

面向伪装场景理解的深度学习研究近况

范登平¹  · 季葛鹏²  · 徐鹏³  ·
程明明⁴  · Christos Sakaridis¹  · Luc Van Gool¹ 

Received: date / Accepted: date

摘要 伪装场景理解 (Concealed Scene Understanding, CSU), 旨在对伪装物体进行感知, 是一个热门的计算机视觉课题。目前该领域内先进技术和新颖应用蓬勃发展, 亟需一个新的综述, 以方便研究人员能够了解领域内的整体概况, 包括当前进展和仍存在的挑战。本文主要有以下四点贡献: (1) 本文是首篇面向伪装场景理解的深度学习技术综述, 包括不同任务分类总结、各任务所面临的挑战以及该领域的发展情况。(2) 为了对最先进的技术进行权威地量化, 本文为伪装目标分割 (Concealed Object Segmentation, COS) 提供了最大且最新的基准评测。(3) 为了评估 CSU 领域深度学习方法在实际应用中的可迁移性, 本文收集了当前规模最大的隐蔽缺陷分割数据集 (CDS2K), 包含了从多样化的工业场景中挑选出的困难案例, 并在此基础上构建了一个全面的基准评测。(4) 讨论了 CSU 社区的开放问题以及潜在的研究方向。¹

Keywords 伪装场景理解 · 分割 · 检测 · 综述 · 导论 · 分类学 · 深度学习 · 机器学习

¹ CVL, ETH Zurich, Zurich 8092, Switzerland.

² CECC, ANU, Canberra 0200, Australia.

³ EE, Tsinghua University, Beijing 100084, China.

⁴ CS, Nankai University, Tianjin 300350, China.

¹ 本文的代码和数据集公开在 <https://github.com/DengPingFan/CSU>, 相关内容将持续更新, 以观察和总结该领域的最新进展情况。本文是论文 [1] 的中文翻译稿, 译者: 赵攀诚, 并由季葛鹏、徐鹏和范登平进行校稿。

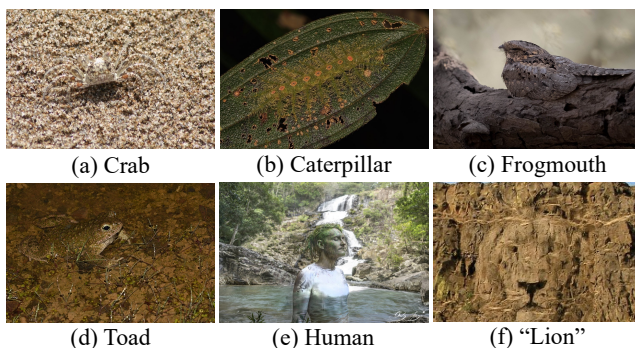


图 1 伪装样例展示。样例 (a-d) 展示了选自 [2] 的动物在他们的自然栖息地的图像。样例 (e) 展示了通过艺术手法伪装起来的人 [3]。样例 (f) 展示了合成的“狮子”图像 [4]。

1 引言

伪装场景理解 (Concealed Scene Understanding, CSU) 旨在识别表现出伪装模式的物体。就其本质而言, 相较于传统的目标检测 [5, 6], CSU 显然更具挑战性。CSU 有着非常广泛的现实应用, 包括搜索和救援工作、发现稀有物种、医疗健康 (例如自动化诊断肠道息肉 [7] 以及肺部病变 [8])、农业 (例如害虫检测 [9] 和水果成熟度检测 [10]) 以及内容创作 (例如娱乐艺术 [11]) 等等。在过去的十年中, 学术界和工业界广泛研究了 CSU 任务, 并且使用传统的计算机视觉和模式识别技术处理了各种具有伪装物体的图像, 包括手工设计的模式 (例如运动线索 [12, 13]、光流 [14, 15])、启发式先验知识 (例如颜色 [16]、纹理 [17]、光照强度 [18, 19]) 以及各种组合类的技术 [20–22]。

近年来, 由于相关的基准数据集 (例如 COD10K [2, 23] 和 NC4K [24]) 的出现以及深度学习迅猛发展,

该领域取得了重要的进展。2020 年, Fan 等人 [2] 发布了第一个大规模公开数据集 COD10K, 旨在推动视觉感知任务向解决伪装问题的方向发展。这也激发了其他相关学科的研究。例如, Mei 等人 [25, 26] 提出了一个基于分心感知来分割伪装物体的框架, 并且该框架可以扩展到识别自然场景中的透明材料任务 [27] 中。2023 年, Ji 等人 [28] 开发了一个高效模型来从目标级别的梯度信息中学习纹理特征, 该模型的可迁移性已经在多种下游任务中得到验证, 包括医学息肉分割以及道路裂缝检测。

尽管许多研究团队已经解决了许多与伪装物体相关的任务, 但本文相信, 进一步加强现有工作间的交互是有益的。因此, 本文主要回顾了 CSU 的现状以及最近基于深度学习的进展。同时提供了一个大规模隐蔽缺陷分割数据集 CDS2K。该数据集包含了从多样化的工业场景中挑选出的困难案例, 能够为 CSU 提供一个有效的基准评测。

现有综述及其范围。 本文的调研结果显示 CSU 社区中至今仅有几篇综述, 主要回顾非深度学习技术 [29, 30], 另外有一些基准评测 [31, 32] 的调研范围相对较窄, 例如聚焦于图像级别的分割任务, 仅评估了少量的深度学习方法。本文面向 CSU 领域中的深度学习技术提供了一个综合性调研, 具有更广的调研范围, 同时贡献了更为全面且面向实际应用的评测基准。

贡献。 本文的贡献总结如下: (1) 本文首先针对 CSU 领域中的深度学习技术进行调研, 包括任务分类及其特定的挑战问题。针对现有数据集和技术方案进行对比分析后, 综述了该方向在深度学习时代中的发展情况。(2) 伪装目标分割 (Concealed Object Segmentation, COS) 是 CSU 领域中一个重要的子任务, 且已经取得了很大进展。为了定量评估当前前沿技术方案, 本文为 COS 任务创建了包含最新方法且评测角度最全面的基准评测。(3) 为了评估深度 CSU 模型应用在实际场景中的性能, 本文重新构建了一个规模最大的伪装缺陷分割数据集 CDS2K, 该数据集将来自各种工业环境具有挑战性的图像样例纳入其中, 并基于此数据集创建了综合性评测基准。(4) 本文深入讨论了 CSU 社区当前研究面临的挑战和未来潜在的研究方向。

2 背景

2.1 任务分类和公式

2.1.1 图像级别的 CSU

本节将介绍五种常用的图像级别 CSU 任务, 它们可以用一个统一的映射函数 $\mathcal{F}: \mathbf{X} \mapsto \mathbf{Y}$ 表示, 该函数将输入空间 \mathbf{X} 转换到目标空间 \mathbf{Y} 。

- **伪装目标分割 (Concealed Object Segmentation, COS)** [23, 28] 是一项与类别无关的密集预测任务, 对未知类别的伪装区域或物体进行分割。如图 2 (a) 所示, 模型 $\mathcal{F}_{\text{COS}}: \mathbf{X} \mapsto \mathbf{Y}$ 表示在二值掩码 \mathbf{Y} 的监督下, 为图像 \mathbf{X} 的每个像素 \mathbf{x} 预测一个概率 $\mathbf{p} \in [0, 1]$ 。 \mathbf{p} 表示像素 \mathbf{x} 属于伪装区域的置信度。

- **伪装目标定位 (Concealed Object Localization, COL)** [24, 33] 旨在识别伪装目标最明显的区域, 这与人类的感知心理学 [33] 是一致的。该任务学习一个密集映射函数 $\mathcal{F}_{\text{COL}}: \mathbf{X} \mapsto \mathbf{Y}$ 。如图 2 (b) 所示, 输出 \mathbf{Y} 是一个由眼球追踪设备捕获的非二值眼球注视图。其关键点在于, 对一个像素 \mathbf{x} 的预测概率 $\mathbf{p} \in [0, 1]$ 表示其伪装的程度大小。

- **伪装实例排序 (Concealed Instance Ranking, CIR)** [24, 33] 是根据伪装场景中不同实例的可检测性来进行排序, 伪装程度被作为这种排序的基础。CIR 任务的目标是学习一个从输入空间 \mathbf{X} 到伪装排序空间 \mathbf{Y} 的密集映射 $\mathcal{F}_{\text{CIR}}: \mathbf{X} \mapsto \mathbf{Y}$, 其中, \mathbf{Y} 表示每个实例的逐像素标注, 用于表示该实例相对应的排序级别。例如, 在图 2 (c) 中, 展示了三只具有不同伪装程度的蟾蜍, 它们的排序标签来自于文献 [24]。为实现这一项任务, 可以使用每个实例的排序标签来替换 Mask R-CNN [34] 等实例分割模型中的类别 ID 标签。

- **伪装实例分割 (Concealed Instance Segmentation, CIS)** [35, 36] 是一项根据语义特征来识别伪装场景中实例的技术。与一般场景中实例分割 [37, 38] 为每个实例分配一个类别标签不同, CIS 识别每个伪装目标的属性以更有效地区分不同的个体。为了实现这一目标, CIS 任务采用了一个映射函数 $\mathcal{F}_{\text{CIS}}: \mathbf{X} \mapsto \mathbf{Y}$,

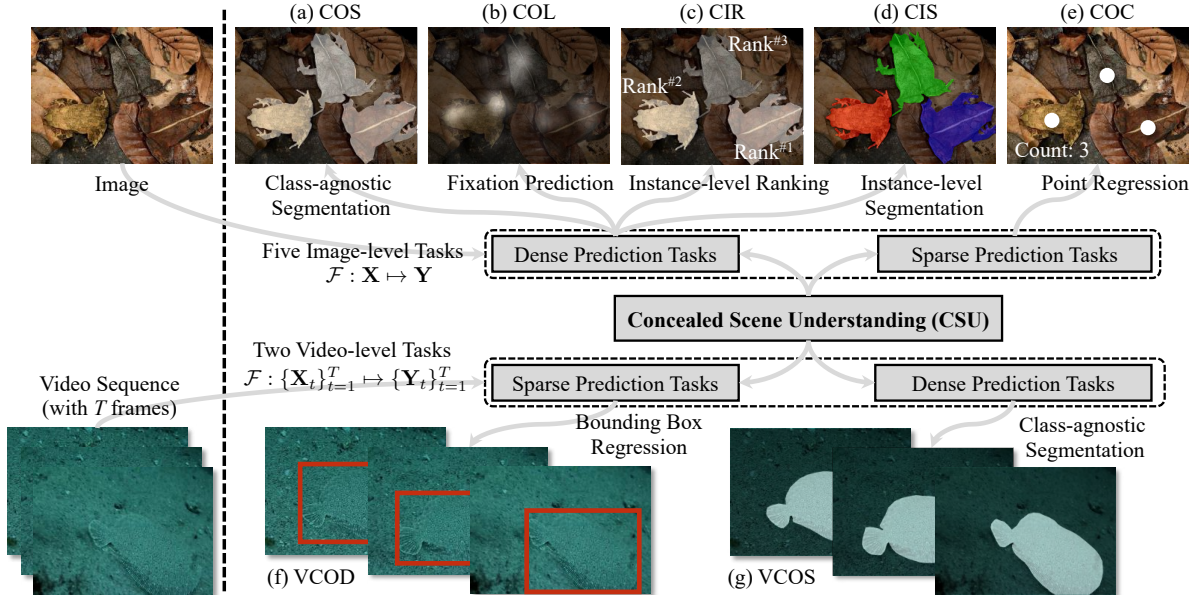


图 2 具有代表性的 CSU 任务的图示。其中包含 5 种图像级别的任务：(a) 伪装目标分割 (COS)、(b) 伪装目标定位 (COL)、(c) 伪装实例排序 (CIR)、(d) 伪装实例分割 (CIS) 以及 (e) 伪装目标计数 (COC)。另外 2 个是视频级别的任务：(f) 视频伪装目标检测 (VCOD) 以及 (g) 视频伪装目标分割 (VCOS)。该图针对每个任务特有的标注方式进行了可视化，详细解释请参见章节 2.1。

其中， \mathbf{Y} 是一个涵盖了不同实体的标量集合，用于解析图像中的每一个像素。图 2 (d) 解释了这一概念。

• **伪装目标计数 (Concealed Object Counting, COC)** [39] 旨在估计隐藏在周围环境中的实例的数量，是 CSU 领域的一个新任务。如图 2 (e) 所示，COC 估计每个实例的中心坐标并生成他们的数量。其可以被公式化表述为 $\mathcal{F}_{\text{COC}} : \mathbf{X} \mapsto \mathbf{Y}$ ，其中， \mathbf{X} 是输入图像， \mathbf{Y} 是模型输出，是一张指示场景中伪装实例的概率密度图。

总的来说，图像级别的 CSU 任务可以根据其语义信息被分为两组：对象级别 (COS 和 COL) 和实例级别 (CIR, COC 和 CIS)。对象级别的任务聚焦于如何感知对象，而实例级别的任务旨在识别语义以区分不同的实体。此外，根据其输出形式，COC 被视作稀疏型预测任务，其余任务则属于密集型预测任务。表 1 所回顾的文献显示，COS 任务已经得到了广泛的研究，至于其他三项任务的研究则正在逐步增加。

2.1.2 视频级别的 CSU

给定一个包含 T 个视频帧的伪装视频片段 $\{\mathbf{X}_t\}_{t=1}^T$ ，视频级别的 CSU 任务可以被表示为映射函数 \mathcal{F} ：

$\{\mathbf{X}_t\}_{t=1}^T \mapsto \{\mathbf{Y}_t\}_{t=1}^T$ ，用于解析密集空间中的空间-时间对应关系，其中， \mathbf{Y}_t 是视频帧 \mathbf{X}_t 所对应的标签。

• **视频伪装目标检测 (Video Concealed Object Detection, VCOD)** [40] 与视频目标检测任务 [41] 相似。该任务通过学习时空映射函数 $\mathcal{F}_{\text{VCOD}} : \{\mathbf{X}_t\}_{t=1}^T \mapsto \{\mathbf{Y}_t\}_{t=1}^T$ 来识别和定位视频中隐藏的目标。该映射预测每一帧 \mathbf{X}_t 中伪装目标的位置并输出为 \mathbf{Y}_t 。如图 2 (f) 所示，位置标签 \mathbf{Y}_t 以边界框的形式给出，每个边界框使用 4 个数字 (x, y, w, h) 表示目标的位置。其中， (x, y) 表示边界框左上角坐标， w 和 h 分别表示边界框的宽度和高度。

• **视频伪装目标分割 (Video Concealed Object Segmentation, VCOS)** [42] 源于伪装目标发现任务 [40]，其针对视频中的伪装目标进行分割。该任务通常利用时空线索来驱动模型学习一系列输入帧 \mathbf{X}_t 和每一帧对应分割掩码标注 \mathbf{Y}_t 之间的映射关系 $\mathcal{F}_{\text{VCOS}} : \{\mathbf{X}_t\}_{t=1}^T \mapsto \{\mathbf{Y}_t\}_{t=1}^T$ 。VCOS 任务的分割掩码可视化如图 2 (g) 所示。

总的来说，与图像级别 CSU 任务相比，视频级别 CSU 任务的发展相对缓慢，其主要原因是收集和标注视频数据耗时费力。但是，随着第一个大规模 VCOS 基准 MoCA-Mask [42] 的建立，这一领域取

得了根本性的进展,并展现出发展潜力,具有很大的探索空间。

2.1.3 任务关联

在图像级别的 CSU 任务中, CIR 任务对理解能力的需求最高,因为他不仅同时涉及到四个任务,例如,对区域进行像素级别的分割 (COS), 计数 (COC) 或者区分不同的实例 (CIS), 而且需要根据这些实例在不同难度下的被注视概率 (COL) 对其进行排序。此外,对于视频级别,由于分割任务需要模型提供像素级的分类概率,因此 VCOS 是 VCOD 的下游任务。

2.2 相关主题

接下来,本文将简要介绍显著目标检测 (Salient Object Detection, SOD)。与 COS 一样, SOD 任务需要提取目标对象的属性,但 SOD 任务侧重于显著性,而 COD 侧重于隐藏的属性。

• **图像级别的 SOD** 旨在识别图像中最吸引注意力的物体,并提取其精确的像素轮廓 [43]。在 SOD 领域中,已经探索了许多基于深度学习的网络架构,例如:多层感知机 [44–47]、全卷积网络 [48–52]、胶囊网络 [53–55]、基于 Transformer 架构的网络 [56] 以及混合网络 [57, 58]。同时,在 SOD 模型中也研究了不同的学习策略,包括:数据高效的方法 (例如:基于分类标签的弱监督方法 [59–63] 以及基于伪标签的无监督方法 [64–66])、多任务范式 (例如:目标计数 [67, 68]、眼动预测 [69, 70]、语义分割 [71, 72]、边缘检测 [73–77]、图像描述 [78]) 以及实例级范式 [79–82] 等。想全面了解这一领域发展概况的读者,可以参考关于以下任务的经典综述文章或者代表性研究,包括:视觉注意力 [83]、显著性预测 [84]、协同显著性检测 [85–87]、RGB SOD [5, 88–90]、RGB-D (深度, Depth) SOD [91, 92]、RGB-T (热力图, Thermal) SOD [93, 94] 以及光场 SOD [95]。

• **视频级别的 SOD**。视频显著目标检测 (Video Salient Object Detection, VSOD) 早期起源于在视频目标分割 (Video Object Segmentation, VOS) 任

务中引入注意力机制。在该阶段,任务场景相对简单,视频中只包含一个移动的物体。由于移动的物体往往会吸引视觉注意力,所以此阶段的 VOS 和 VSOD 是等价任务。例如, Wang 等人 [96] 使用一个全卷积神经网络来解决 VSOD 任务。随着 VOS 技术的发展,研究人员引入了更复杂的场景 (例如:复杂背景、物体移动以及两个目标),这个阶段下两个任务仍然是等价的。随后的工作利用了语义级别的时空特征 [97–100]、循环神经网络 [101, 102] 或者光流 [101, 103–105] 等离线运动线索。但是,随着更具挑战性的视频场景引入 (包括三个及以上的物体、同步传输制摄像机以及物体移动), VOS 和 VSOD 不再是等价任务了。然而,研究人员继续把这两个任务视作等价,忽视了视频场景中多物体运动的视觉注意力分配问题,严重阻碍了该领域的发展。为了解决这个问题,2019 年, Fan 等人 [106] 引入了眼球追踪器来标记多目标运动场景中的视觉注意力变化,并首次强调了 VSOD 任务中的注意力转移问题,构造了第一个大规模的 VSOD 基准-DAVSOD² 以及一个基线模型 SSAV, 这推动 VSOD 任务进入了一个全新的发展阶段。

• **附注:** COS 和 SOD 是不同的任务,但是他们可以通过 CamDiff 方法 [107] 相互受益,这一点已经通过对抗学习 [108] 得到证明。二者的相关性促使了联合研究工作的开展,例如最近提出的二分图像分割 [109]。在章节 6 中,本文将继续探讨这些领域未来潜在研究方向。

3 CSU 领域的深度学习模型

本节系统地回顾了深度 CSU 方法的任务定义和数据类型,同时创建了一个 GitHub 仓库³用于提供该领域最新、最全面的信息。

² <https://github.com/DengPingFan/DAVSOD>

³ https://github.com/GewelsJI/SINet-V2/blob/main/AWESOME_COD_LIST.md

表 1 本文所回顾的图像级别 CSU 方法的基本特点。此总结概述了相关方法主要设计特点，包括：**架构设计 (Arc.)**：模型所使用的框架，可以是多流框架 (MSF)、自底向上 & 自顶向下框架 (BTF) 或者多分支框架 (BF)。**多线索 (M.C.)**：是否提供辅助线索。**监督级别 (S.L.)**：使用全监督 (★) 或弱监督 (◇) 方式学习。**任务级别 (T.L.)**：该方法所解决的任务，包括 COS (●)、CIS (○)、COC (□) 和多任务学习 (■)。N/A 表示源代码不可获取。关于上述基本特点更加详细的描述，请参见章节3.1中关于图像级别 CSU 模型的介绍。

#	Model	Pub.	Core Component	Arc.	M.C.	S.L.	T.L.	Code
1	ANet [3]	CVIU ₁₉	classification & segmentation streams	BF	✓	★	●	Link
2	SINet [2]	CVPR ₂₀	search and identification modules	BTF	-	★	●	Link
3	MirrorNet [110]	Access ₂₁	fuse input and mirror data streams	MSF	-	★	●	Link
4	DCE [111]	arXiv ₂₁	depth contribution exploration, confidence-aware loss	BF	✓	★	●	Link
5	D2CNet [112]	TIE ₂₁	dual-branch, dual-guidance & cross-refine	BTF	-	★	●	Link
6	C2FNet [113]	IJCAI ₂₁	context-aware cross-level fusion	BTF	-	★	●	Link
7	UR-COD [114]	MMA ₂₁	uncertainty of pseudo-edge labels	MSF	-	★	●	Link
8	TINet [115]	AAAI ₂₁	texture perception & feature interaction guidance	BF	✓	★	●	N/A
9	JSCOD [108]	CVPR ₂₁	uncertainty-aware adversarial learning	MSF	-	★	●	Link
10	LSR [24]	CVPR ₂₁	localize, segment, & rank objects simultaneously	BF	✓	★	■	Link
11	MGL [116]	CVPR ₂₁	mutual graph learning	BF	✓	★	●	Link
12	PFNet [25]	CVPR ₂₁	distraction mining, positioning and focus modules	BTF	-	★	●	Link
13	UGTR [117]	ICCV ₂₁	uncertainty-guided transformer reasoning	BF	✓	★	●	Link
14	BAS [118]	arXiv ₂₂	residual refinement module, hybrid loss	BTF	-	★	●	Link
15	OSFormer [35]	ECCV ₂₂	location-sensing transformer, coarse-to-fine fusion	BF	✓	★	○	Link
16	CFL [36]	TIP ₂₂	camouflage fusion learning	BF	✓	★	○	Link
17	NCHIT [119]	CVIU ₂₂	neighbor connection, hierarchical information transfer	BTF	-	★	●	N/A
18	DTC-Net [120]	TMM ₂₂	local bilinear & spatial coherence organization	BTF	-	★	●	N/A
19	C2FNet-V2 [121]	TCSVT ₂₂	context-aware cross-level fusion	BTF	-	★	●	Link
20	CubeNet [122]	PR ₂₂	encoder-decoder framework with \mathcal{X} -connection	BF	✓	★	●	Link
21	ERRNet [123]	PR ₂₂	selective edge aggregation, reversible re-calibration	BF	✓	★	●	Link
22	TPRNet [124]	TVCJ ₂₂	transformer-induced progressive refinement	BTF	-	★	●	Link
23	ANSA-Net [125]	IJCNN ₂₂	attention-based neighbor selective aggregation	BF	✓	★	●	N/A
24	BSANet [126]	AAAI ₂₂	boundary-guided separated attention	BF	✓	★	●	Link
25	FAPNet [127]	TIP ₂₂	boundary guidance, feature aggregation & propagation	BF	✓	★	●	Link
26	FindNet [128]	TIP ₂₂	boundary-and-texture cues (extension of [126])	BF	✓	★	●	N/A
27	PINet [129]	ICME ₂₂	cascaded decamouflage module, label reweighting	BTF	-	★	●	N/A
28	OCENet [130]	WACV ₂₂	online confidence estimation, dynamic uncertainty loss	BF	✓	★	●	Link
29	BGNet [131]	IJCAI ₂₂	edge-guidance feature & context aggregation modules	BF	✓	★	●	Link
30	PreyNet [132]	MM ₂₂	bidirectional bridging interaction, predator learning	BF	✓	★	●	Link
31	DTINet [133]	ICPR ₂₂	dual-task interactive transformer	BF	✓	★	●	Link
32	ZoomNet [134]	CVPR ₂₂	scale integration & hierarchical mixed-scale units	MSF	-	★	●	Link
33	FDNet [135]	CVPR ₂₂	frequency enhancement & high-order relation modules	MSF	-	★	●	N/A
34	SegMaR [136]	CVPR ₂₂	segment, magnify, reiterate in a iterative manner	BTF	-	★	●	Link
35	SINetV2 [23]	TPAMI ₂₂	neighbor connection decoder, group-reversal attention	BTF	-	★	●	Link
36	MGL-V2 [137]	TIP ₂₃	multi-source attention recovery (extension of [116])	BF	✓	★	●	Link
37	FBNet [138]	TMCCA ₂₃	frequency-aware context aggregation & attention	BTF	-	★	●	N/A
38	TANet [139]	TCSVT ₂₃	texture-aware refinement, boundary-consistency loss	BTF	-	★	●	N/A
39	LSR+ [33]	TCSVT ₂₃	triple task learning (extension of [24])	BF	✓	★	■	Link
40	SARNet [140]	TCSVT ₂₃	triple-stage refinement (search-amplify-recognize)	BTF	-	★	●	Link
41	MFFN [141]	WACV ₂₃	co-attention of multi-view, channel fusion unit	MSF	-	★	●	Link
42	CRNet [142]	AAAI ₂₃	feature-guided and consistency losses	MSF	-	◇	●	Link
43	HitNet [143]	AAAI ₂₃	high-resolution iterative feedback	BTF	-	★	●	Link
44	DGNet [28]	MIR ₂₃	gradient-based texture learning, efficient network	BF	✓	★	●	Link
45	FSPNet [144]	CVPR ₂₃	feature shrinkage pyramid with transformer	BTF	-	★	●	Link
46	FEDER [145]	CVPR ₂₃	deep wavelet-like decomposition	BTF	-	★	●	Link
47	DCNet [146]	CVPR ₂₃	pixel-level decoupling, instance-level suppression	BF	✓	★	○	Link
48	IOCFFormer [39]	CVPR ₂₃	unify density- and regression-based strategies	BF	✓	★	□	Link
49	PFNet+ [26]	SCIS ₂₃	extension of PFNet [25]	BTF	-	★	●	Link
50	DQnet [147]	arXiv ₂₃	cross-model detail querying, relation-based querying	MSF	-	★	●	Link
51	CamoFormer [148]	arXiv ₂₃	masked separable attention	BTF	-	★	●	Link
52	PopNet [149]	arXiv ₂₃	source-free depth, object pop-out prior	MSF	-	★	●	Link

3.1 图像级别的 CSU 模型

3.1.1 伪装目标分割

本节从网络架构和学习范式两个维度讨论了现有的伪装目标分割任务 (COS) 的解决方案。

本节回顾了现有的四种图像级别的 CSU 任务，包括：伪装目标分割 (COS)、伪装目标定位 (COL)、伪装实例排序 (CIR) 以及伪装实例分割 (CIS)。表1中总结了这些方法的主要特征。

• **网络架构。**一般来说，全卷积网络 (FCNs [150]) 是图像分割的标准解决方案，因为它可以接受一个可变大小的输入并进行单次前馈传播。因此，基于 FCN 架构的模型在 COS 的解决方案中占主导地位

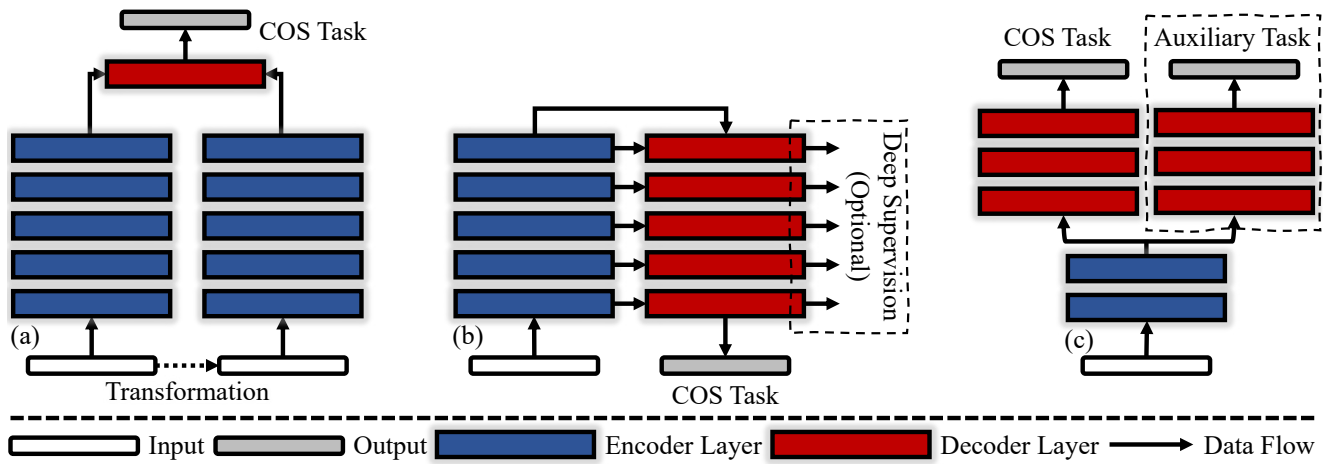


图 3 用于 COS 任务的网络架构概览。从左到右分别介绍了四种类型的框架：(a) 多流框架、(b) 自底向上/自顶向下框架及其带有深度监督的变体（可选项）以及 (c) 分支网络框架。详情参见章节 3.1.1。

位，主要分为以下三类：**a) 多流框架**，包含多个输入流，用以显式地学习多源表征，如图 3 (a) 所示。MirrorNet [110] 受生物启发，率先尝试增加额外的数据流来打破伪装状态。最近的许多工作采用了多流框架来改善他们的结果，包括：提供伪深度图 [149]、伪边缘的不确定性 [114]、对抗性学习范式 [108]、频域增强 [135]、多尺度输入 [134] 或多视角输入 [141]、多个主干骨干网络 [147]。与其他方法的监督设置不同，CRNet [142] 是唯一使用涂鸦标签作为监督的弱监督框架。这种方法有助于缓解有限标注数据导致的过拟合问题。**b) 自底向上和自顶向下框架**，在单次前馈过程中使用较深的特征来逐步增强较浅的特征，如图 3 (b) 所示。例如，C2FNet [113] 采用了这种设计，从粗到细地优化伪装的特征。此外，SegMaR [136] 基于该策略，使用子网络迭代优化的主网络。此外，其它相关研究 [2, 23, 25, 26, 112, 118–121, 124, 125, 129, 138–140, 143, 144, 148] 利用该框架对中间特征层进行深度监督策略 [151, 152]。这种做法也被特征金字塔网络 [153] 所采用，即以密集形式进行自顶向下和自底向上的特征传播，结合了更全面的多层次上下文特征，同时在最终预测之前引入额外的监督信号，这能够为更深层特征提供更可靠的监督指导。**c) 分支框架**，是一个单输入多输出的架构，由分割和辅助任务分支共同组成，如图 3 (c) 所示。值得注意的是，分支架构的分割部分可能与以前的一些框架有一些重叠，如单流 [3] 以及自底向上和自顶向下 [24, 28, 33, 108, 111, 115–117, 122, 123, 125–128, 130–133, 137] 架

构。例如，ERRNet [123] 和 FAPNet [127] 是伪装目标及其边界联合学习的典型案例。由于这些分支框架与多任务学习范式密切相关，接下来章节中将提供更多细节。

• **学习范式**。本节将讨论 COS 任务的两种常见学习范式：单任务和多任务学习。**a) 单任务学习**是 COS 任务中最常用的范式，它只涉及到伪装目标的分割任务。基于这一范式，目前大多数工作设计基于注意力的模块来助力目标区域的识别，例如文献 [2, 23, 121]。**b) 多任务学习**额外引入了一个辅助任务来协调或补充分割任务，使得学习过程更加鲁棒可靠。这些多任务框架可通过以下方式实现：进行置信度估计 [108, 117, 130, 132]、定位或排序 [24, 33]、类别预测 [3] 等辅助任务，或者学习伪装目标的深度线索 [111, 149]、边界线索 [116, 122, 123, 126, 127, 131] 以及纹理线索 [28, 115] 等。

3.1.2 伪装实例排序

对这一主题展开研究的工作有限，Lv 等人 [24] 首次观察到现有的 COS 方法不能量化伪装的难度等级。并针对该问题，使用眼球追踪仪创建了一个新数据集，CAM-LDR [33]，其中包含了实例分割的掩码标注、眼动点掩码以及排序标注。此外，还提出了两个统一的框架 LSR [24] 和其拓展版 LSR+ [33]，用于同时学习定位、分割、排序三个任务。其背后的见解是，具有鉴别性的定位区域可以指导全局范围内的

表 2 本文所回顾的基于视频的 CSU 方法的基本特点。 光流 (O.F.) : 是否预生成光流图。监督级别 (S.L.) : 使用真实数据 (★) 或者合成数据 (♣) 进行全监督学习, 以及使用自监督学习 (♡)。任务级别 (T.L.) : 视频伪装目标检测 (Δ) 以及分割 (▲)。更多细节请参见章节 3.2。

#	Model	Pub.	Core Components	O.F.	S.L.	T.L.	Project
1	FMC [154]	CVPR ₁₉	pixel trajectory recurrent neural network and clustering	✓	★	▲	N/A
2	VRS [40]	ACCV ₂₀	video registration and motion segmentation network	✓	★	Δ	Link
3	SIMO [155]	BMVC ₂₁	dual-head architecture, synthetic dataset	✓	♣	Δ	Link
4	MG [156]	ICCV ₂₁	self-supervised motion grouping	✓	♡	Δ	Link
5	RCF [157]	arXiv ₂₂	rotation-compensated flow, camera motion estimation	✓	★	Δ	N/A
6	OLCR [158]	NeurIPS ₂₂	object-centric layered representation, synthetic dataset	✓	♣	Δ	N/A
7	OFS [159]	TPAMI ₂₂	expectation-maximization method, motion augmentation	✓	♡	Δ	Link
8	QSDI [160]	CVPR ₂₂	quantifying the static and dynamic biases	✓	★	Δ	Link
9	SLTNet [42]	CVPR ₂₂	implicit motion handling, short- and long-term modules	-	★	▲	Link

伪装目标分割任务。进而, 不同伪装目标的可检出性可以通过排序任务进行推断。

3.1.3 伪装实例分割

这项任务将 COS 任务从区域级别推进到实例级别, 相对于 COS 而言是一个比较新的领域。Le 等人 [36] 通过扩充现有数据集 CAMO [3], 建立了一个全新的 CIS 任务基准 CAMO++。此外, 他们还提出了一种伪装融合学习策略, 通过学习图像上下文信息来微调现有的实例分割模型 (例如 Mask R-CNN 模型 [34])。近来, 在 COD10K [2] 以及 NC4K [24] 等伪装实例分割基准的基础上, Pei 等 [35] 通过引入位置感应 Transformer 以及由粗到细的融合这两个关键设计, 形成该领域第一个单阶段 Transformer 框架 OSFormer。最近, Luo 等人 [146] 提出用像素级伪装解耦模块和实例级伪装抑制模块这两种设计来分割伪装实例。

3.1.4 伪装目标计数

最近, Sun 等人 [39] 为社区引入了一个新的挑战, 即针对不可辨认目标的计数问题 (Indiscernible Object Counting, IOC), 来统计那些难以与周围环境区分开目标的数量。为了解决这一领域缺乏合适的数据集的问题, 还建立一个大规模数据集 IOCfish5K, 包括许多难以辨认的目标 (主要类别是鱼) 的水下场景高分辨率图像和密集型标注。并提出了一个名为 IOCFormer 的基线模型, 在一个统一的框架中整合了基于密度和基于回归的方法。

基于上述总结, COS 任务正在经历一个快速发展期, 每年都涌现出许多最新发表的论文。然而, 针

对 COL、CLR 以及 CIS 任务提出的解决方案仍然非常少, 表明这些领域仍然没有得到充分的探索, 为未来研究提供了巨大的空间。值得注意的是, 现有的许多研究工作都可以作为参考 (例如: 显著性预测 [84]、显著目标计数 [68] 以及显著实例分割 [82]), 为从伪装角度理解这些任务提供了一个扎实的基础。

3.2 视频级别的 CSU 模型

视频级别 CSU 任务有两个分支, 包括检测和分割视频中的伪装物体, 详情请参考表 2

3.2.1 视频伪装目标检测

由于逐像素标注的稀缺性, 大多数工作 [156, 158] 将这一主题表述为分割任务的一个退化问题。这些工作通常在 DAVIS [161]、FBMS [162] 等基于分割的数据集上训练, 然后在视频伪装目标检测数据集 MoCA [40] 上测试模型的泛化性。这些方法一致选择离线的方式提取光流作为分割任务的运动信息指导, 但在学习策略上表现更多样化, 如: 使用真实的 [40, 157, 160] 或生成的 [155, 158] 数据进行全监督学习, 以及使用自监督学习 [156, 159]。

3.2.2 视频伪装目标分割

Xie 等人 [154] 提出了第一个发现视频中伪装目标的工作, 即: 使用一个像素轨迹递归神经网络来聚集前景运动线索进行分割任务。然而, 这项工作受限于一个小规模的数据集 CAD [163]。Cheng 等人 [42] 通过创建一个具有像素级掩码的大规模 VCOS 基准数据集集合 MoCA-Mask 来推动这一领域, 还引入了

一个具有两阶段的基线模型 SLTNet 来隐式地利用运动信息。

从上面的回顾来看,目前 VCOS 任务的方法仍处于发展初期。在发展良好的视频分割领域中,已经有几个同期工作(例如自监督关联学习 [164–168]、为不同基于运动任务所设计的统一框架 [169–171]),为进一步的探索指明了方向。此外,高级语义理解还有一个值得补充的研究空白,如伪装场景中的语义分割和实例分割。

4 CSU 数据集

近年来,研究者 CSU 任务收集了许多图像和视频级别的数据集。在表3中,本文总结了具有代表性的数据集的特点。

4.1 图像数据集

- **CAMO-COCO** [3] 是为 COS 任务而收集的,包括来自 8 个类别的 2,500 张图像样本,被划分为带有伪装物体的 CAMO 数据集和带有非伪装物体的 MS-COCO 数据集两部分。CAMO 和 MS-COCO 均包含 1,250 张图像,其中 1,000 张用于训练,剩余 250 张用于测试。

- **NC4K** [24] 是目前评估 COS 模型的最大测试集合。NC4K 由来源于互联网的 4,121 张伪装图像组成,被分为两个大类别:自然场景和人工场景。除了图像之外,这个数据集还提供了包括对象级分割和实例级掩码的定位标签,使其成为该领域研究的宝贵资源。在 Lv 等人 [24] 最近的一项研究中,利用眼球追踪器来收集每个图像的眼动点信息。作者根据收集到的图像创建了一个由 2,280 张图像组成的 CAM-FR 数据集,其中 2,000 张图像用于训练,280 张用于测试。该数据集有三种类型的标签,包括:定位标签、排序标签以及实例级分割标签。

- **CAMO++** [36] 是一个新发布的数据集,包含 5,500 个图像样本和层次化的逐像素标注。数据集被分为两个部分:伪装样本包含 1,700 张训练图像和

1,000 张测试图像,非伪装样本包含 1,800 张训练图像和 1,000 张测试图像。

- **COD10K** [2,23] 是目前规模最大的数据集,具有多样化的伪装场景。该数据集包含来自多个公开获取摄影网站中收集到的 10,000 张图片,涵盖 10 个父类别和 78 个子类别。在这些图片中包括 5,066 张伪装图像和 1,934 张非伪装图像,剩余 3,000 张是背景图片。COD10K 的伪装子集使用不同的标签,如类别标签、边界框、目标级掩码和实例级掩码进行标注,提供了一个多样化的标注集合。

- **CAM-LDR** [33] 包括 4,040 个训练样本和 2,026 个测试样本。这些样本选自常用的混合训练数据集,即:来自 CAMO 数据集的 1,000 张训练样本和来自 COD10K 数据集的 3,040 张训练样本;同时还提供了测试数据集,包含来自于 COD10K 数据集的 2,026 张测试样本。CAM-LDR 数据集是 NC4K 数据集 [24] 的扩展版本,包括四种类型的标注:定位标签、排序标签、目标级分割掩码和实例级分割掩码。排序标签可根据伪装难度分为六个等级:背景、简单、中等 1、中等 2、中等 3 和困难。

- **S-COD** [142] 是第一个在弱监督设置下专门为 COS 任务设计的数据集。该数据集包括 4,040 张训练图像,其中 3,040 个样本选自 COD10K 数据集,1,000 个样本来自 CAMO 数据集。这些样本使用涂鸦形式进行了重新标注,这些标注是根据观察者的第一印象所提供主要结构的粗略轮廓,并没有参考到原始的像素级别真实标注。

- **IOCFish5K** [39] 是一个独特的数据集,重点在于对伪装场景中鱼的实例进行计数。这个用于 COC 任务的数据集来源于 YouTube 网站中所收集的 5,637 张高分辨率图像,包含了 659,024 个目标中心点标注。该数据集被分为三个子集,其中 3,137 张图像用于训练,500 张用于验证,2,000 张用于测试。

附注:综上所述,三个数据集(CAMO、COD10K 和 NC4K)通常被用于评估伪装对象分割(COS)任务方法的基准,涉及到的实验设定在章节5.2中进行了描述。对于伪装实例分割(CIS)任务,可以使用两个

表 3 CSU 数据集的基本特点。 **Train/Test:** 训练集和测试集中样本的数量，例如：图像数据集中指图像数量，视频数据集中指帧的数量。**Task:** 数据集的数据类型。**N.Cam.:** 是否包含非伪装样本。**Cls.:** 是否提供分类标签。**B.Box:** 是否为检测任务提供标注框。**Obj./Ins.:** 是否为分割任务提供目标级别或实例级别的分割掩码。**Rank:** 是否为每个实例提供排序标签。**Scr.:** 是否提供涂鸦形式的弱标签。**Cou.:** 是否提供密集的物体计数标签。详情请参见章节4.1和章节4.2中的描述。

#	Dataset	Year	Pub.	Train	Test	Task	N.Cam.	Cls.	B.Box	Obj.	Ins.	Fix.	Rank	Scr.	Cou.	Website
1	CAD [163]	2016	ECCV	0	836	Video	-	✓	-	✓	-	-	-	-	-	Link
2	CAMO-COCO [3]	2019	CVIU	2000	500	Image	✓	-	-	✓	-	-	-	-	-	Link
3	MoCA [40]	2020	ACCV	0	37,250	Video	-	✓	✓	-	-	-	-	-	-	Link
4	NC4K [24]	2021	CVPR	0	4,121	Image	-	-	✓	✓	✓	-	-	-	-	Link
5	MoCA-Mask [42]	2022	CVPR	19,313	3,626	Video	-	✓	-	✓	-	-	-	-	-	Link
6	CAMO++ [36]	2022	TIP	3,500	2,000	Image	✓	-	✓	✓	✓	-	-	-	-	Link
7	COD10K [2,23]	2022	TPAMI	6,000	4,000	Image	-	✓	✓	✓	✓	-	-	-	-	Link
8	CAM-LDR [33]	2023	TCSVT	4,040	2,026	Image	-	-	-	✓	✓	✓	✓	-	-	Link
9	S-COD [142]	2023	AAAI	4,040	0	Image	-	-	-	-	-	-	-	✓	-	Link
10	IOCFish5K [39]	2023	CVPR	3,637	2,000	Image	-	✓	-	-	-	-	-	-	✓	Link

包含实例级分割掩码的数据集 (COD10K 和 NC4K)。CAM-LDR 数据集提供了从物理眼球追踪设备收集的眼动点信息和三种类型的注释，适用于计算机视觉中各种基于大脑启发的探索。此外，还有两个来自 CSU 的新数据集：S-COD，可以设计用于弱监督的 COS 方法；IOCFish5K，专注于对伪装场景中的物体进行计数。

4.2 视频数据集

- **CAD [163]** 是一个包含 9 个短视频片段和 836 帧的小型数据集。该数据集所使用的标注策略是稀疏的，即：每隔五帧对伪装目标进行一次标注。因此，该数据集有 191 个分割掩码可供使用。

- **MoCA [40]** 是一个来源于 YouTube 网站的综合视频数据库，旨在检测移动的伪装动物。由 67 个类别和 141 个视频片段所组成，具有 37,250 个高分辨率视频帧和 7,617 个实例的相应标注框。

- **MoCA-Mask [42]** 是 MoCA 数据集 [40] 的扩展，每隔五帧提供一次人工标注的分割掩码。MoCA-Mask 数据集被分为两个部分：由 71 个短视频片段组成的训练集 (19,313 个视频帧及其对应的 3,946 个分割掩码) 和包含 16 个短视频片段的验证集 (3,626 个视频帧及其对应的 745 个分割掩码)。对于未获得标注的视频帧，使用基于双向光流的策略 [172] 生成了对应的伪分割标签。

附注：MoCA 数据集是目前最大的伪装目标视频数据集，但是仅提供用于目标检测的标签。因此，社区内的研究人员 [156, 158] 通常通过将分割掩码转换为

检测边界框，来侧面评估一个训练好的分割模型的性能。最近，MoCA-Mask 数据集的引入为伪装场景中的视频分割带来了转机。尽管有这些进展，但数据注释的数量和质量仍然不足以构建可靠的视频模型，能够有效地处理复杂的伪装场景。

5 CSU 评测基准

本次调研结果显示，COS 任务的方法构建相对完善，并且提供了许多有竞争力的方法，因此本节的基准评测会建立在 COS 任务之上。下面几节将详细介绍评测指标 (章节5.1)、基准评测协议 (章节5.2)、定量分析 (章节5.3、章节5.4、章节5.5) 以及定性比较 (章节5.6)。

5.1 评估指标

如文献 [23] 中所建议的，有五个常用的指标⁴可以用作 COS 任务的评估。本文将预测的掩码 \mathbf{P} 与相同分辨率大小的真实掩码 \mathbf{G} 进行比较。

- **MAE** (平均绝对误差, \mathcal{M}) 是一个传统的逐像素级别度量指标，可以定义为：

$$\mathcal{M} = \frac{1}{W \times H} \sum_x \sum_y^H |\mathbf{P}(x, y) - \mathbf{G}(x, y)|, \quad (1)$$

其中， W 和 H 是真实掩码 \mathbf{G} 的宽和高， (x, y) 是 \mathbf{G} 中的像素坐标。

⁴ https://github.com/DengPingFan/CSU/tree/main/cos_eval_toolbox

表 4 CAMO [3] 测试集上的定量结果对比。本文从两个方面对方法进行分类, 包括: 使用基于卷积骨干网络 (例如: ResNet [173], Res2Net [174], EffNet [175] 以及 ConvNext [176]) 的方法和使用基于 Transformer 骨干网络 (例如: MiT [177], PVTv2 [178] 和 Swin [179]) 的方法。本文遵循原论文中预设的输入分辨率测试了两个用于衡量模型效率的指标, 即模型参数 (Para) 和乘积运算 (MACs)。此外, 还报告了 9 个评价指标, 排名前三的数值分别使用**红色**、**绿色**和**蓝色**标注显示, \uparrow/\downarrow 表示分数越高/越低越好。如果由于代码尚未公开问题而无法获取到对应的结果, 表格中用连字符 (-) 来表示。除单独说明, 在随后的表格中将遵循以上描述的符号设置。

Model	Pub/Year	Backbone	Input	Para.	MACs	$\mathcal{S}_\alpha \uparrow$	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$E_\phi^{ad} \uparrow$	$E_\phi^{mn} \uparrow$	$E_\phi^{mx} \uparrow$	$F_\beta^{ad} \uparrow$	$F_\beta^{mn} \uparrow$	$F_\beta^{mx} \uparrow$
• Convolution-based Backbone														
SINet [2]	CVPR ₂₀	ResNet-50	352 ²	48.95M	19.42G	0.745	0.644	0.092	0.825	0.804	0.829	0.712	0.702	0.708
D2CNet [112]	TIE ₂₁	Res2Net-50	320 ²	-	-	0.774	0.683	0.087	0.844	0.818	0.838	0.747	0.735	0.743
C2FNet [113]	IJCAI ₂₁	Res2Net-50	352 ²	28.41M	13.12G	0.796	0.719	0.080	0.865	0.854	0.864	0.764	0.762	0.771
TINet [115]	AAAI ₂₁	ResNet-50	352 ²	28.56M	8.58G	0.781	0.678	0.087	0.847	0.836	0.848	0.729	0.728	0.745
JSCOD [108]	CVPR ₂₁	ResNet-50	352 ²	121.63M	25.20G	0.800	0.728	0.073	0.872	0.859	0.873	0.779	0.772	0.779
LSR [24]	CVPR ₂₁	ResNet-50	352 ²	57.90M	25.21G	0.787	0.696	0.080	0.859	0.838	0.854	0.756	0.744	0.753
R-MGL [116]	CVPR ₂₁	ResNet-50	473 ²	67.64M	249.89G	0.775	0.673	0.088	0.848	0.812	0.842	0.738	0.726	0.740
S-MGL [116]	CVPR ₂₁	ResNet-50	473 ²	63.60M	236.60G	0.772	0.664	0.089	0.850	0.807	0.842	0.733	0.721	0.739
PFNet [25]	CVPR ₂₁	ResNet-50	416 ²	46.50M	26.54G	0.782	0.695	0.085	0.855	0.841	0.855	0.751	0.746	0.758
UGTR [117]	ICCV ₂₁	ResNet-50	473 ²	48.87M	127.12G	0.785	0.686	0.086	0.861	0.823	0.854	0.749	0.738	0.754
BAS [118]	arXiv ₂₁	ResNet-34	288 ²	87.06M	161.19G	0.749	0.646	0.096	0.808	0.796	0.808	0.696	0.692	0.703
NCHIT [119]	CVIU ₂₂	ResNet-50	288 ²	-	-	0.784	0.652	0.088	0.841	0.805	0.840	0.723	0.707	0.739
C2FNet-V2 [121]	TCSVT ₂₂	Res2Net-50	352 ²	44.94M	18.10G	0.799	0.730	0.077	0.869	0.859	0.869	0.777	0.770	0.779
CubeNet [122]	PR ₂₂	ResNet-50	352 ²	-	-	0.788	0.682	0.085	0.852	0.838	0.860	0.734	0.732	0.750
ERRNet [123]	PR ₂₂	ResNet-50	352 ²	69.76M	20.05G	0.779	0.679	0.085	0.855	0.842	0.858	0.731	0.729	0.742
TPRNet [124]	TVCJ ₂₂	Res2Net-50	352 ²	32.95M	12.98G	0.807	0.725	0.074	0.880	0.861	0.883	0.777	0.772	0.785
FAPNet [127]	TIP ₂₂	Res2Net-50	352 ²	29.52M	29.69G	0.815	0.734	0.076	0.877	0.865	0.880	0.776	0.776	0.792
BSANet [126]	AAAI ₂₂	Res2Net-50	384 ²	32.58M	29.70G	0.794	0.717	0.079	0.866	0.851	0.867	0.768	0.763	0.770
OCENet [130]	WACV ₂₂	ResNet-50	480 ²	60.31M	59.70G	0.802	0.723	0.080	0.866	0.852	0.865	0.776	0.766	0.777
BGNet [131]	IJCAI ₂₂	Res2Net-50	416 ²	79.85M	58.45G	0.812	0.749	0.073	0.876	0.870	0.882	0.786	0.789	0.799
PreyNet [132]	MM ₂₂	ResNet-50	448 ²	38.53M	58.10G	0.790	0.708	0.077	0.856	0.842	0.857	0.763	0.757	0.765
ZoomNet [134]	CVPR ₂₂	ResNet-50	384 ²	32.38M	95.50G	0.820	0.752	0.066	0.883	0.877	0.892	0.792	0.794	0.805
FDNet [135]	CVPR ₂₂	Res2Net-50	416 ²	-	-	0.841	0.775	0.063	0.901	0.895	0.908	0.803	0.807	0.826
SegMaR [136]	CVPR ₂₂	ResNet-50	352 ²	56.21M	33.63G	0.815	0.753	0.071	0.881	0.874	0.884	0.795	0.795	0.803
SINetV2 [23]	TPAMI ₂₂	Res2Net-50	352 ²	26.98M	12.28G	0.820	0.743	0.070	0.884	0.882	0.895	0.779	0.782	0.801
CamoFormer-C [148]	arXiv ₂₃	ConvNeXt-B	384 ²	96.69M	50.77G	0.859	0.812	0.050	0.919	0.913	0.920	0.842	0.842	0.855
CamoFormer-R [148]	arXiv ₂₃	ResNet-50	384 ²	54.25M	78.85G	0.816	0.712	0.076	0.863	0.874	0.916	0.735	0.745	0.813
PopNet [149]	arXiv ₂₃	Res2Net-50	512 ²	188.05M	154.88G	0.808	0.744	0.077	0.871	0.859	0.874	0.790	0.784	0.792
CRNet [142]	AAAI ₂₃	ResNet-50	320 ²	32.65M	11.83G	0.735	0.641	0.092	0.829	0.815	0.830	0.709	0.701	0.707
PFNet+ [26]	SCIS ₂₃	ResNet-50	480 ²	-	-	0.791	0.713	0.080	0.862	0.850	0.865	0.764	0.761	0.770
DGNet-S [28]	MIR ₂₃	EffNet-B1	352 ²	7.02M	2.77G	0.826	0.754	0.063	0.896	0.893	0.907	0.786	0.792	0.810
DGNet [28]	MIR ₂₃	EffNet-B4	352 ²	19.22M	1.20G	0.839	0.769	0.057	0.906	0.901	0.915	0.804	0.806	0.822
• Transformer-based Backbone														
DTINet [133]	ICPR ₂₂	MiT-B5	256 ²	266.33M	144.68G	0.856	0.796	0.050	0.918	0.916	0.927	0.821	0.823	0.843
CamoFormer-S [148]	arXiv ₂₃	Swin-B	384 ²	97.27M	64.13G	0.876	0.832	0.043	0.935	0.930	0.938	0.856	0.856	0.871
CamoFormer-P [148]	arXiv ₂₃	PVTv2-B4	384 ²	71.40M	39.74G	0.872	0.831	0.046	0.931	0.929	0.938	0.853	0.854	0.868
HitNet [143]	AAAI ₂₃	PVTv2-B2	704 ²	25.73M	55.95G	0.849	0.809	0.055	0.910	0.906	0.910	0.833	0.831	0.838

• **F-measure** (F 度量, F_β) 定义为:

$$F_\beta = \frac{(1 + \beta^2) \text{Precision} \times \text{Recall}}{\beta^2 \text{Precision} + \text{Recall}}, \quad (2)$$

其中, 根据 [90] 中的推荐, $\beta^2 = 0.3$ 用来强调准确率重要性大于召回率。另外两个指标的定义如下:

$$\text{Precision} = \frac{|\mathbf{P}(T) \cap \mathbf{G}|}{|\mathbf{P}(T)|}, \quad \text{Recall} = \frac{|\mathbf{P}(T) \cap \mathbf{G}|}{|\mathbf{G}|}, \quad (3)$$

其中, $\mathbf{P}(T)$ 是使用阈值 $T \in [0, 255]$ 对非二值的预测图 \mathbf{P} 进行阈值化处理后得到的一个二值掩码。 $|\cdot|$ 用于计算图内部的掩码的总面积。因此, 可以使用 0 到 255 之间变化的阈值将非二值预测掩码转换为一系列的二值掩码。对所有阈值进行迭代计算后, 得到了三个指标, 分别是 F 度量的最大值 (F_β^{mx})、平均值 (F_β^{mn}) 和自适应值 (F_β^{ad})。

• **Enhanced-alignment measure** (增强匹配指标, E_ϕ) [180,181] 是一个最近提出的二值前景评价指标, 它考虑了两个二值图像之间局部相似度和全局相似度。公式定义为:

$$E_\phi = \frac{1}{W \times H} \sum_x \sum_y^H \phi[\mathbf{P}(x, y), \mathbf{G}(x, y)], \quad (4)$$

其中, ϕ 是增强对齐矩阵。与 F_β 类似, 该指标也包括在所有阈值上计算的三个值, 即增强匹配指标的最大值 (E_ϕ^{mx})、平均值 (E_ϕ^{mn}) 以及自适应值 (E_ϕ^{ad})。

• **Structure measure** (结构度量, \mathcal{S}_α) [182,183] 用来衡量非二值预测图和真实掩码之间的结构相似度:

$$S_\alpha = (1 - \alpha)S_o(\mathbf{P}, \mathbf{G}) + \alpha S_r(\mathbf{P}, \mathbf{G}), \quad (5)$$

其中, α 用于平衡目标感知相似度 S_o 和区域感知相似度 S_r 。与原论文设定一致, 本文默认设置 $\alpha = 0.5$ 。

5.2 实验协议

根据 Fan 等人 [23] 建议, 所有参与基准测试的方法都在同一个混合数据集上进行训练, 包括 COD10K [2] 和 CAMO [3] 两个数据集的训练部分, 共计 4,040 个样本。然后在三个常用的基准数据集上进行模型评估: 2,026 个样本的 COD10K 测试集合 [2]、250 个样本的 CAMO 测试集合 [3]、4,121 个样本的 NC4K 测试集合 [24]。

5.3 关于 CAMO 的定量分析

如表4中所述, 在 CAMO 测试数据集 [3] 上使用各种指标评估了 36 种基于深度学习的模型。依据所使用的骨干网络, 可以将这些模型分为两组: 32 个基于卷积神经网络的和 4 个基于 Transformer 的。对于那些使用卷积骨干网络的模型, 本文观察到如下几个有趣的发现:

- CamoFormer-C [148] 借助于 ConvNeXt [176] 骨干网络, 在 CAMO 数据集上取得了最好的性能, 在一些评估指标上甚至部分超越了基于 Transformer 的方法, 如 S_α 指标的数值: 0.859 (CamoFormer-C 模型) 对比 0.856 (DTINet 模型 [133]) 对比 0.849 (HitNet 模型 [143])。然而, 以 ResNet-50 为骨干网络的 CamoFormer-R 模型 [148] 无法超越与之具有相同骨干网络的竞争对手, 例如: 使用多尺度缩放策略的 ZoomNet 模型 [134] 和使用迭代细化策略的 SegMaR 模型 [136]。
- 基于 Res2Net 为骨干网络的模型中, FDNet [135] 使用 416^2 大小的高分辨率输入, 在 CAMO 上取得了最佳性能。此外, SINetV2 模型 [23] 和 FAPNet 模型 [127] 也使用相同的骨干网络在 352^2 的小尺寸输入下也取得了令人满意的结果。
- DGNet [28] 是一个高效的模型, 尽管它只有 19.22M 参数量和 1.20G 计算成本, 但与 JSCOD 模型 [108]

(121.63M) 和 PopNet 模型 [149] (181.05M) 等大重量模型相比, DGNet 的性能也能排进前 3 名。该模型在性能和效率之间寻求了一个平衡, 这使它成为一个非常有前景的架构, 值得进一步探索其潜力。

- 有趣的是, CRNet 模型 [142] 作为一个弱监督模型, 与早期的全监督模型 SINet [2] 在性能方面旗鼓相当。这说明仍有足够的空间开发新模型, 来弥补数据高效学习方法 (例如自监督或半监督学习) 与全监督方法之间的差距。

此外, 基于 Transformer 的方法因其优异的长距离建模能力, 大大提升了模型性能。本文在 CAMO 测试数据集上测试了四个基于 Transformer 的模型, 得出了三个值得关注的发现:

- CamoFormer-S 模型 [148] 利用 Swin Transformer 增强了对伪装内容的分层建模能力, 在 CAMO 基准测试中取得了最优异的表现。本文观察到, 基于 PVT 的变体 CamoFormer-P 模型 [148] 在参数量较少的情况下 (71.40M (CamoFormer-P) 对比 97.27M (CamoFormer-R)), 取得了与 CamoFormer-S 相似的结果。
- DTINet 模型 [133] 是一个双流分支网络, 利用 SegFormer [177] 中的 MiT-B5 语义分割模型作为主干。尽管它有 266.33M 的模型参数量, 但由于难以在训练阶段平衡如此大重量的两个分支, 其性能表现并没有令人印象深刻。尽管如此, 这一尝试也打破了我们的成见, 激励我们可以研究语义分割模型在伪装场景中的可移植性。
- 本文还研究了输入图像的分辨率对不同模型性能的影响。HitNet 模型 [143] 使用大小为 704^2 的高分辨率图像, 这可以提高模型检测小目标的能力, 但代价是增加了计算成本。类似的, 基于卷积的方法, 如 ZoomNet [134] 通过采用多个不同分辨率的输入 (最大输入为 576^2) 来提高分割性能, 也取得了令人印象深刻的性能。但是, 不是所有模型都能从这种方法中受益。例如使用 480^2 分辨率图像作为输入的 PopNet 模型 [149] 在所有评价指标上都没有超过使用 352^2 分辨率输入的 SINetV2 模型 [23]。这一观察结果引发两个关键性的问题: 高分辨率图像是否应该用于伪装场景, 以及我们如何制定有效的学习策略来检测不同大小的伪装物体? 本文将提出这些问

题的潜在解决方案，并会在章节5.5中基于 COD10K 测试数据集提供一些有趣的分析。

5.4 关于 NC4K 的定量分析

与 CAMO 数据集相比，NC4K 数据集 [24] 的数据规模更大、样本多样性更丰富，因此结论可能会发生微妙的变化。表5介绍了在目前最大的 COS 测试数据集 NC4K 的定量结果，其中有 4,121 个样本。该基准测评包括 28 个基于卷积的方法和 4 个基于 Transformer 的方法。观察结果如下：

- CamoFormer-C [148] 在 NC4K 测试数据集上仍优于所有方法。但与章节5.3中描述的在 CAMO 上观察到的尴尬情况相反的是，基于 ResNet-50 的 CamoFormer-R [148] 在 NC4K 上的表现优于其他两个竞争对手（即：SegMaR [136] 和 ZoomNet [134]）。这些结果证实了 CamoFormer 解码器设计在将潜特征空间映射回预测空间方面的有效性，特别是处理那些更为复杂的场景时。
- DGNet 模型 [28] 在具有挑战性的 NC4K 数据集上表现平平，这可能是由于模型参数较少，导致其建模能力受限。然而，这个缺点让模型拥有了一个轻量级和简单的结构，更有利于后续模型的修改和拓展。
- 尽管 PopNet [149] 在小规模数据集 CMAO 上可能表现不佳，但在更具挑战性的 NC4K 数据集上展现了优势。这表明使用额外网络来生成深度图先验信息，有助于模型检测更具挑战性的样本。与基于 Res2Net-50 骨干网络的 SINetV2 模型相比，PopNet 模型被设计的更重量（即：188.05M 对比 26.98M），并且输入图像的分辨率更大（即：512² 对比 352²），但这些改变仅仅使 PopNet 模型在 E_{ϕ}^{nn} 指标的数值上提高了 0.6%。
- 关于 CamoFormer 模型 [148]，其两个变体模型之间存在明显的性能差异。具体来说，基于 Swin-B 骨干网络的 CamoFormer-S 模型性能稍逊一筹，而基于 PVTv2-B4 骨干网络的 CamoFormer-P 模型表现更好。

5.5 关于 COD10K 的定量分析

在表6中，展示了 36 个模型在具有多样化伪装图像案例的 COD10K 测试数据集上的性能比较，其中包括 32 个基于卷积的模型和 4 个基于 Transformer 的模型。

- CamoFormer-C 模型 [148] 因具有强大的骨干网络，仍是所有基于卷积的方案中表现最佳的。与其在 NC4K 上的表现类似，CamoFormer-R 模型 [148] 再次超越了具有相同骨干网络的强大竞争对手，例如：SegMaR 模型 [136] 和 ZoomNet 模型 [134]。
- 与它在 NC4K 数据集上的表现相似，PopNet [149] 在 COD10K 测试数据集上也同样取得了优异的表现，排名仅次于 CamoFormer-C 模型 [148]。本文认为，场景的深度先验知识在提高对伪装环境理解能力方面起着至关重要的作用。这一见解将会推进研究更为智能的方法来学习结构性先验，例如：将多任务学习或启发式方法纳入模型之中。
- 值得注意的是，HitNet 模型 [143] 在 COD10K 基准测试中取得了最高的性能，超过了使用更强的骨干网络（例如：Swin-B 和 PVTv2-B4 等）的模型。为了弄清该现象的起因，本文计算了三个测试数据集中所有图像样本分辨率的平均值，分别为：CAMO (W=693.89, H=564.22)、NC4K (W=709.19, H=529.61) 以及 COD10K (W=963.34, H=740.54)。统计结果显示，COD10K 测试集合的整体样本分辨率最高，这表明利用高分辨率输入或多尺度建模的模型将更容易从中受益。因此，在存在高分辨率图像的情况下，HitNet 是检测伪装物体的不错选择。

5.6 定性比较

本节从视觉上评估了当前表现最优的模型在具有挑战性和复杂的图像样本上的表现。本文选取了 10 组表现最好模型所预测的结果，包括：6 个基于卷积的模型（即：CamoFormer-C [148]、DGNet [28]、PopNet [149]、ZoomNet [134]、FDNet [135] 以及 SINetV2 [23]）、2 个基于 Transformer 的模型（即：CamoFormer-S [148] 和 HitNet [143]）以及 2 个其他类型的模型（包括最早的基线模型模型 SINet [2] 以及一个弱监督模型 CRNet [142]）。所有的样本都

表 5 NC4K [24] 测试集上的定量比较。

Model	Pub/Year	Backbone	$\mathcal{S}_\alpha \uparrow$	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$E_\phi^{ad} \uparrow$	$E_\phi^{mn} \uparrow$	$E_\phi^{mx} \uparrow$	$F_\beta^{ad} \uparrow$	$F_\beta^{mn} \uparrow$	$F_\beta^{mx} \uparrow$
• Convolution-based Backbone											
SINet [2]	CVPR ₂₀	ResNet-50	0.808	0.723	0.058	0.883	0.871	0.883	0.768	0.769	0.775
C2FNet [113]	IJCAI ₂₁	Res2Net-50	0.838	0.762	0.049	0.901	0.897	0.904	0.788	0.795	0.810
TINet [115]	AAAI ₂₁	ResNet-50	0.829	0.734	0.055	0.882	0.879	0.890	0.766	0.773	0.793
JSCOD [108]	CVPR ₂₀	ResNet-50	0.842	0.771	0.047	0.906	0.898	0.907	0.803	0.806	0.816
LSR [24]	CVPR ₂₁	ResNet-50	0.840	0.766	0.048	0.904	0.895	0.907	0.802	0.804	0.815
R-MGL [116]	CVPR ₂₁	ResNet-50	0.833	0.740	0.052	0.890	0.867	0.893	0.778	0.782	0.800
S-MGL [116]	CVPR ₂₁	ResNet-50	0.829	0.731	0.055	0.885	0.863	0.893	0.771	0.777	0.797
PFNet [25]	CVPR ₂₁	ResNet-50	0.829	0.745	0.053	0.884	0.887	0.898	0.779	0.784	0.799
UGTR [117]	ICCV ₂₁	ResNet-50	0.839	0.747	0.052	0.889	0.874	0.899	0.779	0.787	0.807
BAS [118]	arXiv ₂₁	ResNet-34	0.817	0.732	0.058	0.868	0.859	0.872	0.767	0.772	0.782
NCHIT [119]	CVIU ₂₂	ResNet-50	0.830	0.710	0.058	0.872	0.851	0.894	0.751	0.758	0.792
C2FNet-V2 [121]	TCSVT ₂₂	Res2Net-50	0.840	0.770	0.048	0.900	0.896	0.904	0.799	0.802	0.814
ERRNet [123]	PR ₂₂	ResNet-50	0.827	0.737	0.054	0.892	0.887	0.901	0.769	0.778	0.794
TPRNet [124]	TVCJ ₂₂	Res2Net-50	0.846	0.768	0.048	0.901	0.898	0.911	0.798	0.805	0.820
FAPNet [127]	TIP ₂₂	Res2Net-50	0.851	0.775	0.047	0.903	0.899	0.910	0.804	0.810	0.826
BSANet [126]	AAAI ₂₂	Res2Net-50	0.841	0.771	0.048	0.906	0.897	0.907	0.805	0.808	0.817
OCENet [130]	WACV ₂₂	ResNet-50	0.853	0.785	0.045	0.908	0.902	0.913	0.812	0.818	0.831
BGNet [131]	IJCAI ₂₂	Res2Net-50	0.851	0.788	0.044	0.911	0.907	0.916	0.813	0.820	0.833
PreyNet [132]	MM ₂₂	ResNet-50	0.834	0.763	0.050	0.899	0.887	0.899	0.805	0.803	0.811
ZoomNet [134]	CVPR ₂₂	ResNet-50	0.853	0.784	0.043	0.907	0.896	0.912	0.814	0.818	0.828
FDNet [135]	CVPR ₂₂	Res2Net-50	0.834	0.750	0.052	0.895	0.893	0.905	0.774	0.784	0.804
SegMaR [136]	CVPR ₂₂	ResNet-50	0.841	0.781	0.046	0.905	0.896	0.907	0.821	0.821	0.826
SINetV2 [23]	TPAMI ₂₂	Res2Net-50	0.847	0.770	0.048	0.901	0.903	0.914	0.792	0.805	0.823
CamoFormer-C [148]	arXiv ₂₃	ConvNeXt-B	0.883	0.834	0.032	0.937	0.933	0.940	0.851	0.857	0.870
CamoFormer-R [148]	arXiv ₂₃	ResNet-50	0.855	0.788	0.042	0.913	0.900	0.914	0.820	0.821	0.830
PopNet [149]	arXiv ₂₃	Res2Net-50	0.861	0.802	0.042	0.915	0.909	0.919	0.830	0.833	0.843
DGNet-S [28]	MIR ₂₃	EfficientNet-B1	0.845	0.764	0.047	0.902	0.902	0.913	0.789	0.799	0.819
DGNet [28]	MIR ₂₃	EfficientNet-B4	0.857	0.784	0.042	0.910	0.911	0.922	0.803	0.814	0.833
• Transformer-based Backbone											
DTINet [133]	ICPR ₂₂	MiT-B5	0.863	0.792	0.041	0.914	0.917	0.926	0.809	0.818	0.836
CamoFormer-S [148]	arXiv ₂₃	Swin-B	0.888	0.840	0.031	0.941	0.937	0.946	0.857	0.863	0.877
CamoFormer-P [148]	arXiv ₂₃	PVTv2-B4	0.892	0.847	0.030	0.941	0.939	0.946	0.863	0.868	0.880
HitNet [143]	AAAI ₂₃	PVTv2-B2	0.875	0.834	0.037	0.928	0.926	0.929	0.854	0.853	0.863

是根据七个不同细粒度的属性，从 COD10K 测试数据集集中挑选的。在图4中展现了上述方法预测结果的定性比较，并提供了几处有趣的发现。

- 多个对象的属性 (MO) 具有一定挑战性，因为当前表现最佳的模型存在较高的假阴性率。正如图4的第一列所示，十个模型中只有两个能够大致定位白色飞鸟，请观察图中真值掩码的红圈标注。这两个模型是 CamoFormer-S [148] 和 FDNet [135]，前者采用强大的基于 Transformer 编码器的方法，后者利用了基于频域的学习策略。
- 本文测试的模型可以通过精确定位目标的主体部分，来准确检测大物体 (BO)。然而，这些模型难以识别更细微的细节，例如图4中第二列被红色圆圈标识出的蟾蜍爪子。
- 小物体 (SO) 属性具有挑战性，因为它在图像中仅占据很小的区域，如 COD10K [2] 中所报告的，通常小于总像素的 10%。如图4中第三列所示，只有两个模型 (CamoFormer-S 和 CamoFormer-C [148]) 可以检测到远处地上躺着一只可爱的猫。出现这样的困难主要有两个原因：首先，模型很难将小物体与图

像中复杂的背景或其他不相关的物体区分开来；其次，由于低分辨率输入和下采样操作，造成了检测器可能会错过微小区域。

- 视野外 (OV) 属性指的是部分超出图像边界的对象，导致图像中物体显示不完整。为了解决这个问题，模型应该对伪装场景具有更好的整体理解。如图4第四列所示，CamoFormer-C 模型 [148] 和 FDNet 模型 [135] 都可以处理 OV 属性并保持物体完整性。然而，两个基于 Transformer 的模型未能做到这一点。这一观察启发研究人员探索更有效率的方法，例如在卷积框架内进行局部建模和跨域学习策略。
- 形状复杂 (SC) 属性表示一个物体包含的细窄的部分，例如动物的脚。在图4的第 5 列中，竹节虫的脚作为一个具体案例可以很好地展现了这种属性，因为其细长的特点，很难准确地预测。而 HitNet 模型 [143] 在输入高分辨率的图像时，能准确预测出用红圈标注出的竹节虫的脚部细节。
- 遮挡 (OC) 属性指的是物体部分被遮挡，这也是一般场景中常见的挑战 [184]。如图4所示，第 6 列显示的两只猫头鹰被铁丝网遮挡了部分区域，导致它

表 6 COD10K [2] 测试集上的定量比较。

Model	Pub/Year	Backbone	$\mathcal{S}_\alpha \uparrow$	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$E_\phi^{ad} \uparrow$	$E_\phi^{mn} \uparrow$	$E_\phi^{mx} \uparrow$	$F_\beta^{ad} \uparrow$	$F_\beta^{mn} \uparrow$	$F_\beta^{mx} \uparrow$
• Convolution-based Backbone											
SINet [2]	CVPR ₂₀	ResNet-50	0.776	0.631	0.043	0.867	0.864	0.874	0.667	0.679	0.691
D2CNet [112]	TIE ₂₁	ResNet-50	0.807	0.680	0.037	0.879	0.876	0.887	0.702	0.720	0.736
C2FNet [113]	IJCAI ₂₁	Res2Net-50	0.813	0.686	0.036	0.886	0.890	0.900	0.703	0.723	0.743
TINet [115]	AAAI ₂₁	ResNet-50	0.793	0.635	0.042	0.848	0.861	0.878	0.652	0.679	0.712
JSCOD [108]	CVPR ₂₀	ResNet-50	0.809	0.684	0.035	0.882	0.884	0.891	0.705	0.721	0.738
LSR [24]	CVPR ₂₁	ResNet-50	0.804	0.673	0.037	0.883	0.880	0.892	0.699	0.715	0.732
R-MGL [116]	CVPR ₂₁	ResNet-50	0.814	0.666	0.035	0.865	0.852	0.890	0.681	0.711	0.738
S-MGL [116]	CVPR ₂₁	ResNet-50	0.811	0.655	0.037	0.851	0.845	0.889	0.667	0.702	0.733
PFNet [25]	CVPR ₂₁	ResNet-50	0.800	0.660	0.040	0.868	0.877	0.890	0.676	0.701	0.725
UGTR [117]	ICCV ₂₁	ResNet-50	0.818	0.667	0.035	0.850	0.853	0.891	0.671	0.712	0.742
BAS [118]	arXiv ₂₁	ResNet-34	0.802	0.677	0.038	0.869	0.855	0.870	0.707	0.715	0.729
NCHIT [119]	CVIU ₂₂	ResNet-50	0.792	0.591	0.046	0.794	0.819	0.879	0.596	0.649	0.698
C2FNet-V2 [121]	TCSVT ₂₂	Res2Net-50	0.811	0.691	0.036	0.890	0.887	0.896	0.718	0.725	0.742
CubeNet [122]	PR ₂₂	ResNet-50	0.795	0.643	0.041	0.862	0.865	0.883	0.669	0.692	0.715
ERRNet [123]	PR ₂₂	ResNet-50	0.786	0.630	0.043	0.845	0.867	0.886	0.646	0.675	0.702
TPRNet [124]	TVCJ ₂₂	Res2Net-50	0.817	0.683	0.036	0.869	0.887	0.903	0.694	0.724	0.748
FAPNet [127]	TIP ₂₂	Res2Net-50	0.822	0.694	0.036	0.875	0.888	0.902	0.707	0.731	0.758
BSANet [126]	AAAI ₂₂	Res2Net-50	0.818	0.699	0.034	0.894	0.891	0.901	0.723	0.738	0.753
OCENet [130]	WACV ₂₂	ResNet-50	0.827	0.707	0.033	0.885	0.894	0.905	0.718	0.741	0.764
BGNet [131]	IJCAI ₂₂	Res2Net-50	0.831	0.722	0.033	0.902	0.901	0.911	0.739	0.753	0.774
PreyNet [132]	MM ₂₂	ResNet-50	0.813	0.697	0.034	0.894	0.881	0.891	0.731	0.736	0.747
ZoomNet [134]	CVPR ₂₂	ResNet-50	0.838	0.729	0.029	0.893	0.888	0.911	0.741	0.766	0.780
FDNet [135]	CVPR ₂₂	Res2Net-50	0.840	0.729	0.030	0.906	0.919	0.935	0.728	0.757	0.788
SegMaR [136]	CVPR ₂₂	ResNet-50	0.833	0.724	0.034	0.893	0.899	0.906	0.739	0.757	0.774
SINetV2 [23]	TPAMI ₂₂	Res2Net-50	0.815	0.680	0.037	0.864	0.887	0.906	0.682	0.718	0.752
CamoFormer-C [148]	arXiv ₂₃	ConvNeXt-B	0.860	0.770	0.024	0.926	0.926	0.935	0.778	0.798	0.818
CamoFormer-R [148]	arXiv ₂₃	ResNet-50	0.838	0.724	0.029	0.900	0.916	0.930	0.721	0.753	0.786
PopNet [149]	arXiv ₂₃	Res2Net-50	0.851	0.757	0.028	0.910	0.910	0.919	0.771	0.786	0.802
CRNet [142]	AAAI ₂₃	ResNet-50	0.733	0.576	0.049	0.845	0.832	0.845	0.637	0.633	0.636
PFNet+ [26]	Ssis ₂₃	ResNet-50	0.806	0.677	0.037	0.880	0.884	0.895	0.698	0.716	0.734
DGNet-S [28]	MIR ₂₃	EfficientNet-B1	0.810	0.672	0.036	0.869	0.888	0.905	0.680	0.710	0.743
DGNet [28]	MIR ₂₃	EfficientNet-B4	0.822	0.693	0.033	0.879	0.896	0.911	0.698	0.728	0.759
• Transformer-based Backbone											
DTINet [133]	ICPR ₂₂	MiT-B5	0.824	0.695	0.034	0.881	0.896	0.911	0.702	0.726	0.754
CamoFormer-S [148]	arXiv ₂₃	Swin-B	0.862	0.772	0.024	0.932	0.931	0.941	0.780	0.799	0.818
CamoFormer-P [148]	arXiv ₂₃	PVTv2-B4	0.869	0.786	0.023	0.931	0.932	0.939	0.794	0.811	0.829
HitNet [143]	AAAI ₂₃	PVTv2-B2	0.871	0.806	0.023	0.936	0.935	0.938	0.818	0.823	0.838

们的视觉区域被分隔开。不幸的是，现有的大多数模型都无法很好的处理这种情况。

- 不可定义的边界（IB）由于前景和背景难以区分，具有很大的不确定性，因此很难被解决。如图4的最后一列具有抠图级别的图像样本所示，模型均无法正确区分目标边界。

- 图4的最后两行展示了由 SINet 模型 [2] 产生的预测结果，这是最早的基线模型。目前的模型在定位精度、边界细节和其他方面都有了显著改善。此外，CRNet 模型 [142] 是一种由弱标签进行监督的方法，可以有效地定位目标物体并达到令人满意的标准。

6 讨论与展望

基于前文的文献回顾和实验分析，在本节中继续讨论 CSU 领域所面临的五个挑战以及潜在的相关发展方向。

- **高效标签学习**: 深度学习技术已经显著推动了 CSU 领域的发展。然而，传统的有监督深度学习需要大量数据和资源消耗。在实际场景中，我们希望模型能够在有限的资源下工作，并具有良好的可移植性。因此，开发有效的 CSU 任务学习策略是一个有前途的方向，例如 CRNet [142] 中使用弱监督策略。

- **域自适应**: 伪装样本通常是从自然场景中收集的。因此，在自动驾驶场景中部署模型以检测隐藏伪装物体极具挑战。近期实践表明，可以使用各种技术来缓解这个问题，例如域自适应 [185]、迁移学习 [186]、少样本学习 [187] 和元学习 [188]。

- **高保真的合成数据集**: 为了缓解算法偏见，增加数据的多样性和规模至关重要。人工智能内容生成（AIGC）[189] 和深度生成模型，如生成对抗网络 [190–192] 和扩散模型 [193, 194] 的快速发展，使得创建通用领域的合成数据更容易。最近，为了解决多模式训练图像稀缺性问题，Luo 等人 [107] 提出了

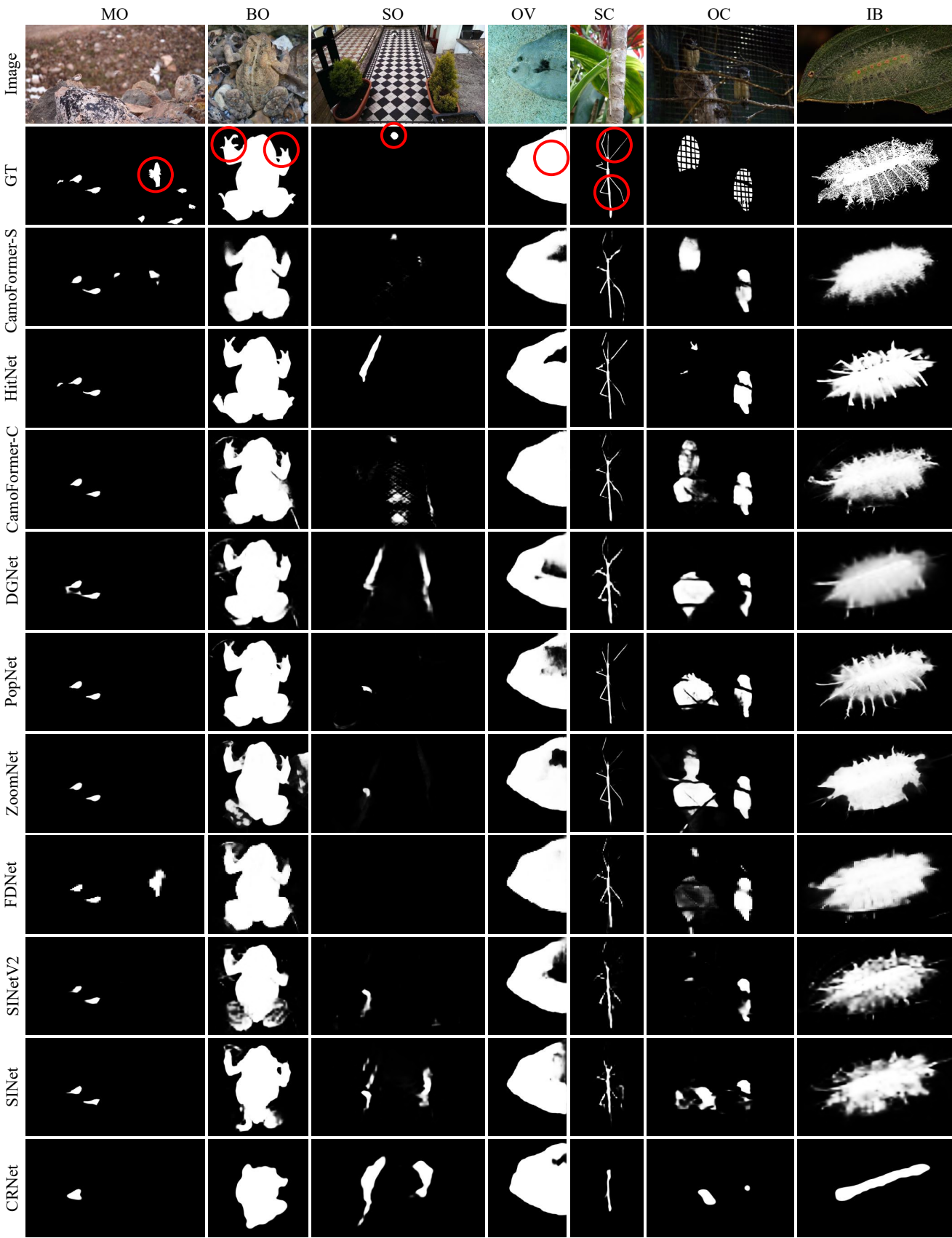


图 4 十种 COS 方法的定性结果对比。每一列中视觉属性的描述描述请参见章节5.6。

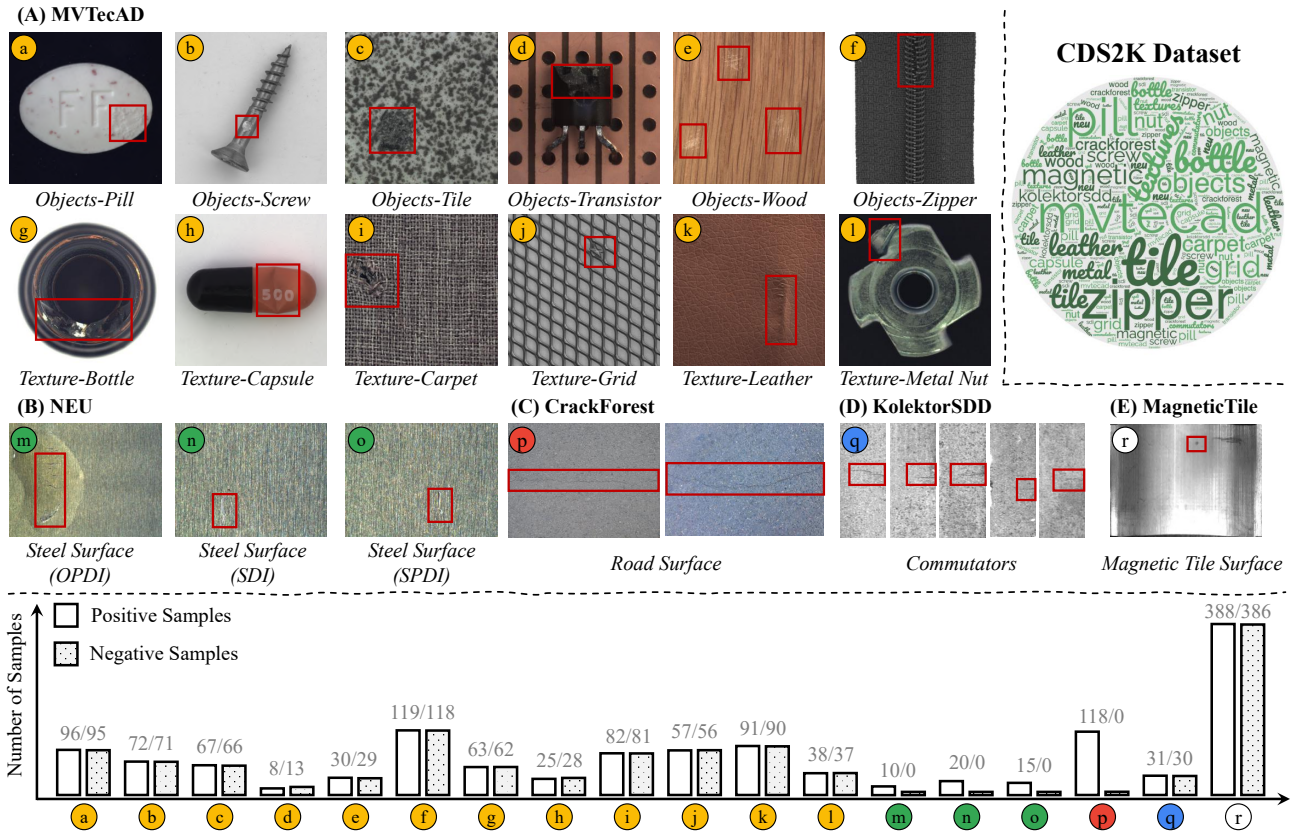


图 5 CDS2K 的样本图库。它收集自五个子数据库：(a-l) MVTecAD、(m-o) NEU、(p) CrackForest、(q) KolektorSDD 和 (r) MagneticTile。有缺陷的区域用红色矩形标出。右上角展示 CDS2K 的词云可视化结果。下图展示 CDS2K 中每个类别的阳性/阴性样本的统计数量。

一种基于扩散的图像生成框架，该框架在保留原始标签的同时，在伪装样本上生成显著对象。由此，在该数据集上训练的模型能够区分伪装目标和显著目标，以获得稳健的特征表示。

• **神经架构搜索：**自动网络架构搜索 (NAS) 是一个很有前景的研究方向，可以发现在给定任务上具有优越性能的最佳网络架构。在伪装环境中，NAS 可以发现更有效的网络架构来处理具有复杂背景的场景、高度可变的物体外观和有限数据标注。这可以开发出更高效和有效的网络架构，从而提高准确性和效率。将 NAS 与其他研究方向相结合，如域适应和数据高效学习，可以进一步加强对伪装场景的理解。这些探索方向具有显著推动技术水平不断提升的潜力，并值得未来进一步研究。

• **大模型和提示工程：**这个研究主题已经变得越来越受欢迎，甚至成为了自然语言处理社区的一个新

的方向。最近，“分割一切模型” (SAM) [195] 被提出，虽然它在几个伪装场景的无提示设置上存在局限性 [196]，但是仍为计算机视觉领域带有很大的革新。人们可以利用提示工程范式来简化工作流，使用训练有素的强大编码器和特定任务的适应性，例如特定任务的提示和多任务预测头。这种方法有望将成为计算机视觉社区未来的趋势。大型语言模型 (LLMs) 给 AI 带来了新机遇和挑战，并进一步推动了人工智能朝着通用人工智能迈进。然而，对于学术界来说训练资源消耗巨大的大模型是具有挑战性的。这可能会出现一个有前景的范式：最前沿的深度 CSU 模型被用作领域专家模型，同时大模型可以作为外部组件来提供辅助决策、表征等方式来协助专家模型。

表 7 CDS2K 数据集中正样本的数据统计。 对于一个给定的图像，区域比率由 $r = \text{缺陷像素} / \text{所有像素}$ 计算而得。值得注意的是，这里只计算五个子数据集的阳性样本数量。

Category		$0\% < r < 1\%$	$1\% \leq r < 10\%$	$10\% \leq r < 20\%$	$20\% \leq r < 30\%$	$30\% \leq r < 40\%$	$40\% \leq r < 50\%$	Total
MVTecAD	Objects-Pill	41	55	0	0	0	0	96
	Objects-Screw	71	1	0	0	0	0	72
	Objects-Tile	0	30	28	7	2	0	67
	Objects-Transistor	1	7	0	0	0	0	8
	Objects-Wood	2	26	2	0	0	0	30
	Objects-Zipper	16	102	1	0	0	0	119
	Texture-Bottle	3	39	20	1	0	0	63
	Texture-Capsule	17	8	0	0	0	0	25
	Texture-Carpet	37	45	0	0	0	0	82
	Texture-Grid	39	18	0	0	0	0	57
	Texture-Leather	70	21	0	0	0	0	91
	Texture-Metal Nut	6	31	1	0	0	0	38
NEU	OPDI	10	0	0	0	0	0	10
	SDI	20	0	0	0	0	0	20
	SPDI	15	0	0	0	0	0	15
CrackForest		28	90	0	0	0	0	118
KolektorSDD		31	0	0	0	0	0	31
Magnetic Tile Defect		216	70	27	27	24	24	388
Total		623	543	79	35	26	24	1330

表 8 对 CDS2K 的正样本进行定量比较。

Model	Pub/Year	Backbone	$S_{\alpha} \uparrow$	$F_{\beta}^w \uparrow$	$\mathcal{M} \downarrow$	$E_{\phi}^{ad} \uparrow$	$E_{\phi}^{mn} \uparrow$	$E_{\phi}^{mx} \uparrow$	$F_{\beta}^{ad} \uparrow$	$F_{\beta}^{mn} \uparrow$	$F_{\beta}^{mx} \uparrow$
SINetV2 [23]	TPAMI ₂₂	Res2Net-50	0.551	0.215	0.102	0.509	0.567	0.597	0.223	0.248	0.258
HitNet [143]	AAAI ₂₃	PVTv2-B2	0.563	0.276	0.118	0.574	0.564	0.570	0.298	0.298	0.299
DGNet [28]	MIR ₂₃	EfficientNet-B4	0.578	0.258	0.089	0.552	0.569	0.579	0.274	0.291	0.297
CamoFormer-P [148]	arXiv ₂₃	PVTv2-B4	0.589	0.298	0.100	0.590	0.588	0.596	0.330	0.329	0.339

7 缺陷分割数据集

工业缺陷通常源于不良的生产过程，例如机械冲击、工件摩擦、化学腐蚀和其他不可避免的物理因素，其外部视觉形式通常表现为特殊的模式或异常值，例如工业设备表面划痕、斑点、孔洞；织物表面色差、凹痕；材料表面杂质、断裂和污渍等。尽管以前的研究通过基于视觉技术（如分类 [197–199]、检测 [200–202] 和分割 [203–205]）取得了有希望的成果来识别视觉缺陷。这些技术基于缺陷容易被检测到这一假设，但它们忽略了那些“无缝”嵌入其材料环境中的具有挑战性的缺陷。为此，本文精心地收集了一个新的多场景基准，名为 CDS2K，用于伪装缺陷分割任务，其样本选自现有的工业缺陷数据库。

7.1 数据集组织

为了创建一个高质量的数据集，本文确立了三个数据选取原则：(a) 所选择的样本应包括至少一个有缺陷的区域，这将作为一个正样本。(b) 缺陷区域应具

有与背景相似的模式，使其难以识别。(c) 本文还选择正常样例作为负样本，以提供与正样本的对比。这些样本选自以下著名的缺陷分割数据库。

- MVTecAD⁵ [206, 207] 包含若干正负样本，用于无监督的异常检测。本文手动选择了 748 个正样本和 746 个负样本，这些样本有两个主要的伪装模式：(a) 目标类别，如图5的第一行：药丸、螺丝、瓷砖、晶体管、木材和拉链；(b) 纹理类别，如图5的第二行：瓶子、胶囊、地毯、网格、皮革和金属螺母。正/负样本的数量在图5中用黄色圆圈指示。

- NEU⁶ 提供了三个不同的数据库：油污缺陷图像 [208] (Oil Pollution Defect Images, OPDI)、斑点缺陷图像 [209] (Spot Defect Images, SDI) 和钢凹缺陷缺陷图像 [210] (Steel Pit Defect Images, SPDI)。如图5中第三行绿色圆圈标注的子图所示，本文从这些数据库中分别选择了 10、20 和 15 个正样本。

⁵ <https://www.mvtec.com/company/research/datasets/mvtec-ad>

⁶ http://faculty.neu.edu.cn/songkechen/zh_CN/zdylm/263270/list/index.htm

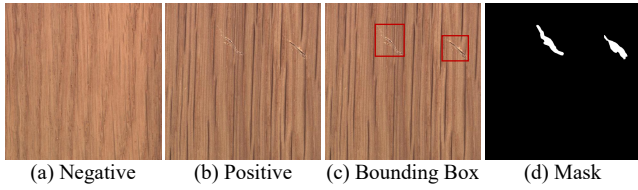


图 6 不同标注的可视化。从 MVTecAD 数据库中选择了一组图像，包括一个负样本 (a) 和一个正样本 (b)。提供了相应的标注：类别（木材上的划痕）、缺陷位置：边界框 (c) 和分割掩码 (d)。

- CrackForest⁷ [211, 212] 是一个具有密集型注释的道路裂缝图像数据库，用于城市道路表面的健康监测。本文从中选择了 118 个具有伪装性质的样本，这些样本显示在图5的第三行红圈标注的子图中。
- KolektorSDD⁸ [204] 是由 Kolektor 课题组收集和标注的，其中包含了一些来自真实世界案例中受控工业环境中的缺陷以及非缺陷表面图像。本文手动选择了 31 个正样本和 30 个负样本，详见图5的第三行蓝色圆圈标注的子图。
- Magnetic Tile Defect⁹ [213] 数据集包含六个常见的磁砖缺陷和对应的密集型标注。本文挑选了 388 个正样本和 386 个负样本，详见图5中白色圆圈标注的子图。

7.2 数据集描述

数据集 CDS2K 包括 2,492 个样本，由 1,330 个正样本和 1,162 个负样本组成。每个样本都包含三个不同的人工标注的标签：类别、边界框和逐像素的分割掩码。图6举例说明了这些不同的标注。每个类别的缺陷区域的平均比例详见表7，这表明大多数缺陷区域都相对较小。

7.3 对 CDS2K 数据集的评估

本节，评估了现阶段前沿的 COS 模型在 CDS2K 的正样本上的泛化能力。根据代码的可获取性，本文选择了四种表现最好的 COS 方法：SINetV2 [23]、DGNet [28]、CamoFormer-P [148] 以及 HitNet [143]。

正如在表8中报告的那样，这些模型在处理跨域样本时并不有效，这表明需要进一步探索自然场景和下游应用之间的域差异问题。

8 总结

本文旨在提供一个为伪装场景理解 (CSU) 量身定制的深度学习技术综述。为了帮助读者全面了解该领域的整体发展情况，本文做出了四项贡献：首先，本文提供了详细的 CSU 领域调研，包括其背景、分类介绍、特定任务下的技术挑战性以及在深度学习时代取得的进展。据我们所知，这项调查是目前此领域内最全面的综述。其次，本文创建了体量最大和方法最新的伪装目标分割 (COS) 的测评基准，该基准展示了对最先进技术的定量比较。第三，本文收集了相关数据集进而构建出了一个规模最大的伪装工业缺陷分割数据集 CDS2K，包含来自多种工业场景中的困难样例，并在此基础上构建了一个综合性基准评测来评估深度 CSU 模型方法在实际场景中的通用性。最后，本文讨论了这个社区的开放问题和潜在方向，旨在促进这一领域未来进一步的研究和发展。

本文得出以下几方面结论。(1) **模型**：最常见的思路是基于 UNet 架构进行模型设计，并添加注意力模块进行特征增强。考虑额外的先验信息和/或引入辅助任务参与训练，可以在一定程度上提高性能，这方面还有许多潜在的问题值得去探索；(2) **训练**：在 COS 任务中，全监督学习仍是当前伪装目标分割领域的主流策略，但很少有研究解决数据不足或标签不足条件下的学习问题。CRNet 模型 [142] 是解决这一问题的一次好的尝试。(3) **数据集**：现有的数据集规模不足且缺乏多样性，社区内也需要更多专业化领域的伪装样本，例如自动驾驶和临床诊断领域；(4) **性能**：基于 Transformer 架构和 ConvNext 架构的模型在性能上明显优于其他模型。计算代价和模型性能的权衡仍未得到充分研究。在这方面 DGNet 模型 [28] 是一个很好的尝试。(5) **评测指标**：当前没有专门适用于伪装领域的评价指标，以考虑不同数据样本的不同伪装程度，从而给出一个无偏的评价结果。

⁷ <https://github.com/cuilimeng/CrackForest-dataset>

⁸ <https://www.vicos.si/resources/kolektorsdd/>

⁹ <https://github.com/abin24/Magnetic-tile-defect-datasets>

现有的伪装场景理解技术方案侧重于利用外观属性（如颜色、纹理、边界）来识别伪装物体，而没有从语义的角度（例如物体之间的关系）进行充分的场景感知和输出。语义层面的认知是下一代伪装视觉感知的关键。因此在不久未来，伪装场景理解模型应该包含各种语义能力，包括整合高层次的语义、学习视觉语言知识 [214] 以及模拟物体间的交互关系。

我们希望这份综述能够为新加入此方向的研究人员提供一个详细的概述，也为相关专家提供一定的参考，并促进未来的研究。

参考文献

- Deng-Ping Fan, Ge-Peng Ji, Peng Xu, Ming-Ming Cheng, Christos Sakaridis, and Luc Van Gool. Advances in deep concealed scene understanding. *arXiv preprint:2304.11234*, 2023.
- Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *CVPR*, 2020.
- Trung-Nghia Le, Tam V Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabran network for camouflaged object segmentation. *CVIU*, 184:45–56, 2019.
- Qing Zhang, Gelin Yin, Yongwei Nie, and Wei-Shi Zheng. Deep camouflage images. In *AAAI*, 2020.
- Deng-Ping Fan, Jing Zhang, Gang Xu, Ming-Ming Cheng, and Ling Shao. Salient objects in clutter. *IEEE TPAMI*, 45(2):2344–2366, 2023.
- Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.
- Ge-Peng Ji, Guobao Xiao, Yu-Cheng Chou, Deng-Ping Fan, Kai Zhao, Geng Chen, and Luc Van Gool. Video polyp segmentation: A deep learning perspective. *MIR*, 19(6):531–549, 2022.
- Deng-Ping Fan, Tao Zhou, Ge-Peng Ji, Yi Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Infnet: Automatic covid-19 lung infection segmentation from ct images. *IEEE TMI*, 39(8):2626–2637, 2020.
- Liu Liu, Rujing Wang, Chengjun Xie, Po Yang, Fangyuan Wang, Sud Sudirman, and Wancai Liu. Pestnet: An end-to-end deep learning approach for large-scale multi-class pest detection and classification. *IEEE Access*, 7:45301–45312, 2019.
- Matteo Rizzo, Matteo Marcuzzo, Alessandro Zangari, Andrea Gasparetto, and Andrea Albarelli. Fruit ripeness classification: A survey. *AIA*, 7:44–57, 2023.
- Hung-Kuo Chu, Wei-Hsin Hsu, Niloy J Mitra, Daniel Cohen-Or, Tien-Tsin Wong, and Tong-Yee Lee. Camouflage images. *ACM TOG*, 29(4):51–1, 2010.
- Terrance E Boulton, Ross J Micheals, Xiang Gao, and Michael Eckmann. Into the woods: Visual surveillance of noncooperative and camouflaged targets in complex outdoor settings. *Proceedings of the IEEE*, 89(10):1382–1402, 2001.
- Donatello Conte, Pasquale Foggia, Gennaro Percannella, Francesco Tufano, and Mario Vento. An algorithm for detection of partially camouflaged people. In *IEEE AVSS*, 2009.
- Jianqin Yin Yanbin Han Wendi Hou and Jinping Li. Detection of the mobile object with camouflage color under dynamic background based on optical flow. *Proceedia Engineering*, 15:2201–2205, 2011.
- Sungho Kim. Unsupervised spectral-spatial feature selection-based camouflaged object detection using vnir hyperspectral camera. *TSWJ*, 2015, 2015.
- Xiang Zhang, Ce Zhu, Shuai Wang, Yipeng Liu, and Mao Ye. A bayesian approach to camouflaged moving object detection. *IEEE TCSVT*, 27(9):2001–2013, 2016.
- Meirav Galun, Eitan Sharon, Ronen Basri, and Achi Brandt. Texture segmentation by multiscale aggregation of filter responses and shape elements. In *ICCV*, 2003.
- Ariel Tankus and Yehezkel Yeshurun. Detection of regions of interest and camouflage breaking by direct convexity estimation. In *IEEE WVS*, 1998.
- Ariel Tankus and Yehezkel Yeshurun. Convexity-based visual camouflage breaking. *CVIU*, 82(3):208–237, 2001.
- Anurag Mittal and Nikos Paragios. Motion-based background subtraction using adaptive kernel density estimation. In *CVPR*, 2004.
- Zhou Liu, Kaiqi Huang, and Tieniu Tan. Foreground object detection using top-down information based on em framework. *IEEE TIP*, 21(9):4204–4217, 2012.

22. Shuai Li, Dinei Florencio, Yaqin Zhao, Chris Cook, and Wanqing Li. Foreground detection in camouflaged scenes. In *ICIP*, 2017.
23. Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE TPAMI*, 44(10):6024–6042, 2022.
24. 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 *CVPR*, 2021.
25. Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *CVPR*, 2021.
26. Haiyang Mei, Xin Yang, Yunduo Zhou, Ge-Peng Ji, Xiaopeng Wei, and Deng-Ping Fan. Distraction-aware camouflaged object segmentation. *SCIS*, 2023.
27. Letian Yu, Haiyang Mei, Wen Dong, Ziqi Wei, Li Zhu, Yuxin Wang, and Xin Yang. Progressive glass segmentation. *IEEE TIP*, 31:2920–2933, 2022.
28. 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. *MIR*, 20:92–108, 2023.
29. Jaya S Kulchandani and Kruti J Dangarwala. Moving object detection: Review of recent research trends. In *IEEE ICPC*, 2015.
30. Ajoy Mondal. Camouflaged object detection and tracking: A survey. *IJIG*, 20(04):2050028, 2020.
31. Hongbo Bi, Cong Zhang, Kang Wang, Jinghui Tong, and Feng Zheng. Rethinking camouflaged object detection: Models and datasets. *IEEE TCSVT*, 32(9):5708–5724, 2022.
32. SHI Caijuan, REN Bijuan, WANG Ziwen, YAN Jinwei, and SHI Ze. Survey of camouflaged object detection based on deep learning. *IFCST*, 16(12):2734, 2022.
33. Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Nick Barnes, and Deng-Ping Fan. Towards deeper understanding of camouflaged object detection. *IEEE TCSVT*, 2023.
34. Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.
35. Jialun Pei, Tianyang Cheng, Deng-Ping Fan, He Tang, Chuanbo Chen, and Luc Van Gool. Osformer: One-stage camouflaged instance segmentation with transformers. In *ECCV*, 2022.
36. Trung-Nghia Le, Yubo Cao, Tan-Cong Nguyen, Minh-Quan Le, Khanh-Duy Nguyen, Thanh-Toan Do, Minh-Triet Tran, and Tam V Nguyen. Camouflaged instance segmentation in-the-wild: Dataset, method, and benchmark suite. *IEEE TIP*, 31:287–300, 2022.
37. Enze Xie, Wenhai Wang, Mingyu Ding, Ruimao Zhang, and Ping Luo. Polarmask++: Enhanced polar representation for single-shot instance segmentation and beyond. *IEEE TPAMI*, 44(9):5385–5400, 2021.
38. Hao Chen, Kunyang Sun, Zhi Tian, Chunhua Shen, Yongming Huang, and Youliang Yan. Blendmask: Top-down meets bottom-up for instance segmentation. In *CVPR*, 2020.
39. Guolei Sun, Zhaochong An, Yun Liu, Ce Liu, Christos Sakaridis, Deng-Ping Fan, and Luc Van Gool. Indiscernible object counting in underwater scenes. In *CVPR*, 2023.
40. Hala Lamdouar, Charig Yang, Weidi Xie, and Andrew Zisserman. Betrayed by motion: Camouflaged object discovery via motion segmentation. In *ACCV*, 2020.
41. Licheng Jiao, Ruohan Zhang, Fang Liu, Shuyuan Yang, Biao Hou, Lingling Li, and Xu Tang. New generation deep learning for video object detection: A survey. *IEEE TNNLS*, 33(8):3195–3215, 2022.
42. 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 *CVPR*, 2022.
43. Deng-Ping Fan, Ming-Ming Cheng, Jiang-Jiang Liu, Shang-Hua Gao, Qibin Hou, and Ali Borji. Salient objects in clutter: Bringing salient object detection to the foreground. In *ECCV*, 2018.
44. Shengfeng He, Rynson WH Lau, Wenxi Liu, Zhe Huang, and Qingxiong Yang. Supercnn: A superpixelwise convolutional neural network for salient object detection. *IJCV*, 115(3):330–344, 2015.
45. Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *CVPR*, 2015.
46. Lijun Wang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Deep networks for saliency detection via local estimation and global search. In *CVPR*, 2015.

47. Jongpil Kim and Vladimir Pavlovic. A shape-based approach for salient object detection using deep learning. In *ECCV*, 2016.
48. Yi Zeng, Pingping Zhang, Jianming Zhang, Zhe Lin, and Huchuan Lu. Towards high-resolution salient object detection. In *ICCV*, 2019.
49. Nian Liu and Junwei Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *CVPR*, 2016.
50. Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *CVPR*, 2019.
51. Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Baocai Yin. Learning uncertain convolutional features for accurate saliency detection. In *ICCV*, 2017.
52. Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip H. S. Torr. Deeply supervised salient object detection with short connections. *IEEE TPAMI*, 41(4):815–828, 2019.
53. Mingchen Zhuge, Deng-Ping Fan, Nian Liu, Dingwen Zhang, Dong Xu, and Ling Shao. Salient object detection via integrity learning. *IEEE TPAMI*, 45(3):3738–3752, 2023.
54. Yi Liu, Qiang Zhang, Dingwen Zhang, and Jungong Han. Employing deep part-object relationships for salient object detection. In *ICCV*, 2019.
55. Qi Qi, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. Multi-scale capsule attention-based salient object detection with multi-crossed layer connections. In *ICME*, 2019.
56. Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer. In *ICCV*, 2021.
57. Guanbin Li and Yizhou Yu. Deep contrast learning for salient object detection. In *CVPR*, 2016.
58. Youbao Tang and Xiangqian Wu. Saliency detection via combining region-level and pixel-level predictions with cnns. In *ECCV*, 2016.
59. Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, 2017.
60. Guanbin Li, Yuan Xie, and Liang Lin. Weakly supervised salient object detection using image labels. In *AAAI*, 2018.
61. Chunshui Cao, Yongzhen Huang, Zilei Wang, Liang Wang, Ninglong Xu, and Tieniu Tan. Lateral inhibition-inspired convolutional neural network for visual attention and saliency detection. In *AAAI*, 2018.
62. Bo Li, Zhengxing Sun, and Yuqi Guo. Supervae: Superpixelwise variational autoencoder for salient object detection. In *AAAI*, 2019.
63. Yu Zeng, Yunzhi Zhuge, Huchuan Lu, Lihe Zhang, Mingyang Qian, and Yizhou Yu. Multi-source weak supervision for saliency detection. In *CVPR*, 2019.
64. Dingwen Zhang, Junwei Han, and Yu Zhang. Supervision by fusion: Towards unsupervised learning of deep salient object detector. In *ICCV*, 2017.
65. Jing Zhang, Tong Zhang, Yuchao Dai, Mehrtash Harandi, and Richard Hartley. Deep unsupervised saliency detection: A multiple noisy labeling perspective. In *CVPR*, 2018.
66. Gyungin Shin, Samuel Albanie, and Weidi Xie. Unsupervised salient object detection with spectral cluster voting. In *CVPR*, 2022.
67. Shengfeng He, Jianbo Jiao, Xiaodan Zhang, Guoqiang Han, and Rynson WH Lau. Delving into salient object subitizing and detection. In *ICCV*, 2017.
68. Md Amirul Islam, Mahmoud Kalash, and Neil DB Bruce. Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects. In *CVPR*, 2018.
69. Wenguan Wang, Jianbing Shen, Xingping Dong, and Ali Borji. Salient object detection driven by fixation prediction. In *CVPR*, 2018.
70. Srinivas SS Kruthiventi, Vennela Gudisa, Jaley H Dhokakiya, and R Venkatesh Babu. Saliency unified: A deep architecture for simultaneous eye fixation prediction and salient object segmentation. In *CVPR*, 2016.
71. Yu Zeng, Yunzhi Zhuge, Huchuan Lu, and Lihe Zhang. Joint learning of saliency detection and weakly supervised semantic segmentation. In *ICCV*, 2019.
72. Linzhao Wang, Lijun Wang, Huchuan Lu, Pingping Zhang, and Xiang Ruan. Saliency detection with recurrent fully convolutional networks. In *ECCV*, 2016.

73. Xin Li, Fan Yang, Hong Cheng, Wei Liu, and Ding-gang Shen. Contour knowledge transfer for salient object detection. In *ECCV*, 2018.
74. Wenguan Wang, Shuyang Zhao, Jianbing Shen, Steven CH Hoi, and Ali Borji. Salient object detection with pyramid attention and salient edges. In *CVPR*, 2019.
75. 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 *CVPR*, 2019.
76. Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnnet: Edge guidance network for salient object detection. In *ICCV*, 2019.
77. Jinming Su, Jia Li, Yu Zhang, Changqun Xia, and Yonghong Tian. Selectivity or invariance: Boundary-aware salient object detection. In *ICCV*, 2019.
78. Lu Zhang, Jianming Zhang, Zhe Lin, Huchuan Lu, and You He. Capsal: Leveraging captioning to boost semantics for salient object detection. In *CVPR*, 2019.
79. Guanbin Li, Yuan Xie, Liang Lin, and Yizhou Yu. Instance-level salient object segmentation. In *CVPR*, 2017.
80. Xin Tian, Ke Xu, Xin Yang, Baocai Yin, and Rynson WH Lau. Learning to detect instance-level salient objects using complementary image labels. *IJCV*, 130(3):729–746, 2022.
81. Ruochen Fan, Ming-Ming Cheng, Qibin Hou, Tai-Jiang Mu, Jingdong Wang, and Shi-Min Hu. S4net: Single stage salient-instance segmentation. In *CVPR*, 2019.
82. Yu-Huan Wu, Yun Liu, Le Zhang, Wang Gao, and Ming-Ming Cheng. Regularized densely-connected pyramid network for salient instance segmentation. *IEEE TIP*, 30:3897–3907, 2021.
83. Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE TPAMI*, 35(1):185–207, 2012.
84. Ali Borji. Saliency prediction in the deep learning era: Successes and limitations. *IEEE TPAMI*, 43(2):679–700, 2019.
85. 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 TPAMI*, 44(8):4339–4354, 2022.
86. Deng-Ping Fan, Zheng Lin, Ge-Peng Ji, Dingwen Zhang, Huazhu Fu, and Ming-Ming Cheng. Taking a deeper look at co-salient object detection. In *CVPR*, 2020.
87. Dingwen Zhang, Huazhu Fu, Junwei Han, Ali Borji, and Xuelong Li. A review of co-saliency detection algorithms: fundamentals, applications, and challenges. *ACM TIST*, 9(4):1–31, 2018.
88. Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li. Salient object detection: A survey. *CVMJ*, 5(2):117–150, 2019.
89. Wenguan Wang, Qiuxia Lai, Huazhu Fu, Jianbing Shen, Haibin Ling, and Ruigang Yang. Salient object detection in the deep learning era: An in-depth survey. *IEEE TPAMI*, 44(6):3239–3259, 2021.
90. Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *IEEE TIP*, 24(12):5706–5722, 2015.
91. Tao Zhou, Deng-Ping Fan, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Rgb-d salient object detection: A survey. *CVMJ*, 7(1):37–69, 2021.
92. Deng-Ping Fan, Zheng Lin, Zhao Zhang, Menglong Zhu, and Ming-Ming Cheng. Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks. *IEEE TNNLS*, 32(5):2075–2089, 2020.
93. Runmin Cong, Kepu Zhang, Chen Zhang, Feng Zheng, Yao Zhao, Qingming Huang, and Sam Kwong. Does thermal really always matter for rgb-t salient object detection? *IEEE TMM*, 2022.
94. Zhengzheng Tu, Zhun Li, Chenglong Li, Yang Lang, and Jin Tang. Multi-interactive dual-decoder for rgb-thermal salient object detection. *IEEE TIP*, 30:5678–5691, 2021.
95. Keren Fu, Yao Jiang, Ge-Peng Ji, Tao Zhou, Qijun Zhao, and Deng-Ping Fan. Light field salient object detection: A review and benchmark. *CVMJ*, pages 1–26, 2022.
96. Wenguan Wang, Jianbing Shen, and Ling Shao. Video salient object detection via fully convolutional networks. *IEEE TIP*, 27(1):38–49, 2017.
97. Trung-Nghia Le and Akihiro Sugimoto. Deeply supervised 3d recurrent fcn for salient object detection in videos. In *BMVC*, 2017.

98. Chenglizhao Chen, Guotao Wang, Chong Peng, Yuming Fang, Dingwen Zhang, and Hong Qin. Exploring rich and efficient spatial temporal interactions for real-time video salient object detection. *IEEE TIP*, 30:3995–4007, 2021.
99. Trung-Nghia Le and Akihiro Sugimoto. Video salient object detection using spatiotemporal deep features. *IEEE TIP*, 27(10):5002–5015, 2018.
100. Miao Zhang, Jie Liu, Yifei Wang, Yongri Piao, Shunyu Yao, Wei Ji, Jingjing Li, Huchuan Lu, and Zhongxuan Luo. Dynamic context-sensitive filtering network for video salient object detection. In *ICCV*, 2021.
101. Guanbin Li, Yuan Xie, Tianhao Wei, Keze Wang, and Liang Lin. Flow guided recurrent neural encoder for video salient object detection. In *CVPR*, 2018.
102. Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. Pyramid dilated deeper convlstm for video salient object detection. In *ECCV*, 2018.
103. Ge-Peng Ji, Deng-Ping Fan, Keren Fu, Zhe Wu, Jianbing Shen, and Ling Shao. Full-duplex strategy for video object segmentation. *CVMJ*, 8:155–175, 2022.
104. Haofeng Li, Guanqi Chen, Guanbin Li, and Yizhou Yu. Motion guided attention for video salient object detection. In *ICCV*, 2019.
105. Runmin Cong, Weiyu Song, Jianjun Lei, Guanghui Yue, Yao Zhao, and Sam Kwong. Psnet: Parallel symmetric network for video salient object detection. *IEEE TETCI*, 7(2):402–414, 2023.
106. Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *CVPR*, 2019.
107. Xue-Jing Luo, Shuo Wang, Zongwei Wu, Christos Sakaridis, Yun Cheng, Deng-Ping Fan, and Luc Van Gool. Camdiff: Camouflage image augmentation via diffusion model. *arXiv preprint arXiv:2304.05469*, 2023.
108. Aixuan Li, Jing Zhang, Yunqiu Lv, Bowen Liu, Tong Zhang, and Yuchao Dai. Uncertainty-aware joint salient object and camouflaged object detection. In *CVPR*, 2021.
109. Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, and Luc Van Gool. Highly accurate dichotomous image segmentation. In *ECCV*, 2022.
110. Jinnan Yan, Trung-Nghia Le, Khanh-Duy Nguyen, Minh-Triet Tran, Thanh-Toan Do, and Tam V Nguyen. Mirrornet: Bio-inspired camouflaged object segmentation. *IEEE Access*, 9:43290–43300, 2021.
111. Mochu Xiang, Jing Zhang, Yunqiu Lv, Aixuan Li, Yiran Zhong, and Yuchao Dai. Exploring depth contribution for camouflaged object detection. *arXiv preprint arXiv:2106.13217*, 2021.
112. Kang Wang, Hongbo Bi, Yi Zhang, Cong Zhang, Ziqi Liu, and Shuang Zheng. D²c-net: A dual-branch, dual-guidance and cross-refine network for camouflaged object detection. *IEEE TIE.*, 69(5):5364–5374, 2022.
113. Yujia Sun, Geng Chen, Tao Zhou, Yi Zhang, and Nian Liu. Context-aware cross-level fusion network for camouflaged object detection. In *IJCAI*, 2021.
114. Nobukatsu Kajiura, Hong Liu, and Shin’ichi Satoh. Improving camouflaged object detection with the uncertainty of pseudo-edge labels. In *ACM MM Asia*, 2021.
115. Jinchao Zhu, Xiaoyu Zhang, Shuo Zhang, and Junnan Liu. Inferring camouflaged objects by texture-aware interactive guidance network. In *AAAI*, 2021.
116. Qiang Zhai, Xin Li, Fan Yang, Chenglizhao Chen, Hong Cheng, and Deng-Ping Fan. Mutual graph learning for camouflaged object detection. In *CVPR*, 2021.
117. 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 *ICCV*, 2021.
118. 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*, 2022.
119. Cong Zhang, Kang Wang, Hongbo Bi, Ziqi Liu, and Lina Yang. Camouflaged object detection via neighbor connection and hierarchical information transfer. *CVIU*, 221:103450, 2022.
120. Wei Zhai, Yang Cao, HaiYong Xie, and Zheng-Jun Zha. Deep texton-coherence network for camouflaged object detection. *IEEE TMM*, 2022.
121. Geng Chen, Si-Jie Liu, Yu-Jia Sun, Ge-Peng Ji, Ya-Feng Wu, and Tao Zhou. Camouflaged object de-

- tection via context-aware cross-level fusion. *IEEE TCSVT*, 32(10):6981–6993, 2022.
122. Mingchen Zhuge, Xiankai Lu, Yiyao Guo, Zhihua Cai, and Shuhan Chen. Cubenet: X-shape connection for camouflaged object detection. *PR*, 127:108644, 2022.
 123. Ge-Peng Ji, Lei Zhu, Mingchen Zhuge, and Keren Fu. Fast camouflaged object detection via edge-based reversible re-calibration network. *PR*, 123:108414, 2022.
 124. Qiao Zhang, Yanliang Ge, Cong Zhang, and Hongbo Bi. Tprnet: camouflaged object detection via transformer-induced progressive refinement network. *TVCI*, pages 1–15, 2022.
 125. Yao Cheng, Hao-Zhou Hao, Yi Ji, Ying Li, and Chun-Ping Liu. Attention-based neighbor selective aggregation network for camouflaged object detection. In *IJCNN*, 2022.
 126. 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*, 2022.
 127. Tao Zhou, Yi Zhou, Chen Gong, Jian Yang, and Yu Zhang. Feature aggregation and propagation network for camouflaged object detection. *IEEE TIP*, 31:7036–7047, 2022.
 128. Peng Li, Xuefeng Yan, Hongwei Zhu, Mingqiang Wei, Xiao-Ping Zhang, and Jing Qin. Findnet: Can you find me? boundary-and-texture enhancement network for camouflaged object detection. *IEEE TIP*, 31:6396–6411, 2022.
 129. Mu-Chun Chou, Hung-Jen Chen, and Hong-Han Shuai. Finding the achilles heel: Progressive identification network for camouflaged object detection. In *ICME*, 2022.
 130. Jiawei Liu, Jing Zhang, and Nick Barnes. Modeling aleatoric uncertainty for camouflaged object detection. In *WACV*, 2022.
 131. Yujia Sun, Shuo Wang, Chenglizhao Chen, and Tian-Zhu Xiang. Boundary-guided camouflaged object detection. In *IJCAI*, 2022.
 132. Miao Zhang, Shuang Xu, Yongri Piao, Dongxiang Shi, Shusen Lin, and Huchuan Lu. Preynet: Preying on camouflaged objects. In *ACM MM*, 2022.
 133. Zhengyi Liu, Zhili Zhang, Yacheng Tan, and Wei Wu. Boosting camouflaged object detection with dual-task interactive transformer. In *ICPR*, 2022.
 134. 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 *CVPR*, 2022.
 135. Yijie Zhong, Bo Li, Lv Tang, Senyun Kuang, Shuang Wu, and Shouhong Ding. Detecting camouflaged object in frequency domain. In *CVPR*, 2022.
 136. Qi Jia, Shuilian Yao, Yu Liu, Xin Fan, Risheng Liu, and Zhongxuan Luo. Segment, magnify and reiterate: Detecting camouflaged objects the hard way. In *CVPR*, 2022.
 137. Qiang Zhai, Xin Li, Fan Yang, Zhicheng Jiao, Ping Luo, Hong Cheng, and Zicheng Liu. Mgl: Mutual graph learning for camouflaged object detection. *IEEE TIP*, 32:1897–1910, 2023.
 138. Jiaying Lin, Xin Tan, Ke Xu, Lizhuang Ma, and Rynson WH Lau. Frequency-aware camouflaged object detection. *ACM TMCCA*, 19(2):1–16, 2023.
 139. 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 TCSVT*, 33(3):1157–1167, 2023.
 140. Haozhe Xing, Yan Wang, Xujun Wei, Hao Tang, Shuyong Gao, and Wenqiang Zhang. Go closer to see better: Camouflaged object detection via object area amplification and figure-ground conversion. *IEEE TCSVT*, 2023.
 141. Dehua Zheng, Xiaochen Zheng, Laurence T Yang, Yuan Gao, Chenlu Zhu, and Yiheng Ruan. Mffn: Multi-view feature fusion network for camouflaged object detection. In *WACV*, 2023.
 142. Ruozhen He, Qihua Dong, Jiaying Lin, and Rynson WH Lau. Weakly-supervised camouflaged object detection with scribble annotations. In *AAAI*, 2023.
 143. Xiaobin Hu, Deng-Ping Fan, Xuebin Qin, Hang Dai, Wenqi Ren, Ying Tai, Chengjie Wang, and Ling Shao. High-resolution iterative feedback network for camouflaged object detection. In *AAAI*, 2023.
 144. Zhou Huang, Hang Dai, Tian-Zhu Xiang, Shuo Wang, Huai-Xin Chen, Jie Qin, and Huan Xiong. Feature

- shrinkage pyramid for camouflaged object detection with transformers. In *CVPR*, 2023.
145. Chunming He, Kai Li, Yachao Zhang, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Camouflaged object detection with feature decomposition and edge reconstruction. In *CVPR*, 2023.
 146. Naisong Luo, Yuwen Pan, Rui Sun, Tianzhu Zhang, Zhiwei Xiong, and Feng Wu. Camouflaged instance segmentation via explicit de-camouflaging. In *CVPR*, 2023.
 147. Wei Sun, Chengao Liu, Linyan Zhang, Yu Li, Pengxu Wei, Chang Liu, Jialing Zou, Jianbin Jiao, and Qixiang Ye. Dqnet: Cross-model detail querying for camouflaged object detection. *arXiv preprint arXiv:2212.08296*, 2022.
 148. Bowen Yin, Xuying Zhang, Qibin Hou, Bo-Yuan Sun, Deng-Ping Fan, and Luc Van Gool. Camoformer: Masked separable attention for camouflaged object detection. *arXiv preprint arXiv:2212.06570*, 2023.
 149. 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*, 2023.
 150. Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.
 151. Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *ICCV*, 2015.
 152. Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *AISTATS*, 2015.
 153. Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.
 154. Christopher Xie, Yu Xiang, Zaid Harchaoui, and Dieter Fox. Object discovery in videos as foreground motion clustering. In *CVPR*, 2019.
 155. Hala Lamdouar, Weidi Xie, and Andrew Zisserman. Segmenting invisible moving objects. In *BMVC*, 2021.
 156. Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie. Self-supervised video object segmentation by motion grouping. In *ICCV*, 2021.
 157. Pia Bideau, Erik Learned-Miller, Cordelia Schmid, and Karteek Alahari. The right spin: Learning object motion from rotation-compensated flow fields. *arXiv preprint arXiv:2203.00115*, 2022.
 158. Junyu Xie, Weidi Xie, and Andrew Zisserman. Segmenting moving objects via an object-centric layered representation. In *NeurIPS*, 2022.
 159. Etienne Meunier, Anaïs Badoual, and Patrick Bouthemy. Em-driven unsupervised learning for efficient motion segmentation. *IEEE TPAMI*, 45(4):4462–4473, 2023.
 160. Matthew Kowal, Mennatullah Siam, Md Amirul Islam, Neil DB Bruce, Richard P Wildes, and Konstantinos G Derpanis. A deeper dive into what deep spatiotemporal networks encode: Quantifying static vs. dynamic information. In *CVPR*, 2022.
 161. 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 *CVPR*, 2016.
 162. Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. *IEEE TPAMI*, 36(6):1187–1200, 2013.
 163. Pia Bideau and Erik Learned-Miller. It’s moving! a probabilistic model for causal motion segmentation in moving camera videos. In *ECCV*, 2016.
 164. Liulei Li, Tianfei Zhou, Wenguan Wang, Lu Yang, Jianwu Li, and Yi Yang. Locality-aware inter-and intra-video reconstruction for self-supervised correspondence learning. In *CVPR*, 2022.
 165. Nikita Araslanov, Simone Schaub-Meyer, and Stefan Roth. Dense unsupervised learning for video segmentation. In *NeurIPS*, 2021.
 166. Runtao Liu, Zhirong Wu, Stella Yu, and Stephen Lin. The emergence of objectness: Learning zero-shot segmentation from videos. In *NeurIPS*, 2021.
 167. Xiankai Lu, Wenguan Wang, Jianbing Shen, Yu-Wing Tai, David J Crandall, and Steven CH Hoi. Learning video object segmentation from unlabeled videos. In *CVPR*, 2020.
 168. Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

169. Zhongdao Wang, Hengshuang Zhao, Ya-Li Li, Shengjin Wang, Philip Torr, and Luca Bertinetto. Do different tracking tasks require different appearance models? In *NeurIPS*, 2021.
170. Bin Yan, Yi Jiang, Peize Sun, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. Towards grand unification of object tracking. In *ECCV*, 2022.
171. Hao-fei Xu, Jing Zhang, Jian-fei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *arXiv preprint arXiv:2211.05783*, 2022.
172. Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020.
173. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
174. 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 TPAMI*, 43(2):652–662, 2019.
175. Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019.
176. Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022.
177. 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 *NeurIPS*, 2021.
178. 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. *CVMIJ*, 8(3):415–424, 2022.
179. 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 *ICCV*, 2021.
180. Deng-Ping Fan, Ge-Peng Ji, Xuebin Qin, and Ming-Ming Cheng. Cognitive vision inspired object segmentation metric and loss function. *SCIS*, 51(9):1475–1489, 2021.
181. Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *IJ-CAI*, 2018.
182. Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *ICCV*, 2017.
183. Ming-Ming Cheng and Deng-Ping Fan. Structure-measure: A new way to evaluate foreground maps. *IJCV*, 129(9):2622–2638, 2021.
184. Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip HS Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *IJCV*, 130(8):2022–2039, 2022.
185. Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *NC*, 312:135–153, 2018.
186. Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.
187. Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM CSUR*, 53(3):1–34, 2020.
188. Timothy Hospedales, Antreas Antoniou, Paul Miccaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *IEEE TPAMI*, 44(9):5149–5169, 2021.
189. Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S Yu, and Lichao Sun. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv preprint arXiv:2303.04226*, 2023.
190. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *CACM*, 63(11):139–144, 2020.
191. Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016.
192. Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018.
193. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

194. Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.
195. Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.
196. Ge-Peng Ji, Deng-Ping Fan, Peng Xu, Ming-Ming Cheng, Bowen Zhou, and Luc Van Gool. Sam struggles in concealed scenes—empirical study on “segment anything”. *arXiv preprint arXiv:2304.06022*, 2023.
197. Jonathan Masci, Ueli Meier, Dan Ciresan, Jürgen Schmidhuber, and Gabriel Fricout. Steel defect classification with max-pooling convolutional neural networks. In *IJCNN*, 2012.
198. Arnaz Malhi and Robert X Gao. Pca-based feature selection scheme for machine defect classification. *IEEE TIM*, 53(6):1517–1525, 2004.
199. Qiwu Luo, Xiaoxin Fang, Jiaojiao Su, Jian Zhou, Bingxing Zhou, Chunhua Yang, Li Liu, Weihua Gui, and Lu Tian. Automated visual defect classification for flat steel surface: a survey. *IEEE TIM*, 69(12):9329–9349, 2020.
200. Henry YT Ngan, Grantham KH Pang, and Nelson HC Yung. Automated fabric defect detection—a review. *IVC*, 29(7):442–458, 2011.
201. Ajay Kumar. Computer-vision-based fabric defect detection: A survey. *IEEE TIE*, 55(1):348–363, 2008.
202. Santanu Ghorai, Anirban Mukherjee, M Gangadaran, and Pranab K Dutta. Automatic defect detection on hot-rolled flat steel products. *IEEE TIM*, 62(3):612–621, 2012.
203. Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. In *VISIGRAPP*, 2018.
204. Domen Tabernik, Samo Šela, Jure Skvarč, and Danijel Skočaj. Segmentation-based deep-learning approach for surface-defect detection. *JIM*, 31(3):759–776, 2020.
205. Du-Ming Tsai, Shu-Kai S Fan, and Yi-Hsiang Chou. Auto-annotated deep segmentation for surface defect detection. *IEEE TIM*, 70:1–10, 2021.
206. Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. The mvtec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection. *IJCV*, 129(4):1038–1059, 2021.
207. Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad—a comprehensive real-world dataset for unsupervised anomaly detection. In *CVPR*, 2019.
208. Ke-Chen Song, Shao-Peng Hu, Yun-Hui Yan, and Jun Li. Surface defect detection method using saliency linear scanning morphology for silicon steel strip under oil pollution interference. *ISIJ International*, 54(11):2598–2607, 2014.
209. Yanqi Bao, Kechen Song, Jie Liu, Yanyan Wang, Yunhui Yan, Han Yu, and Xingjie Li. Triplet-graph reasoning network for few-shot metal generic surface defect segmentation. *IEEE TIM*, 70:1–11, 2021.
210. Yu He, Kechen Song, Qinggang Meng, and Yunhui Yan. An end-to-end steel surface defect detection approach via fusing multiple hierarchical features. *IEEE TIM*, 69(4):1493–1504, 2019.
211. Yong Shi, Limeng Cui, Zhiquan Qi, Fan Meng, and Zhensong Chen. Automatic road crack detection using random structured forests. *IEEE TITS*, 17(12):3434–3445, 2016.
212. Limeng Cui, Zhiquan Qi, Zhensong Chen, Fan Meng, and Yong Shi. Pavement distress detection using random decision forests. In *ICDS*, 2015.
213. Yibin Huang, Congying Qiu, and Kui Yuan. Surface defect saliency of magnetic tile. *TVCJ*, 36(1):85–96, 2020.
214. Ge-Peng Ji, Mingchen Zhuge, Dehong Gao, Deng-Ping Fan, Christos Sakaridis, and Luc Van Gool. Masked vision-language transformer in fashion. *MIR*, 2023.