2 显著目标检测

显著目标检测的目标是在给定图像中识别出最具吸引力的显著物体。该研究方向的发展大致经历了两个阶段：1）传统方法阶段；2）深度学习方法阶段。

在早期阶段，主要依赖于图像块、候选目标及超像素所提取的手工特征 [5], [22], [31], [80]。尽管这些方法在部分场景中表现良好，但其特征提取过程通常较为耗时，且在复杂环境下性能容易大幅下降。

随着全卷积网络 [48]和Transformer架构 [54], [78], [83], [94], [103]的兴起，显著目标检测研究逐步进入深度学习阶段。在这一阶段，U 型结构网络 [40], [65]、多阶段监督机制 [17], [23], [96]和注意力机制 [4], [44]被广泛应用于显著目标检测方法中，从而显著提升了像素级预测的准确性。

值得一提的是，显著目标检测在多个研究领域中具有广泛的应用，例如视觉跟踪 [1]、无监督分割 [38]、图像压缩 [19]以及内容感知图像编辑 [6]。这一应用广度主要得益于显著目标检测能有效发现能够代表场景语义的关键目标或区域。显著目标检测的这一特性也启发了本文提出的“参考信息引导”这一研究思路。

2.3 指向性目标检测

指向性目标分割是指在某种参考信息（如图像或文本）的引导下，从给定图像中分割出特定的视觉目标。

小样本分割（Few-shot Segmentation, FSS）研究的是利用包含相同类别目标的标注图像作为参考，引导目标分割任务。在训练阶段，模型在大量具有基础类别像素标注的图像（查询集）上进行训练；在测试阶段，仅需少量标注样本（支持集）即可对未见类别进行像素级预测。大多数现有的小样本分割网络通常采用双分支结构，即支持分支与查询分支，分别提取支持图像和查询图像的特征，并实现两者之间的信息交互。小样本分割的开创性工作由 [66]提出，其支持分支直接预测查询分支中用于分割的最后一层的权重。随后， [95] 引入了掩码平均池化操作以提取具有代表性的支持特征，这一设计也被众多后续工作广泛采用。近年来，大量研究工作 [33], [75], [90]在冻结的主干网络上构建强大的模块，以提升模型对未见类别的适应能力。

指向性表达分割（Referring Expression Segmentation, RES）则是利用文本表达对图像中的目标进行分割。指向性表达分割的核心目标是根据描述性的文本表达来分割视觉目标，网络结构同样普遍采用双分支架构。该领域的开创性研究由 [24]提出，方法中分别使用视觉编码器和语言编

图 2. R2C7K 数据集中的示例图像。Camo子集中的伪装目标使用橙色掩码标注显示其标注信息。

码器提取视觉特征和文本特征，然后通过特征拼接生成分割掩码。随后，研究者相继引入多层视觉特征融合 [37]、多模态LSTM [41]、注意力机制 [67], [100]以及协同网络 [51]等方法，不断提升分割精度。此外，[70] 提出利用文本描述作为图像内容丰富性的参考信息以获得更准确的视觉关注预测。

本文提出的指向性伪装目标检测也属于指向性目标分割任务的一种。但与上述已有方法不同的是，本文的参考信息获取过程更加便捷：我们无需额外收集包含稀有伪装目标的难标注图像，也无需为现有伪装目标检测数据集编写细致的文本描述。因此，本文提出的指向性伪装目标检测在学术与工业落地方面具有更高的可行性与实用价值。

3 数据集

一系列数据集的出现为人工智能研究，尤其是在当前深度学习高度依赖数据的时代，奠定了基础。此外，如文献 [61], [81] 所指出的，一个数据集的质量在其作为基准使用的生命周期中起着重要作用。基于以上考虑，我们构建了一个大规模数据集，命名为R2C7K，用于本文提出的指向性伪装目标检测任务。在本节中，我们将分别介绍该数据集的构建过程和统计信息。

3.1 数据收集与标注

为了构建R2C7K 数据集，第一步是确定要检测哪些伪装目标。为此，我们调查了伪装目标检测研究中最常用的几个数据集，即COD10K [11]、CAMO [34] 和NC4K [53]。鉴于COD10K 是目前最大且注释最全面的伪装目标检测数据集，我们主要基于COD10K 构建了Camo子集作为R2C7K 的一部分。具体来说，我们剔除了一些不常见的类别，例如寄居蟹、鲽鱼等，最终获得了覆盖64 个类别的4,966 张伪装图像。对于仅包含一个伪装目标的图像，我们直接采用COD10K 提供的标注；而对于包含多个伪装目标的图像，我们仅保留与参考类别对应的标注像素，其余部分予以清除。值得注意的是，对于一些样本数量极少的类别，我们还从NC4K 中补充了49 个样本。

接下来，我们根据选定的64 个类别构建了Ref子集。我们将这些类别名称作为关键词，在互联网上搜索每类25 张包含所需显著目标的真实场景图像，作为参考图像。特别地，这些无版权争议的参考图像主要来自Flickr 和Unsplash。关于图像收集方案的更多细节，推荐读者参考文献 [98]。

最后，我们在图 2 中展示了R2C7K 数据集的图像与标注示例。

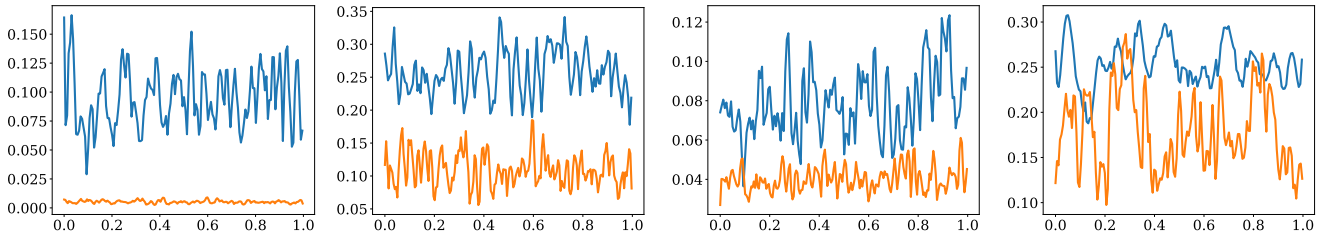
3.2 数据统计

子集对比. 图 3 展示了Ref子集与Camo子集之间在四个属性上的比较。具体而言，目标面积指的是图像中目标的实际大小，目标占比表示目标在图像中所占的比例，目标距离是目标中心点到图像中心的距离，而全局对比度则是衡量目标检测难度的一个指标。从图中可以观察到，Ref子集中的目标相较于Camo子集中的目标更大，同时Ref中的图像包含了更多对比度信息。因此，Ref中的目标更容易被检测出来，这种类型的参考信息非常适合用于指向性伪装目标检测 研究任务。

类别与数量. R2C7K 数据集共包含6,615 个样本，涵盖64 个类别。其中，Camo子集包含5,015 个样本，Ref子集包含1,600 个样本。值得注意的是，该数据集中每个类别都固定包含25 张参考图像，而伪装目标检测图像的数量则在不同类别之间分布不均，如图图 4 所示。

分辨率分布. 图 5a 和图 5b 分别展示了Camo和Ref子集中图像的分辨率分布情况。可以看出，这两个子集中都包含大量的高清图像，这有助于提供更清晰的目标边界与纹理信息。

数据划分. 为便于开展指向性伪装目标检测 研究任务的模型开发，我们为R2C7K 数据集提供了标准的参考式划分策略。对于Ref子集，我们从每个类别中随机选择20 张图像作为训练集，其余5 张图像作为测试集；而在Camo子集中，来源于COD10K 训练集的样本被用作训练集，来源于测试集的样本用于测试；至于来自NC4K 的样本，我们将其随机分配到训练集和测试集中，以确保每个类别在两个划分中至少拥有6个样本。



(a) 目标面积 (b) 目标占比 (c) 目标距离 (d) 全局对比度

图 3. Camo子集与Ref子集在四个属性上的对比，即目标面积、目标占比、目标距离和全局对比度。前者的结果以橙色显示，后者的结果以蓝色显示。

表 1
我们在多个维度上对R2C7K 与以往的伪装目标检测数据集进行了对比。Cate.: 类别数量; Camo-Img.: 包含伪装目标的图像数量; Ref-Img.: 作为参照的图像数量; Loc.: 定位; Det.: 检测; Cls.: 分类; WS.: 弱监督; RefSeg.: 指向性目标分割.

Datasets	Statistics				Tasks				
	Year	Cate.	Camo-Img.	Ref-Img.	Loc.	Det.	Cls.	WS.	RefSeg.
CHAMELEON [69]	2018	N/A	76	N/A	✓	✓	✗	✗	✗
CAMO [34]	2019	N/A	2500	N/A	✓	✓	✗	✗	✗
COD10K [11]	2020	68	5066	N/A	✓	✓	✓	✓	✗
NC4K [53]	2021	N/A	4121	N/A	✓	✓	✗	✗	✗
R2C7K (Ours)	2023	64	5015	1600	✓	✓	✓	✓	✓

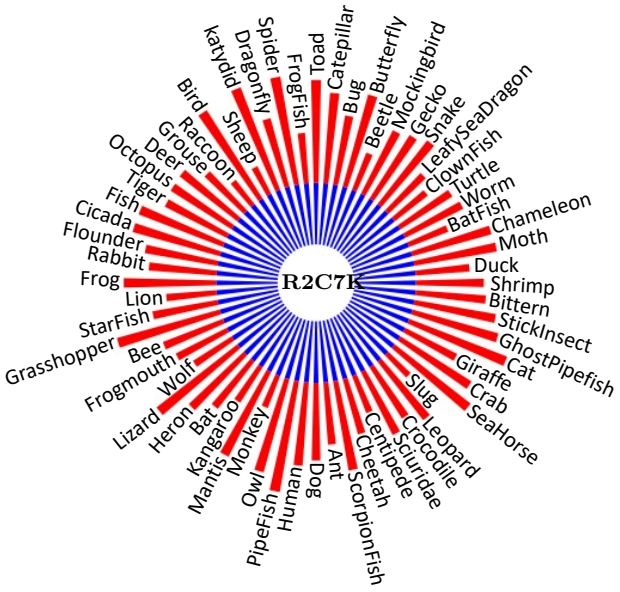
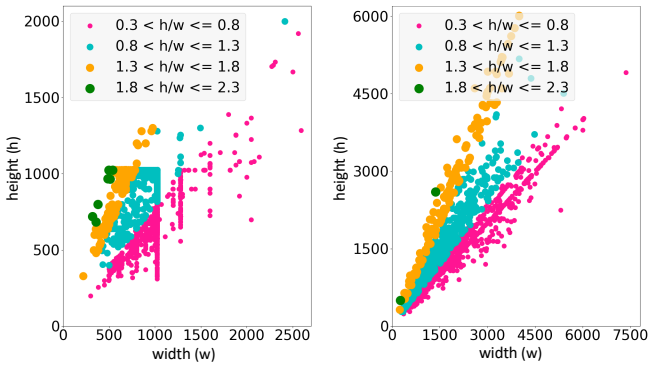


图 4. R2C7K 数据集的分类体系与日志数量分布。Camo子集的结果以红色显示，Ref子集的结果以蓝色显示。



(a) Camosubset (b) Refsubset

图 5. Camo子集与Ref子集中的图像分辨率分布情况。

与现有数据集的比较. 我们将所提出的R2C7K数据集与现有的伪装目标检测数据集进行了多维度的对比，包括CHAMELEON [69]、CAMO [34]、COD10K [11]和NC4K [53]。如表 1所示，CHAMELEON 数据集是通过在Google 搜索引擎中使用关键词“concealed animal” 收集得到的，仅包含76 张图像。CAMO 数据集包含2,500 张图像，涵盖8 个伪装物体类别。COD10K 和NC4K 数据集包含更多伪装目标图像，分别为5,066 和4,121 张。这些数据集极大地促进了伪装目标检测的发展。然而，当面对真实场景中包含多种类型伪装目标的情况时，这些数据集仍存在一定的局限性，从而影响其实际应用。相比之下，我们的R2C7K数据集包含大量的简单场景图像（共1,600 张），涵盖多个物体类别（64 类），用户可根据需求选择这些图像作为参考，从而在复杂场景图像（共5,015 张）中搜索指定的伪装目标。因此，该数据集具有更广泛的应用潜力，可服务于更多的任务。

4 网络架构

在本节中，我们首先在节 4.1 中简要描述我们提出的指向性伪装目标检测任务的定义。随后，在节 4.2 中介绍我们方

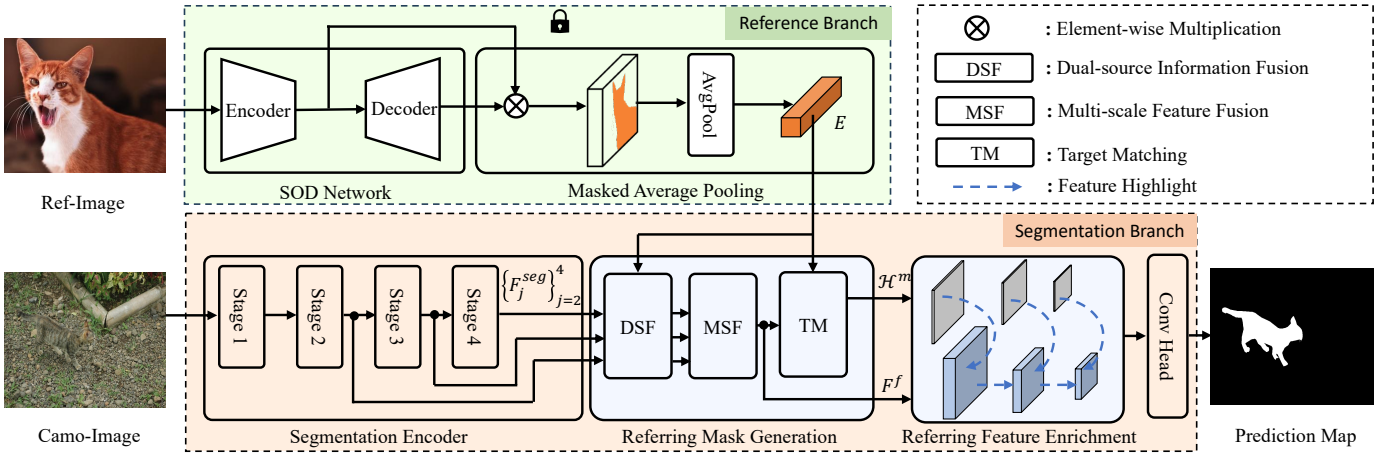


图 6. R2CNet 框架的整体架构，由两个分支组成，分别是绿色的参考分支(Reference Branch)和橙色的分割分支(Segmentation Branch)。在参考分支中，指定目标的通用表征由一组参考图像中提取。具体做法是，使用显著目标网络生成的前景图，对视觉特征进行掩码和池化操作，得到目标的语义表征。在分割分支中，我们提取编码器最后三层的视觉特征，用以表示待处理的图像。随后，这两类特征表示被送入设计精巧的RMG 模块中进行融合与匹配，生成一个掩码先验。该掩码先验在RFE 模块中用于增强不同尺度间的视觉特征，以突出图像中的伪装目标。最后，增强后的特征被送入解码器，生成最终的分割图。

法**R2CNet**的整体架构。接着，我们将在节 4.3 和节 4.4 中分别详细介绍参考掩码生成（**RMG**）模块和参考特征增强（**RFE**）模块的实现细节。

4.1 任务描述

我们提出的指向性伪装目标检测旨在在多个参照图像的引导下，对指定的伪装目标进行分割。与传统的伪装目标检测不同，后者旨在检测图像中所有的伪装目标，而指向性伪装目标检测则尝试通过匹配参照中提供的目标，仅识别出特定类别的对象。具体来说，指向性伪装目标检测系统的输入由两部分组成：第一部分是包含伪装目标的图像，记作$I^{camo} \in \mathbb{R}^{3\times H\times W}$；另一部分是包含显著目标对象的若干参照图像，记作$I^{ref} = \{I_i^{ref}\}_{i=1}^{K}, I_i^{ref} \in \mathbb{R}^{3\times H\times W}$，其中$H$和$W$分别表示图像的高和宽，$K$表示参照图像的数量。需要指出的是，$I^{camo}$来自Camo子集，且包含特定类别$c$的伪装目标；同时，$I_i^{ref}$则从Ref子集中采样，其显著目标同属于类别$c$。指向性伪装目标检测 的输出为一个二值掩码$M^{seg}$，用于标注$I^{camo}$中属于类别$c$的伪装目标区域。

4.2 整体架构

图 6 展示了我们提出的R2CNet的整体架构。如图所示，该框架由两个分支组成，即参考分支和分割分支，其具体细节将在下文中详细介绍。

参考分支（Reference Branch） 该分支用于从参照图像中提取通用表征，其流程由一个基于编码器-解码器结构的显著目标检测网络和一个掩码平均池化函数串联构成。其中，显著目标检测网络用于提取参照图像的视觉特征和前景预测结果，而掩码平均池化函数（MAP）则用于过滤掉参照图像中的无关信息。默认情况下，我们选择使用基于ResNet-50 [20]主干的预训练ICON [103] 模型作为显著目标检测网络。

给定K张尺寸为$H \times W$的参照图像，其包含显著目标对象，我们首先通过显著目标检测网络的编码器与解码器，提取空间尺寸为$\frac{H}{32} \times \frac{W}{32}$的视觉特征$\{F_k^{ref}\}_{k=1}^{K}$，以及空间尺寸为$H \times W$的前景掩码$\{M_k^{ref}\}_{k=1}^{K}$。随后，这两组特征被输入掩码平均池化函数以计算与前景对象相关的表示，记为$F_k^{obj} \in \mathbb{R}^{c_d\times 1\times 1}$。其计算公式如下：

$$F_k^{obj} = \mathcal{F}_{\text{conv1}\times 1}\left(\frac{\sum_{2d}(\mathcal{F}_{\text{down}}(M_k^{ref}) \otimes F_k^{ref})}{\sum_{2d}(F_k^{ref})}\right), \quad (1)$$

其中$\otimes$表示逐元素乘法操作，$\mathcal{F}_{\text{down}}(\cdot)$为双线性下采样操作，用于形状对齐，$\sum_{2d}(\cdot)$表示对空间维度求和操作，$\mathcal{F}_{\text{conv1}\times 1}(\cdot)$表示将通道数变换为$c_d$的$1 \times 1$卷积操作，以在效率和性能之间取得更优的平衡。最终，我们通过对上述目标表示进行平均，获得目标对象在嵌入空间中的通用表征，记为$E \in \mathbb{R}^{c_d\times 1\times 1}$。

分割分支（Segmentation Branch） 我们构建的分割分支同样采用了编码器-解码器结构，这在伪装目标检测研究中被广泛使用。需要指出的是，本文的重点在于提出一种新范式，用于定向分割伪装目标，因此我们在架构设计上并未投入过多精力。事实上，我们发现即使使用一个简单的分割网络，在所提出的范式下依然可以取得良好性能。因此，我们采用ResNet-50 [20] 作为编码器，并按照已有研究 [11], [26]，选取其后三个阶段的输出作为视觉表示。解码器则由两个卷积层组成，负责伪装目标的识别。我们还提出了两个新模块，即参考掩码生成（RMG）和参考特征增强（RFE），它们插入于编码器和解码器之间，用于利用参考分支提供的通用表征来显式分割伪装目标。

给定一张包含伪装目标的图像（尺寸为$H \times W$），我们首先从编码器的后三个阶段提取多尺度特征，并通过1×1卷积将通道数统一转换为c_d。我们将这些特征表示记为$\{F_j^{seg}\}_{j=2}^{4}$，其中$F^f \in \mathbb{R}^{c_d\times \frac{H}{8}\times \frac{W}{8}}$ and $\mathcal{H}^m \in \mathbb{R}^{1\times \frac{H}{8}\times \frac{W}{8}}$。

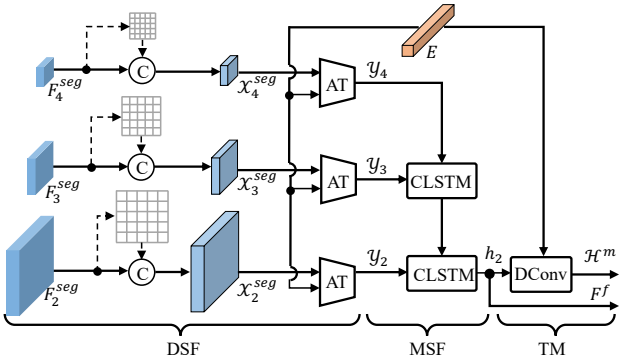


图 7. 参考掩码生成（RMG）模块的结构细节。其中，‘C’表示拼接操作，‘AT’表示仿射变换，‘CLSTM’表示卷积长短时记忆网络，‘Dconv’表示动态卷积。

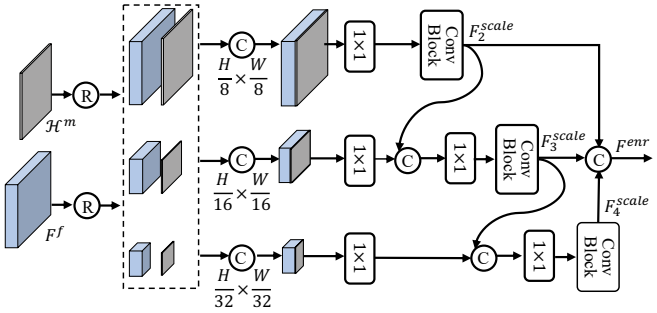


图 8. 参考特征增强（RFE）模块的结构细节。其中，‘R’表示尺寸调整操作，‘1 × 1’表示卷积核大小为1的卷积层，‘Conv Block’由两个卷积核大小为3的卷积层组成。

然后，将这些特征与参考分支提取的通用表征E 一起输入至RMG 模块，以生成融合特征F^f 和参照掩码$\mathcal{H}^m$，其公式如下：

$$F^f, \mathcal{H}^m = \mathrm{RMG}(\{F_j^{seg}\}_{j=2}^4, E). \tag{2}$$

其中$\mathcal{H}^m$ 是一个热力图掩码，其分数越高表示对应位置与通用表征的相关性越强，反之亦然。

接着，我们在RFE 模块中利用该掩码引导融合特征进行多尺度增强，以突出图像中的伪装目标，定义如下：

$$F^{enr} = \mathrm{RFE}(F^f, \mathcal{H}^m). \tag{3}$$

最后，将增强后的特征F^{enr} 输入解码器，以生成最终的分割掩码$M^{seg} \in \mathbb{R}^{1\times H\times W}$ 。

4.3 参考掩码生成

为了根据图像引导准确识别高度相似背景中的伪装物体，需要在目标物体的通用表征和视觉特征之间进行像素级的对比。然而，即使属于同一类别，伪装物体的外观可能与参考图像中的目标显著不同。此外，通用表征和视觉特征来自不同的信息源。这种较大的信息差异可能会干扰对比过程，使得伪装目标的定位变得困难。

受近期多模态融合工作的启发 [25], [37]，我们通过在通用表征和视觉特征之间进行双源信息融合（DSF）来解决这一问题，如图 7所示。为了促进这两类信息的交互，首先将空间位置信息注入到视觉特征中。具体而言，视觉特征的每个位置都被拼接上一个8维的嵌入向量，类似于 [86]中的实现。我们将注入空间信息后的视觉特征记为$\{x_j^{seg}\}_{j=2}^4$，其中$x_j^{seg} \in \mathbb{R}^{(c_d+8)\times \frac{H}{2^{j+1}}\times \frac{W}{2^{j+1}}}$ 。

接着，我们在通用表征的指导下，对视觉特征施加仿射变换。具体来说，使用两个线性层将通用表征映射成两个系数向量，这些系数向量随后分别应用于视觉特征，随后通过卷积和ReLU操作，得到结果$\{y_j\}_{j=2}^4$，其中$y_j \in \mathbb{R}^{c_d\times \frac{H}{2^{j+1}}\times \frac{W}{2^{j+1}}}$ 。

该过程定义如下：

$$y_j = \mathcal{F}_{relu}(\mathcal{F}_{conv3\times3}(\mathcal{F}_{relu}(\gamma_j \otimes x_j^c \oplus \beta_j))), \tag{4}$$

$$\gamma_j = \mathcal{F}_{mlp1}(E), \beta_j = \mathcal{F}_{mlp2}(E), \tag{5}$$

其中$\otimes$和$\oplus$分别表示逐元素乘法和加法，$\mathcal{F}_{conv3\times3}(\cdot)$是用于通道恢复的3 × 3卷积，$\gamma_j$和$\beta_j$是两个系数向量，$\mathcal{F}_{mlp1}(\cdot)$和$\mathcal{F}_{mlp2}(\cdot)$分别表示两个线性层。

此外，为增强方法对不同尺度目标的鲁棒性，基于卷积LSTM [68]的多尺度融合（MSF）在自底向上的路径上被应用，过程定义为：

$$h_j, c_j = \mathcal{F}_{clstm}(y_j, [\mathcal{F}_{up}(h_{j+1}), \mathcal{F}_{up}(c_{j+1})])), \tag{6}$$

其中$\mathcal{F}_{\mathrm{up}}(\cdot)$是双线性上采样操作用于尺寸匹配，$\mathcal{F}_{clstm}$表示卷积LSTM单元。注意初始状态为$h_4 = c_4 = y_4$，最终隐藏状态$h_2 \in \mathbb{R}^{c_d\times \frac{H}{8}\times \frac{H}{8}}$被用作融合特征$F^f$ 。

最后，将通用表征与融合特征中每个位置的表示进行比较，生成指示目标的掩码$\mathcal{H}^m$。受 [29], [42]的启发，该目标匹配（TM）过程通过动态卷积实现。具体来说，通用表征和融合特征分别作为参考引导的动态卷积核和输入。

4.4 参考特征增强

基于上文生成的参考掩码，我们设计了一个参考特征增强（RFE）模块，用于在不同尺度上增强视觉特征，如图 8所示。

具体来说，先将先验掩码和融合特征分别调整为前述三种特征的尺寸，即$\{\frac{H}{2^{j+1}} \times \frac{W}{2^{j+1}}\}_{j=2}^4$。然后，在相同尺度下，将调整后的掩码与特征通过拼接融合，不同尺度的输出特征记为$\{F_j^{scale}\}_{j=2}^4$，其中$F_j^{scale} \in \mathbb{R}^{c_d\times \frac{H}{2^{j+1}}\times \frac{W}{2^{j+1}}}$，这些特征再被整体拼接起来，以增强对伪装目标的识别能力。我们将增强后的特征表示为$F^{enr} \in \mathbb{R}^{c_d\times \frac{H}{8}\times \frac{W}{8}}$ 。正如PFENet [75] 中所提，尺寸较小的目标在下采样特征图中可能变得模糊不清。因此，我们构建了一个类似的跨尺度路径，从细粒度特征到粗粒度特征，实现尺度间的交互。此外，我们还对$\{F_j^{scale}\}_{j=2}^4$ 施加监督，使得最终得到的增强特征F^{enr}更加鲁棒，对应的监督掩码记为$\{M_j^{scale}\}_{j=2}^4$，其中$M_j^{scale} \in \mathbb{R}^{1\times H\times W}$ 。

表 2
流行的伪装目标检测模型与其指向性伪装目标检测对应模型的比较。所有模型均在NVIDIA RTX 3090 GPU 上进行评估。'R-50': ResNet-50 [20], 'E-B4': EfficientNet-B4 [74], 'R2-50': Res2Net-50 [18], 'R^3-50': Triple ResNet-50 [57], 'Swin-S': SwinTransformer-S [47], '-Ref': 包含显著目标图像作为参考的模型, 'Attribute': 每个网络的属性, 'Single-obj': 单个伪装目标的场景, 'Multi-obj': 多个伪装目标的场景, 'Overall': 所有包含伪装目标的场景, '↑': 数值越大越好, '↓': 数值越小越好.

Models	Attribute				Overall				Single-obj				Multi-obj			
	Backbone	Params (M)	Macs (G)	Speed (FPS)	$S_m\uparrow$	$\alpha E\uparrow$	$wF\uparrow$	M↓	$S_m\uparrow$	$\alpha E\uparrow$	$wF\uparrow$	M↓	$S_m\uparrow$	$\alpha E\uparrow$	$wF\uparrow$	M↓
Baseline	R-50	25.97	21.02	185.74	0.772	0.847	0.604	0.044	0.777	0.847	0.611	0.043	0.711	0.849	0.531	0.054
R2CNet	R-50	27.15	23.23	151.47	0.805	0.879	0.669	0.036	0.810	0.880	0.674	0.035	0.747	0.872	0.602	0.046
PFNet$_{2021}$ [56]	R-50	48.55	52.99	80.33	0.791	0.876	0.651	0.040	0.795	0.876	0.656	0.039	0.74	0.868	0.594	0.051
PFNet-Ref	R-50	57.58	59.59	72.48	0.811	0.885	0.687	0.036	0.815	0.886	0.691	0.035	0.764	0.873	0.632	0.045
PreyNet$_{2022}$ [93]	R-50	38.53	116.01	59.78	0.806	0.890	0.690	0.034	0.811	0.892	0.696	0.033	0.749	0.878	0.618	0.042
PreyNet-Ref	R-50	38.70	117.60	57.04	0.817	0.900	0.704	0.032	0.822	0.900	0.709	0.032	0.763	0.898	0.645	0.041
SINetV2$_{2022}$ [11]	R2-50	26.98	24.48	98.12	0.813	0.874	0.678	0.036	0.818	0.874	0.684	0.035	0.763	0.864	0.615	0.045
SINetV2-Ref	R2-50	27.70	26.01	86.60	0.823	0.888	0.700	0.033	0.828	0.889	0.705	0.032	0.771	0.874	0.634	0.043
BSANet$_{2022}$ [101]	R2-50	32.59	59.29	71.75	0.818	0.893	0.702	0.034	0.823	0.895	0.707	0.033	0.766	0.873	0.643	0.041
BSANet-Ref	R2-50	33.07	66.08	67.18	0.830	0.912	0.727	0.030	0.827	0.913	0.733	0.030	0.774	0.895	0.655	0.039
BGNet$_{2022}$ [72]	R2-50	79.85	116.76	66.29	0.818	0.901	0.679	0.036	0.822	0.901	0.683	0.035	0.775	0.886	0.626	0.044
BGNet-Ref	R2-50	151.06	171.03	50.69	0.840	0.909	0.738	0.029	0.844	0.910	0.742	0.029	0.792	0.887	0.679	0.036
ZoomNet$_{2022}$ [57]	R^3-50	32.38	203.50	22.89	0.813	0.884	0.688	0.032	0.818	0.885	0.695	0.031	0.747	0.870	0.605	0.042
ZoomNet-Ref	R^3-50	33.30	218.24	20.82	0.834	0.886	0.720	0.029	0.839	0.887	0.726	0.029	0.781	0.876	0.652	0.038
DGNet$_{2023}$ [26]	E-B4	19.22	5.53	110.57	0.816	0.883	0.684	0.034	0.826	0.885	0.700	0.032	0.744	0.873	0.588	0.047
DGNet-Ref	E-B4	20.10	7.24	95.06	0.821	0.891	0.696	0.032	0.827	0.890	0.703	0.031	0.748	0.879	0.607	0.045
VSCode$_{2024}$ [52]	Swin-S	74.72	59.81	76.81	0.819	0.879	0.702	0.033	0.825	0.880	0.706	0.032	0.750	0.868	0.651	0.043
VSCode-Ref	Swin-S	76.63	64.28	65.26	0.832	0.891	0.714	0.030	0.838	0.892	0.718	0.029	0.766	0.880	0.662	0.041
ZoomNext$_{2024}$ [58]	R^3-50	28.46	185.79	66.29	0.838	0.897	0.742	0.032	0.843	0.898	0.750	0.031	0.777	0.880	0.655	0.040
ZoomNext-Ref	R^3-50	30.32	197.43	52.81	0.850	0.909	0.755	0.027	0.859	0.910	0.762	0.026	0.788	0.892	0.675	0.037

5 实验

在本节中，我们首先介绍本文的实验设置，包括训练与测试协议、超参数细节以及评价指标。随后，我们对指向性伪装目标检测方法与对应的伪装目标检测方法进行了定量对比。特别地，我们将所提出的R2CNet 与基础模型在不同伪装场景下进行了比较，同时将指向性伪装目标检测 的设计应用于7种最新的最优的伪装目标检测方法，以验证其通用性。接下来，我们报告消融实验的结果，以分析各组件的有效性以及设计选择的合理性。最后，我们进行可视化比较，以更直观地展示指向性伪装目标检测的优势。

5.1 实验设置

训练与测试. 鉴于指向性伪装目标检测的目标是生成与标注接近的二值前景图，因此我们采用结构损失函数 [82]作为指向性伪装目标检测模型的优化目标。该损失函数由BCE（Binary Cross Entropy）损失与IoU（Intersection over Union）损失组成，已被广泛应用于多种二值分割任务中。

具体而言，该损失函数可定义为：

$$\mathcal{L}(P,G)=\sum_{i=1}^{4}\mathcal{L}_{bce}(P_i,G)+\mathcal{L}_{iou}(P_i,G), \tag{7}$$

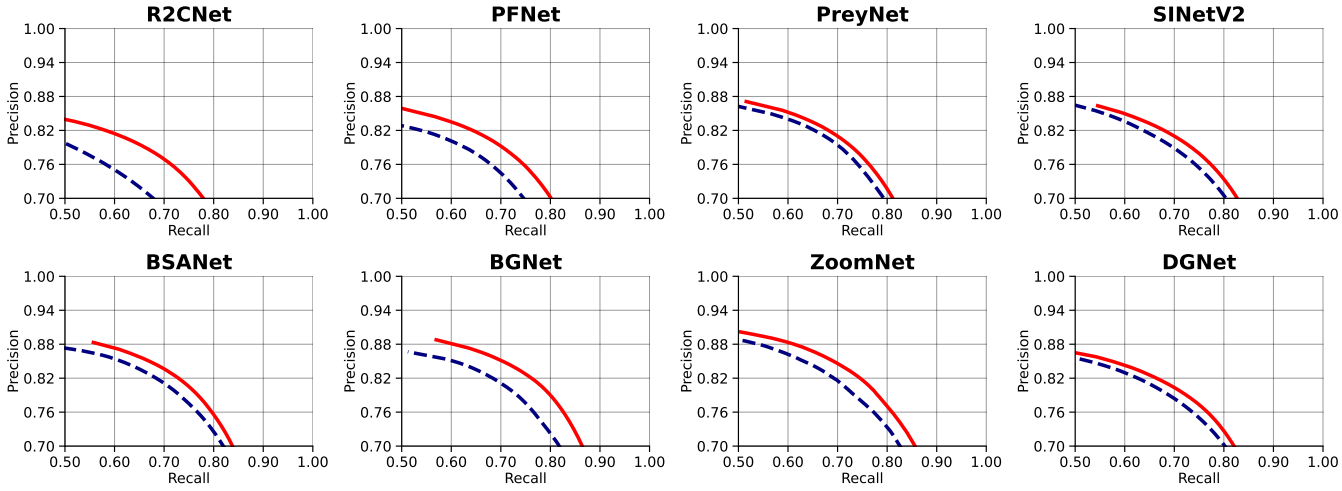
其中$P=\{M_2^{scale},M_3^{scale},M_4^{scale},M^{seg}\}$表示模型生成的预测图，$G$表示对应的真实标注图（ground truth）。

超参数细节. 在默认设置下，指向性伪装目标检测中参考图像的数量设为5。在训练阶段，我们冻结前景预测网络的参数，批量大小设置为32，优化器采用Adam [32]，初始学习率设为5e-4，并根据余弦退火策略 [49]逐步衰减。在推理阶段，所有输入图像首先被调整为352 × 352的大小，然后输入到训练好的R2CNet中以生成最终预测结果，整个过程中不使用任何后处理操作。所有实验均在PyTorch [59]框架下实现。

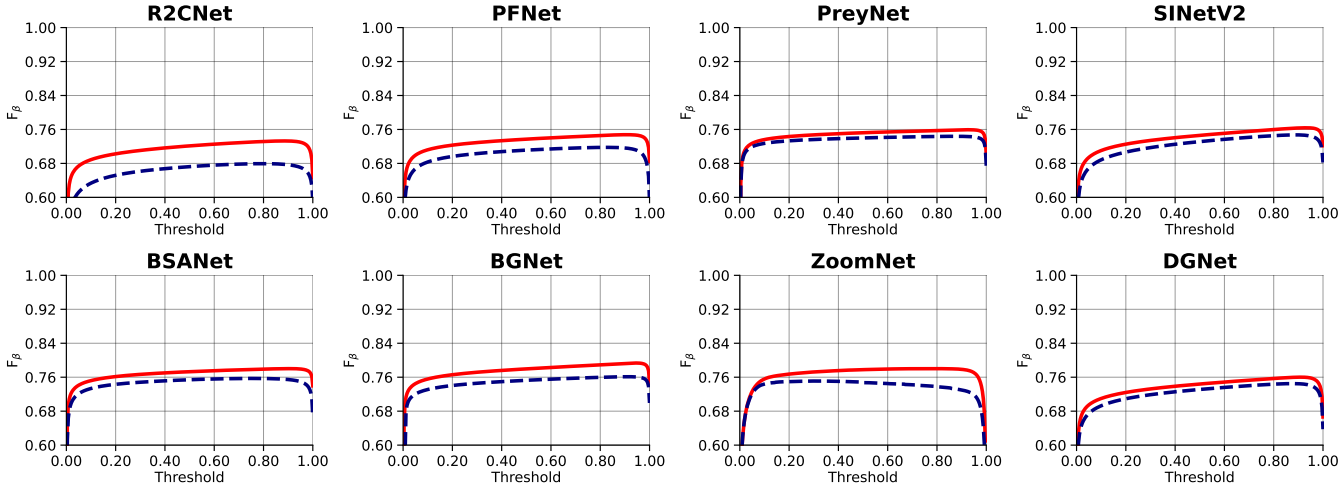
评价指标. 按照伪装目标检测领域的标准评估协议，我们采用四种常用指标进行性能评估，包括平均绝对误差（Mean Absolute Error, M）[60]、结构相似度指标（Structure-measure, S_m）[9]、自适应增强度量（Adaptive E-measure, αE）[10] 以及加权F-度量（Weighted F-measure, wF）[55]。具体而言，M用于衡量预测掩码与真实标注图之间的绝对差异；S_m用于评估预测图与真实标注图在区域感知与目标感知上的结构相似性；αE衡量逐像素与整图级别的相似性；而wF综合衡量了预测与真实标注图之间的召回率与精确率。此外，我们还绘制精确率-召回率（Precision-Recall, PR）曲线和F_β-阈值（F_β）曲线，以便进行更全面的比较分析。

5.2 定量评估

与基线模型的比较. 为了验证我们提出的R2CNet的有效性，我们首先将其与其基线变体进行比较，该基线模型是一个基



(a) PR curves



(b) F_β curves

图 9. 伪装目标检测方法及其指向性伪装目标检测变体的PR 曲线和F_β曲线。标准伪装目标检测方法的结果用蓝色虚线表示，指向性伪装目标检测变体的结果用红色实线表示。

表 3
将指向性伪装目标检测范式应用于近期发表的显著目标检测模型中。

Methods	Publication	$S_m \uparrow$	$\alpha E \uparrow$	$wF \uparrow$	$M \downarrow$	Params (M)	MACs (G)	FPS
VST++	TPAMI2024	0.846	0.872	0.795	0.063	53.6	28.4	185.5
VST++-Ref		0.858	0.885	0.810	0.057	55.2	34.1	166.9
VSCode	CVPR2024	0.861	0.893	0.811	0.054	74.7	59.8	144.6
VSCode-Ref		0.874	0.906	0.820	0.052	76.6	64.3	128.3

于编码器-解码器架构的标准伪装目标检测模型。在该基线模型中，编码器输出的最后三个特征按照FPN的方式在自底向上的路径中进行融合，融合后的特征被输入到解码器中以直接分割伪装目标，而不使用任何参考图像。如表 2 所示，我们的R2CNet在R2C7K的整体场景（即测试集）中在所有指标上均大幅超越了基线模型。为了更全面地评估模型性能，我们进一步将整体场景划分为两类：单个伪装物体场景和多个伪装物体场景。结果表明，我们的R2CNet在所有设置中仍然优于其伪装目标检测基线模型，充分说明参考图像的引入有助于模型从复杂背景中识别伪装目标。

在现有伪装目标检测方法上的应用. 为了验证我们提出的指向性伪装目标检测设计的通用性，我们还将该思想应用于现有的伪装目标检测方法。这些方法的选择基于三个标准：a）近期发表，b）具有代表性，c）开源代码可用。具

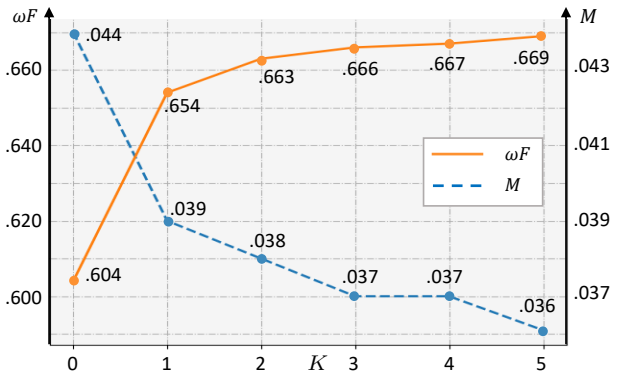


图 10. 对参考图像数量的消融研究。蓝色虚线与橙色实线分别表示两个伪装目标检测指标（即wF 与M）随参考图像数量变化的趋势。

体而言，参与适配的模型包括：PFNet [56], PreyNet [93], SINetV2 [11], BSANet [101], BGNet [72], ZoomNet [57], DGNet [26], VSCode [52], 以及ZoomNext [58]。如表 2 的第3行至最后一行以及图 9 所示，这些模型在采用指向性伪装目标检测思路进行适配后，在所有指标上均优于其原始版本。值得注意的是，我们的轻量级基线模型在引入参考图像后，其性能甚至可与近期的伪装目标检测方法相媲美。这表明我们的方法并不依赖于强大的分割网络。即使是性能不突出的伪装目标检测模型，在加入参考分支之后，也能获得显著的性能提升，体现了我们提出的指向性伪装目标检测思路的有效性。此外，指向性伪装目标检测方法的各项指标的相对排序几乎与其对应的伪装目标检测方法一致，说明在应用我们的设计后，这些伪装目标检测算法原本的性能优势仍然能够得到保留。

在现有显著目标检测方法上的应用. 为了进一步验证我们指向性伪装目标检测设计在显著目标检测任务上的通用性，我们在CoSOD3K [15] 数据集上进行了实验，将指向性伪装目标检测的设计应用于最新的显著目标检测模型，即VSCode [52] 和VST++ [45]。如表 3 所示，我们的指向性版本在所有指标上均优于其原始的显著目标检测方法。

5.3 消融实验

参考图像数量. 我们对参考图像数量（即K）对指向性伪装目标检测的影响进行了消融实验。考虑到在R2C7K测试集中，每个类别包含5 张参考图像，因此K的取值范围为0到5。当$0 < K < 5$时，我们通过三次评估的平均结果来计算R2CNet的性能，在每次评估中，该模型由从测试集中某一类别中随机采样的K张参考图像进行引导。如图 10 所示，随着参考图像数量的增加，R2CNet的性能持续提升。我们认为这是因为在此过程中获得的共性信息受样本差异影响较小，从而实现了对伪装目标更加准确的定位。

模型组件. 我们分析了R2CNet中各个组件的重要性，实验结果如表 4 所示。从表中可以看出，我们提出的两个模块均能显著提升模型性能。其中，RMG 模块将wF 指标从0.604 提升至0.661，RFE 模块将该指标提升至0.644。当两个模块联合使用时，模型性能进一步提升至0.669（相较于初始提升了10.8%），表现更加突出。这些结果验证了我们所提出的两个模块能够在计算开销有限的情况下，通过参考图像有效提升伪装目标的分割性能。

RMG模块. 我们对RMG模块中的参考特征与视觉特征的跨源融合和跨尺度融合部分分别设计了两种变体。如表 5 所示，我们最终采用的融合策略在所有四种组合中取得了最佳性能。

RFE模块. 我们探索了四种可行的特征增强策略，其中包括两种即插即用方式以及两种我们提出的多尺度增强变体。如表 6 所示，我们精心设计的增强方案在保持合理推理速度的前提下，取得了更优的性能，尤其是包含跨尺度路径的变体表现最为突出。

模型维度. 模型的通道维度（c_d）对其参数量和推理速度具有重要影响，因此我们设计了一系列实验以选择合适的维度大小。具体而言，我们将c_d从32 以倍增的方式逐步增加至256，并在此过程中统计了模型的参数规模、计算开销与性能变化，结果如表 7 所示。从表中可以看出，随着c_d的增加，R2CNet 的性能持续提升。然而，当c_d大于64 后，性能提升的幅度逐渐减缓，而模型的参数量和计算开销却急剧上升。例如，当c_d从64 提升至128 时，性能仅有轻微提升，但计算量却增加了约35%（No.2 *vs.* No.3）。因此，我们最终将c_d设置为64，以在模型性能与计算效率之间取得权衡。

参考形式. 我们还探究了三种不同形式的参考信息，即文本描述、包含伪装目标的图像以及包含显著目标的图像。如节 1 与节 2 所述，获取与目标图像中伪装目标相关的详细文本描述或标注好的伪装图像在现实中较为困难。因此，我们仅考虑易获取版本的文本描述和伪装图像，并与我们采用的参考形式（即显著目标图像）进行性能上限对比。

在文本参考方面，受近期提示词工程的启发，我们基于‘a photo of [CLASS]’ 的模板构建文本描述，其中‘[CLASS]’为R2C7K中的64 个类别之一。然后，我们将这64 条描述输入采用ResNet-50 作为视觉主干的预训练模型CLIP [63]，按照CLIPSeg [50] 的方式提取其文本特征，作为目标类别的通用表征。需要注意的是，训练集与测试集中属于同一类别的样本共享相同的文本特征。

在伪装图像参考方面，我们从目标对象所属类别的Camo子集中随机采样若干图像，并用ResNet-50 提取其视觉特征。这些特征通过GT 掩码进行筛选和池化，获取该类别的通用表征。注意在该设置下，训练与测试阶段的图像参考来自不同图像。

实验结果如表 8 所示。需指出的是，文本参考使用了大型预训练模型，并且测试时的类别文本已在训练中出现；而伪装图像参考则使用了GT掩码获得更精确的通用表征。尽管如此，我们采用显著目标图像作为参考的R2CNet仍然在所有评估指标上取得更优的表现。这些实验结果验证了我们所选参考形式的合理性。与文本和伪装图像相比，显著目标图像更易获取，且在提升伪装目标定位与分割性能方面更有效。

表 4
对我们方法各组成模块的消融研究。RMG: 参考掩码生成模块. RFE: 参考特征增强模块. 该实验同时考虑了性能与计算开销两个方面。

No.	RMG	RFE	$S_m \uparrow$	$\alpha E \uparrow$	$wF \uparrow$	$M \downarrow$	Params (M)	MACs (G)	FPS
1	✗	✗	0.772	0.847	0.604	0.044	25.97	21.02	185.74
2	✓	✗	0.800	0.870	0.661	0.038	26.40	22.30	169.36
3	✗	✓	0.792	0.869	0.644	0.040	26.55	21.95	166.57
4	✓	✓	**0.805**	**0.879**	**0.669**	**0.036**	27.15	23.23	151.47

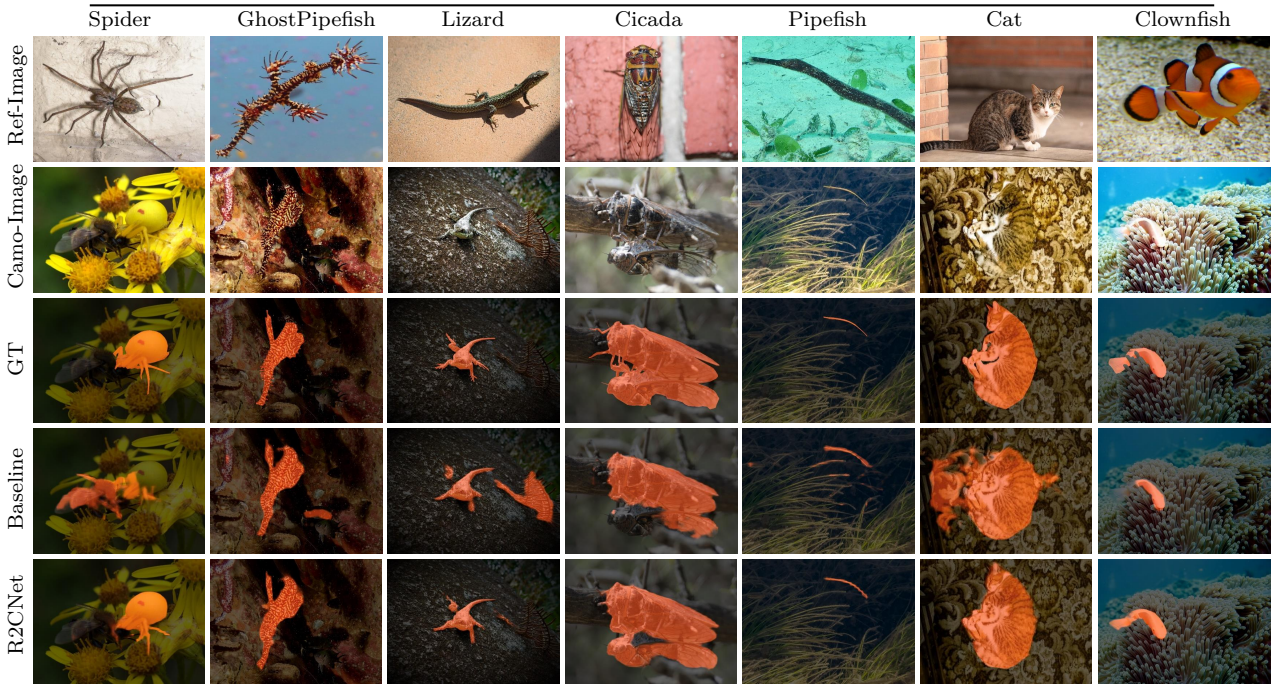


图 11. 所提指向性伪装目标检测方法（R2CNet）与标准伪装目标检测方法（基线模型）预测结果的可视化对比。分割掩码以紫色显示。

表 5
RMG模块中不同融合策略的消融实验。'DSF': 通用表征与视觉特征之间的融合；'MSF': 不同尺度特征之间的融合。'$\mathcal{F}_{multiply}$': 逐元素乘法；'$\mathcal{F}_{at}(\cdot)$': 仿射变换；'$\mathcal{F}_{concate}$': 特征拼接；'$\mathcal{F}_{clstm}(\cdot)$': 卷积LSTM单元。

No.	DSF	MSF	$S_m \uparrow$	$\alpha E \uparrow$	$wF \uparrow$	$M \downarrow$
1	$\mathcal{F}_{multiply}$	$\mathcal{F}_{concate}$	0.801	0.870	0.656	0.039
2	$\mathcal{F}_{multiply}$	$\mathcal{F}_{clstm}$	0.802	0.872	0.659	0.038
3	$\mathcal{F}_{at}$	$\mathcal{F}_{concate}$	0.804	0.872	0.666	0.037
4	$\mathcal{F}_{at}$	$\mathcal{F}_{clstm}$	**0.805**	**0.879**	**0.669**	**0.036**

表 6
RFE模块中特征增强策略的消融实验。'w/o CSP'或'w/ CSP': 在无交叉尺度路径和有交叉尺度路径的情况下进行多尺度特征增强。

No.	Setting	Speed (FPS)	$S_m \uparrow$	$\alpha E \uparrow$	$wF \uparrow$	$M \downarrow$
1	ASPP [3]	143.07	0.803	0.871	0.663	0.038
2	RFB [46]	**144.03**	0.801	0.872	0.662	0.038
3	w/o CSP	134.13	0.803	0.873	0.665	0.037
4	w/ CSP	130.95	**0.805**	**0.879**	**0.669**	**0.036**

表 7
关于我们的R2CNet中通道数设置的消融实验。

No.	Setting	Macs (G)	Params (M)	$S_m \uparrow$	$\alpha E \uparrow$	$wF \uparrow$	$M \downarrow$
1	$c_d = 32$	**10.6**	**24.0**	0.799	0.871	0.655	0.039
2	$c_d = 64$	11.7	25.0	0.805	0.879	0.669	**0.036**
3	$c_d = 128$	15.8	29.1	0.807	0.875	0.672	0.037
4	$c_d = 256$	32.2	44.4	**0.811**	**0.884**	**0.679**	**0.036**

表 8
指向性伪装目标检测参考形式的消融实验。text-ref: 使用简单提示的文本参考；camo-ref: 含伪装目标的图像参考；sal-ref: 含显著目标的图像参考；LSM: 使用大规模预训练模型；GT: 使用参考图像的真实掩码；

No.	Methods	LSM	GT	Seen	$S_m \uparrow$	$\alpha E \uparrow$	$wF \uparrow$	$M \downarrow$
1	Baseline				0.772	0.847	0.604	0.044
2	+ text-ref	✓		✓	**0.805**	0.872	0.661	0.038
3	+ camo-ref		✓		0.801	0.869	0.656	0.039
4	+ sal-ref				**0.805**	**0.879**	**0.669**	**0.036**

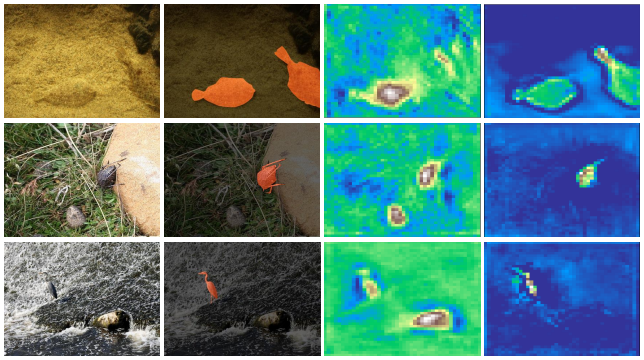


图 12. 不同方法在有无参考信息条件下中间特征表示的可视化对比。

5.4 定性评估

结果可视化. 我们首先在图 11中展示了所提出的基于参考的伪装目标检测方法（即R2CNet）与无类别参考的伪装检测方法（即基线）的预测结果可视化比较。如图中从第1列（蜘蛛）到第4列（蝉）所示，这些场景中存在多个伪装目标。基线模型往往倾向于不加区分地分割所有潜在的伪装目标，因此其输出结果通常不够准确。具体而言，该模型可能漏检部分真实的伪装目标（例如蝉），也可能将其他非目标区域误判为伪装目标（例如蜘蛛和蜥蜴）。相比之下，R2CNet能够根据参考图像有效地定位对应的伪装目标。在包含多个相似干扰物体的复杂场景中（如海龙），该方法对干扰项的鲁棒性更强；而在仅包含单个伪装目标但目标与背景差异极小的场景中（如第6列的猫和第7列的小丑鱼），R2CNet也能更完整地分割出目标的区域。我们认为，这种优势来源于模型从参考图像中学习到的某一类别对象的通用表征，这种表示能够有效地辅助模型定位并分割出该类别的伪装目标主体。

特征可视化. 为了进一步理解指向性伪装目标检测，我们还在图 12 中展示了是否包含参考信息时的中间特征可视化结果。如图所示，未使用参考信息的模型虽然可以大致定位伪装目标，但容易受到其他物体的干扰，导致分割结果不够准确。而在引入参考信息之后，模型能够更加聚焦于指定的伪装目标，从而避免被无关物体干扰。

6 未来工作

基于本文提出的R2C7K数据集与R2CNet框架，关于指向性伪装目标检测的未来研究可以从以下几个方向进行探索：

1）参考信息的其他形式. 近年来，多模态研究在视觉与语言、视觉与语音等领域取得了显著进展。基于这些成果，探索使用其他类型的参考信息（如文本、语音）来扩展指向性伪装目标检测是一个有趣的方向。正如上文所讨论的，这一方向的主要挑战在于如何高效、低成本地获取这些参考信息。

2）参考场景的特殊情况. 本文默认被分割图像中包含与参考图像指定的目标相同的对象。然而，在实际应用中，某些场景中目标可能并不存在。因此，使指向性伪装目标检测能够兼容这些实际场景中的“空目标”情况，将是一项具有前景的拓展方向。

3）相关任务拓展. 在某些应用场景中，用户更关注简单明确的答案以辅助决策，而不是获取伪装目标的完整分割结果。因此，将指向性伪装目标检测拓展至与视觉问答等轻量任务相关的研究方向，也具有一定的实用价值。

7 结论

本文提出了一个新颖的基准任务指向性伪装目标检测，旨在通过简单的图像参考信息，实现对伪装目标的定向分割。首先，我们构建了一个大规模的真实场景图像数据集（即R2C7K），为该任务提供了坚实的数据基础。随后，我们设计了一个端到端的R2CNet框架，该框架采用双分支结构。其中，我们精心设计的RMG 模块与RFE 模块有效促进了参考图像中的通用表征向目标图像中伪装目标的引导，从而显著提升了分割性能。与传统的伪装目标检测方法在背景中无差别分割伪装目标相比，R2CNet在多项主流评估指标上均取得了明显更优的结果，并在视觉效果上更具可解释性与实用性。我们相信，构建一个基于多源信息协同的系统将成为未来的趋势，也希望指向性伪装目标检测所提供的研究视角能够为后续相关研究带来新的启发。

References

[1] Ali Borji, Simone Frintrop, Dicky N Sihite, and Laurent Itti. Adaptive object tracking by learning background context. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2012.

[2] Geng Chen, Si-Jie Liu, Yu-Jia Sun, Ge-Peng Ji, Ya-Feng Wu, and Tao Zhou. Camouflaged object detection via context-aware cross-level fusion. *IEEE Trans. Circuit Syst. Video Technol.*, 32(10):6981–6993, 2022.

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2017.

[4] Shuhan Chen, Xiuli Tan, Ben Wang, and Xuelong Hu. Reverse attention for salient object detection. In *Eur. Conf. Comput. Vis.*, 2018.

[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] Ming-Ming Cheng, Fang-Lue Zhang, Niloy J Mitra, Xiaolei Huang, and Shi-Min Hu. Repfinder: finding approximately repeated scene elements for image editing. *ACM Trans. Graph.*, 29(4):1–8, 2010.

[7] Yao Cheng, Hao-Zhou Hao, Yi Ji, Ying Li, and Chun-Ping Liu. Attention-based neighbor selective aggregation network for camouflaged object detection. In *Inter. Joint Conf. on Neural Net.*, pages 1–8, 2022.

[8] Mu-Chun Chou, Hung-Jen Chen, and Hong-Han Shuai. Finding the achilles heel: Progressive identification network for camouflaged object detection. In *Int. Conf. Multimedia and Expo*, 2022.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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*, 2023.

[14] 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.

[15] Deng-Ping Fan, Tengpeng Li, Zheng Lin, Ge-Peng Ji, Dingwen Zhang, Ming-Ming Cheng, Huazhu Fu, and Jianbing Shen. Re-thinking co-salient object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(8):4339–4354, 2021.

[16] Deng-Ping Fan, Tao Zhou, Ge-Peng Ji, Yi Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Inf-net: Automatic covid-19 lung infection segmentation from ct images. *IEEE Trans. Medical Imaging*, 39(8):2626–2637, 2020.

[17] Qi Fan, Deng-Ping Fan, Huazhu Fu, Chi-Keung Tang, Ling Shao, and Yu-Wing Tai. Group collaborative learning for co-salient object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

[18] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(2):652–662, 2019.

[19] Chenlei Guo and Liming Zhang. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE Trans. Image Process.*, 19(1):185–198, 2009.

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

[21] Ruozhen He, Qihua Dong, Jiaying Lin, and Rynson WH Lau. Weakly-supervised camouflaged object detection with scribble annotations. In *AAAI Conf. on Artif. Intel.*, 2022.

[22] Shengfeng He, Rynson WH Lau, Wenxi Liu, Zhe Huang, and Qingxiong Yang. Supercnn: A superpixelwise convolutional neural network for salient object detection. *Int. J. Comput. Vis.*, 115:330–344, 2015.

[23] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

[24] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *Eur. Conf. Comput. Vis.*, 2016.

[25] Binbin Huang, Dongze Lian, Weixin Luo, and Shenghua Gao. Look before you leap: Learning landmark features for one-stage visual grounding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

[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] Ya Jing, Tao Kong, Wei Wang, Liang Wang, Lei Li, and Tieniu Tan. Locate then segment: A strong pipeline for referring image segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

[30] Nobukatsu Kajiura, Hong Liu, and Shin'ichi Satoh. Improving camouflaged object detection with the uncertainty of pseudo-edge labels. In *ACM MultiMedia Asia*, 2021.

[31] Jongpil Kim and Vladimir Pavlovic. A shape-based approach for salient object detection using deep learning. In *Eur. Conf. Comput. Vis.*, 2016.

[32] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[33] Chunbo Lang, Gong Cheng, Binfei Tu, and Junwei Han. Learning what not to segment: A new perspective on few-shot segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

[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] Xinyi Le, Junhui Mei, Haodong Zhang, Boyu Zhou, and Juntong Xi. A learning-based approach for surface defect detection using small image datasets. *Neurocomputing*, 2020.

[36] 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.

[37] Ruiyu Li, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Referring image segmentation via recurrent refinement networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[38] Zhong-Yu Li, Shanghua Gao, and Ming-Ming Cheng. Exploring feature self-relation for self-supervised transformer. *arXiv preprint arXiv:2206.05184*, 2022.

[39] Jiaying Lin, Xin Tan, Ke Xu, Lizhuang Ma, and Rynson WH Lau. Frequency-aware camouflaged object detection. *ACM Trans. on Multim. Comp., Comm. and App.*, 19(2):1–16, 2023.

[40] 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.

[41] Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. Recurrent multimodal interaction for referring image segmentation. In *Int. Conf. Comput. Vis.*, 2017.

[42] Jie Liu, Yanqi Bao, Guo-Sen Xie, Huan Xiong, Jan-Jakob Sonke, and Efstratios Gavves. Dynamic prototype convolution network for few-shot semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

[43] Jiawei Liu, Jing Zhang, and Nick Barnes. Modeling aleatoric uncertainty for camouflaged object detection. In *IEEE/CVF Winter Conf. on App. of Comp. Vis.*, 2022.

[44] 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.

[45] Nian Liu, Ziyang Luo, Ni Zhang, and Junwei Han. Vst++: Efficient and stronger visual saliency transformer. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024.

[46] Songtao Liu, Di Huang, et al. Receptive field block net for accurate and fast object detection. In *Eur. Conf. Comput. Vis.*, 2018.

[47] 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.*, pages 10012–10022, 2021.

[48] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015.

[49] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[50] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

[51] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[52] 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. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 17169–17180, 2024.

[53] 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.

[54] Yiwei Ma, Jiayi Ji, Xiaoshuai Sun, Yiyi Zhou, and Rongrong Ji. Towards local visual modeling for image captioning. *Pattern Recognition*, 138:109420, 2023.

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

[56] 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.

[57] 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.

[58] Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoomnext: A unified collaborative pyramid network for camouflaged object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024.

[59] 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.

[60] 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.

[61] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

[62] Xuebin Qin, Deng-Ping Fan, Chenyang Huang, Cyril Diagne, Zichen Zhang, Adrià Cabeza Sant'Anna, Albert Suarez, Martin Jagersand, and Ling Shao. Boundary-aware segmentation network for mobile and web applications. *arXiv preprint arXiv:2101.04704*, 2021.

[63] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Inter. Conf. Mach. Learning*, 2021.

[64] Jingjing Ren, Xiaowei Hu, Lei Zhu, Xuemiao Xu, Yangyang Xu, Weiming Wang, Zijun Deng, and Pheng-Ann Heng. Deep texture-aware features for camouflaged object detection. *IEEE Trans. Circuit Syst. Video Technol.*, 2021.

[65] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Inter. Conf. on Med. Image Comp. and Computer-Assisted Interv.*, 2015.

[66] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. *arXiv preprint arXiv:1709.03410*, 2017.

[67] Hengcan Shi, Hongliang Li, Fanman Meng, and Qingbo Wu. Key-word-aware network for referring expression image segmentation. In *Eur. Conf. Comput. Vis.*, 2018.

[68] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Adv. Neural Inform. Process. Syst.*, 2015.

[69] Przemysław Skurowski, Hassan Abdulameer, J Błaszczyk, Tomasz Depta, Adam Kornacki, and P Kozieł. Animal camouflage analysis: Chameleon database. *Unpublished manuscript*, 2(6):7, 2018.

[70] Xiaoshuai Sun, Xuying Zhang, Liujuan Cao, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Exploring language prior for mode-sensitive visual attention modeling. In *ACM Int. Conf. Multimedia*, 2020.

[71] Yujia Sun, Geng Chen, Tao Zhou, Yi Zhang, and Nian Liu. Context-aware cross-level fusion network for camouflaged object detection. *arXiv preprint arXiv:2105.12555*, 2021.

[72] Yujia Sun, Shuo Wang, Chenglizhao Chen, and Tian-Zhu Xiang. Boundary-guided camouflaged object detection. In *Int. Joint Conf. on Artif. Intel.*, 2022.

[73] Domen Tabernik, Samo Šela, Jure Skvarč, and Danijel Skočaj. Segmentation-based deep-learning approach for surface-defect detection. *J. Intell. Manuf.*, 31(3):759–776, 2020.

[74] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Inter. Conf. Mach. Learning*, 2019.

[75] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(2):1050–1065, 2020.

[76] Tom Troscianko, Christopher P Benton, P George Lovell, David J Tolhurst, and Zygmunt Pizlo. Camouflage and visual perception. *Philos. Trans. R. Soc. B: Biol. Sci.*, 364(1516):449–461, 2009.

[77] Muammer Türkoğlu and Davut Hanbay. Plant disease and pest detection using deep learning-based features. *Turk J Elec Eng & Comp Sci*, 27(3):1636–1651, 2019.

[78] 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.*, 2017.

[79] Kang Wang, Hongbo Bi, Yi Zhang, Cong Zhang, Ziqi Liu, and Shuang Zheng. D ^{2} c-net: A dual-branch, dual-guidance and cross-refine network for camouflaged object detection. *IEEE*

Trans. on Industrial Electronics, 69(5):5364–5374, 2021.

[80] Lijun Wang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Deep networks for saliency detection via local estimation and global search. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015.

[81] Wenguan Wang, Jianbing Shen, Fang Guo, Ming-Ming Cheng, and Ali Borji. Revisiting video saliency: A large-scale benchmark and a new model. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[82] Jun Wei, Shuhui Wang, and Qingming Huang. F^3net: fusion, feedback and focus for salient object detection. In *AAAI Conf. on Artif. Intel.*, 2020.

[83] Mingrui Wu, Xuying Zhang, Xiaoshuai Sun, Yiyi Zhou, Chao Chen, Jiaxin Gu, Xing Sun, and Rongrong Ji. Difnet: Boosting visual information flow for image captioning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

[84] Zongwei Wu, Danda Pani Paudel, Deng-Ping Fan, Jingjing Wang, Shuo Wang, Cédric Demonceaux, Radu Timofte, and Luc Van Gool. Source-free depth for object pop-out. *arXiv preprint arXiv:2212.05370*, 2022.

[85] Mochu Xiang, Jing Zhang, Yunqiu Lv, Aixuan Li, Yiran Zhong, and Yuchao Dai. Exploring depth contribution for camouflaged object detection. *arXiv e-prints*, pages arXiv–2106, 2021.

[86] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *Int. Conf. Comput. Vis.*, 2019.

[87] Bowen Yin, Xuying Zhang, Qibin Hou, Bo-Yuan Sun, Deng-Ping Fan, and Luc Van Gool. Camoformer: Masked separable attention for camouflaged object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024.

[88] 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.

[89] Wei Zhai, Yang Cao, HaiYong Xie, and Zheng-Jun Zha. Deep texton-coherence network for camouflaged object detection. *IEEE Trans. Multimedia*, 2022.

[90] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[91] Cong Zhang, Kang Wang, Hongbo Bi, Ziqi Liu, and Lina Yang. Camouflaged object detection via neighbor connection and hierarchical information transfer. *Comp. Vis. and Image Understanding*, 221:103450, 2022.

[92] Jing Zhang, Yunqiu Lv, Mochu Xiang, Aixuan Li, Yuchao Dai, and Yiran Zhong. Depth-guided camouflaged object detection. *arXiv preprint arXiv:2106.13217*, 2021.

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

[94] Xuying Zhang, Xiaoshuai Sun, Yunpeng Luo, Jiayi Ji, Yiyi Zhou, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Rstnet: Captioning with adaptive attention on visual and non-visual words. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

[95] Xiaolin Zhang, Yunchao Wei, Yi Yang, and Thomas S Huang. Sg-one: Similarity guidance network for one-shot semantic segmentation. *IEEE Trans. on Cybern.*, 50(9):3855–3865, 2020.

[96] Zhao Zhang, Wenda Jin, Jun Xu, and Ming-Ming Cheng. Gradient-induced co-saliency detection. In *Eur. Conf. Comput. Vis.*, 2020.

[97] 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.

[98] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6):1452–1464, 2017.

[99] Tao Zhou, Yi Zhou, Chen Gong, Jian Yang, and Yu Zhang. Feature aggregation and propagation network for camouflaged object detection. *IEEE Trans. Image Process.*, 31:7036–7047, 2022.

[100] Yiyi Zhou, Rongrong Ji, Gen Luo, Xiaoshuai Sun, Jinsong Su, Xinghao Ding, Chia-Wen Lin, and Qi Tian. A real-time global inference network for one-stage referring expression comprehension. *IEEE Trans. on Neur. Net. and Learn. Sys.*, 2021.

[101] Hongwei Zhu, Peng Li, Haoran Xie, Xuefeng Yan, Dong Liang, Dapeng Chen, Mingqiang Wei, and Jing Qin. I can find you! boundary-guided separated attention network for camouflaged object detection. In *AAAI Conf. on Artif. Intel.*, 2022.

[102] Jinchao Zhu, Xiaoyu Zhang, Shuo Zhang, and Junnan Liu. Inferring camouflaged objects by texture-aware interactive guidance network. In *AAAI Conf. on Artif. Intel.*, 2021.

[103] 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.*, 45(3):3738–3752, 2023.

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

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.

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.

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.

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.

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.