基于任意来源深度的物体凸显*

吴宗蔚 [1,2,3] Danda Pani Paudel[1,4] 范登平 [1†] 王晶晶 [5] 王硕 [1]
Cédric Demonceaux[2,6] Radu Timofte[3] Luc Van Gool[1,4]

[1] CVL, ETH Zurich [2] University of Burgundy, CNRS, ICB [3] Computer Vision Lab, CAIDAS & IFI, University of Wurzburg
[4] INSAIT, Sofia University [5] AUST [6] University of Lorraine, CNRS, Inria, Loria

摘要

众所周知，深度线索对视觉感知非常有用。然而，直接测量深度往往是不切实际的。幸运的是，当前基于深度学习的方法可以通过推理提供有质量保证的深度图。在这项工作中，本文利用物体在三维空间中的“凸显”先验，将这种深度预测模型用于物体分割。“凸显”是一个简单的构成先验，假设物体位于背景表面。这种构成先验允许本文对三维空间中的物体进行建模。更具体地说，本文调整预测的深度图，以便仅使用三维信息就能定位物体。然而，这种分离需要有关接触面的知识，于是本文使用分割掩膜的弱监督来学习这些知识。本文利用接触面作为中间层的表示，从而对物体进行纯三维推理，这使本文能够更好地将深度知识转化为语义。该文章所提出的适应方法仅需要一个提前训练好的深度预测模型，而不需要用于训练的源数据，使得学习过程高效实用。本文在两个挑战性任务(即显著性物体检测和伪装物体检测)的 8 个数据集上进行了实验，实验结果一致证明了本文的方法在性能和普适性方面的优势。源代码请参见https://github.com/Zongwei97/PopNet。

1. 引言

众所周知，场景的三维知识是可以对视觉感知任务起到重要的补充作用 [12, 15, 54, 64, 88]。但在实际应用中，视觉感知往往只能通过二维图像来实现。给定多幅图像后，可以使用结构-运动技术来恢复三维几何图形 ([24, 39, 53, 83]。然而，当只有一张图像可用时，这种反演方法就不再适用。在这种情况下，通常使用基于学习的方法能将图像反演为深度图 [13, 43, 48, 49, 62]，这些方法在最近几年取得了巨大的成功。遗憾的是，因跨域泛化能力不足，这些方法可能无法提供高质量的深度图。

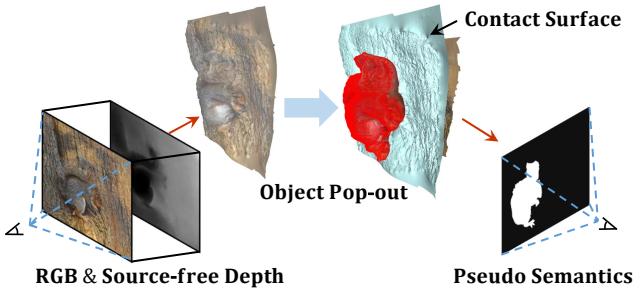

图 1. 使用**物体凸显** (object pop-out) 先验进行深度到语义的转换。对于输入的 RGB 图像和无源深度图像对，本文学习接触表面 (contact surface)。然后，使用获得的接触表面来将物体和背景分开，从而得到伪语义信息用于监督学习。

尽管泛化能力较差，但在一个领域获得的知识被证明在其他相近的领域中是有用的。这种效用通过执行所谓的域自适应 (DA) 来加以利用 [2, 3, 6, 51, 58, 86]。事实上，最近的研究表明，DA 方法可以仅使用预测模型有效地传递知识，即无需访问模型训练所用的数据，这也被称为无源域自适应 (SDA) [23, 26, 37, 69, 75]。SDA 方法因其高效性和隐私保证而备受关注。

大多数现有的无源域自适应 (SDA) 方法都隐

*本文为 ICCV'23 论文 [67] 的中文翻译版。

†通讯作者: 范登平 (dengpfan@gmail.com)

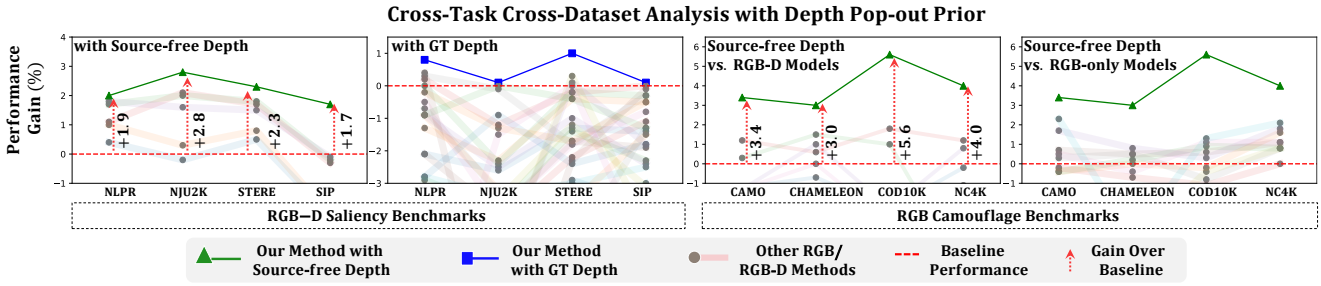

图 2. **使用本文的方法在** F-measure **中获得的性能**与已建立的基线相比。本文的方法（▲,■）在 8 **个数据集**和 2 **个任务**（每个任务在 4 个数据集上–SOD：左边两个；COD：右边两个）上显著提高了基线。本文还比较了 24 **种方法**（●），在这些方法中，本文的方法提供了最先进的结果，尽管它们的任务各有特点。请注意，所有方法都用线连接起来，以说明它们在不同数据集上的性能波动。请参考表 1。更多细节和讨论请参考表 2和章节 4.3。

含了以下一种或两种假设：(a) 在目标域上存在与源域相似的监督任务 [26, 73, 75]；(b) 目标域的离散（且已知）标签空间 [25, 27, 34, 69]。前一假设不仅使得源域和目标域更容易进行比较，而且可能使得这两个域更加接近。后一假设允许通过自训练来进行 SDA，其中离散的标签有助于对模型的置信度进行推理。然后，通过增强目标域上某些可靠示例的置信度来执行自训练。

本研究旨在实现无源深度知识的目标检测传递。这种传递可以帮助通过深度线索定位物体，并在领域差异存在的情况下利用深度知识。所涉及的问题设置与标准无源域自适应（SDA）方法有所不同，具体体现在：(a) 源域和目标域任务的差异；(b) 深度的连续标签空间。这些与标准 SDA 方法的差异使得本文面临的任务非常具有挑战性，在本文目前的知识范围内，本文首次解决了这个问题。为了解决这个具有挑战性的问题，本文依靠“物体凸显（pop-out）”先验，该先验允许本文直接在 3D 空间中推理物体的位置。“物体凸显”是一个简单的组合先验，即假设物体位于背景表面之上。图 1对“物体凸显”先验进行了图形化说明。

Kang 等人在 [22] 中成功地使用了用于图像合成的“凸出先验”。Treisman 的早期工作对这种先验 [59] 进行了深入研究。在这项工作中，本文依赖于这些工作的相同组成基础，并在跨领域转移深度知识之前利用了凸出的表现形式。尽管本文的研究动机来自于这些早期研究，但本文的实验设置在很大程度上不同于这些研究。不同之处不仅在于跨域深度知识转移（任意源数据），还在于仅使用语义进行目标监督。

本文提出的方法利用任意来源的深度将其映射到一个空间中，在该空间中，物体在深度层面相对于背景中更加突出。本文通过学习物体与背景之间的接触表面进行物体与背景分离。这种分离使模型能够得出语义掩码，可直接与真值进行比较，以进行监督。通过对目标的监督，本文可以最大限度地缩小源和目标之间在领域和任务方面的差距，如图 3，本文首先利用物体突出网络鼓励物体从任意来源的深度中突出出来。然后，本文引入另一个网络，即带有接触表面的分割网络，来定位物体并预测接触表面。这些学习模块以端到端的方式进行联合训练，将任意来源的深度转换为适应目标任务（即物体检测）的中间表示。为了评估所提出的方法，本文在两个具有挑战性任务的八个数据集上进行了详尽的实验，分别是显著物体检测和伪装物体检测。在这两个任务中，本文提出的方法显著改进了基线模型，并同时取得了最先进的结果，其概览如图 2 所示。该文的主要贡献如下：

- 本文所面临的跨领域和跨任务的任意源深度知识转移问题既实用又新颖。
- 本文的方法依赖于本文的”凸出先验”来实现视觉理解，这种方法简单而有效。
- 本文的方法在两个不同任务中的结果明显优于基线和现有模型。

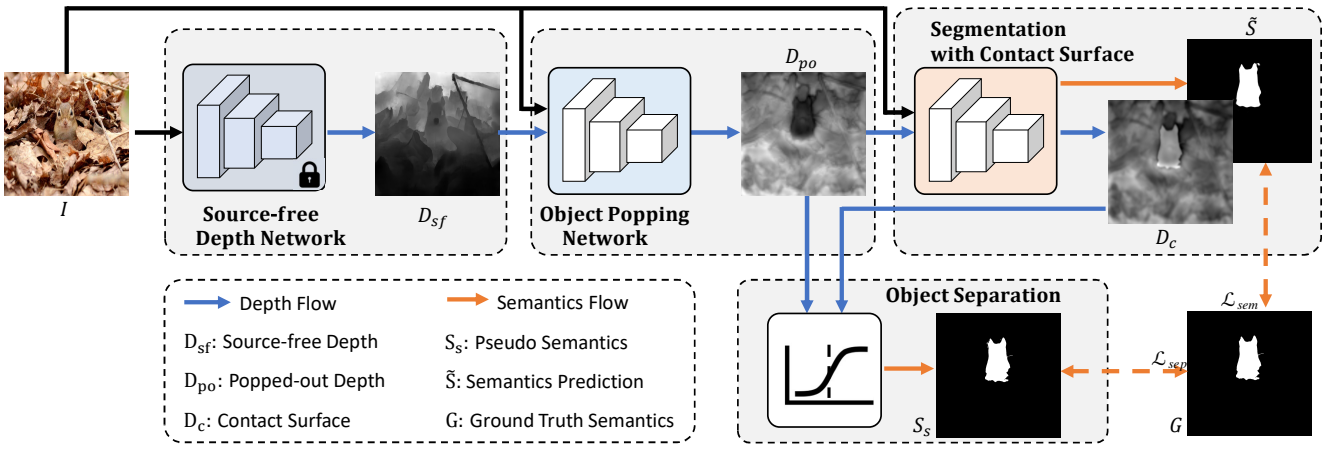

图 3. 本文**提出的框架**被称为 PopNet，由一个任意源网络、一个目标凸出网络、一个接触面分割和一个物体分离模块组成。任意源深度网络以现成的方式生成伪深度（第3.1节）。目标凸出网络将任意源深度转换为目标的凸出深度，弥补了跨领域和跨任务的差距（第3.2节）。分割网络使用这个深度来估计物体的遮罩和接触面（第3.3节）。然后，物体分离模块利用接触面（第3.4节）将凸出深度转换为物体的第二个掩码。本文将两个语义掩码与地面实况进行比较，以端到端的方式对整个管道进行监督。

2. 相关工作

无源适应（Source-free Adaptation）：最近，由于隐私、实用性和高效性等原因，不需要访问源数据的域自适应知识传递方法引起了广泛的关注 [1, 23, 26, 37, 69, 75]。本文观察到，对于通过多阶段训练在各种数据集上学习的现成知识模型（如 Monodepth [13] 和 Midas [49]）来说，为了转移深度知识，访问源数据尤其不切实际。现有的无源自适应方法通常使用生成模型 [26, 27, 32]、伪标签 [23, 35, 61] 或其他定制方法 [52, 74]。本文使用了基于伪标签的方法。然而，使用现有方法并不直接，因为源域和目标域之间的任务存在差异。但通过本文提供的”凸显”技术，本文可以获得伪语义信息，从而更好地在任务之间传递任意来源的深度。

显著性目标检测 (SOD): 显著性检测的目的是检测和分割图像中在视觉上最能吸引人类注意力的区域。大量研究表明，显著性可以作为不同视觉任务的辅助步骤，例如物体跟踪、物体检测等。传统的显著性工作是单模态的，即只需要 RGB 图像作为输入。在普通场景中，基于 RGB 的模型 [36, 68, 81] 已经取得了非常可喜的成果。最近，一些工作 [10, 18, 47, 63, 65, 85, 89] 利用深度图作为三维几何的额外线索，因为深度可以提供更真实的物体边界信息以及尺度感知。这些三维特征进一步提高了在具有挑战性的场景中的检测精度和性能 [20, 31, 66, 77]。

伪装目标检测 (COD): 伪装检测的目的是在图像中发现隐藏的物体。在计算机视觉领域，主要的工作、[9, 30, 45] 经常将 COD 与 SOD 进行比较。一些研究表明，简单地在 COD 上扩展 SOD 模型会导致不理想的结果，这主要是由目标对象的性质所引起的，即目标对象的隐蔽性或突出性。因此，为了将注意力限制在隐蔽物体上，一些研究提出了不同的感知系统来模仿人类对伪装物体的行为，例如三阶段定位-分段-排序策略；迭代细化 [19]，类似于反复观察图像；放大到可能的区域 [45, 56]。其他工作 [16, 38, 57, 79, 80, 87, 90, 91] 借助梯度 [16] 深入探索纹理差异、频率 [87]、边缘 [57, 90] 和概率 [30, 72] 深入探索纹理差异。

最新的心理学研究 [5, 7] 表明，人类感知可以自然地从深度线索中获益来理解场景：（A）物体内部的平滑变化有助于减轻假边缘并保留物体结构；（B）物体边界上的深度不连续性可以使分割更容易。受这些观察结果的启发，本文旨在探索 SOD 和 COD 任务中的任意来源的深度。为了解决任意来源深度图的领域空白问题，本文建议以端到端的方式，在自监督损失和弱语义监督的情况下，联合微调任意来源的深度和语义网络。

3. 所提出的 PopNet

给定一个大小为 $I \in \mathbb{R}^{3\times H\times W}$ 的输入 RGB 图像，其中 H 和 W 分别是图像的高度和宽度，本文的目标是预测用于目标检测的语义掩膜 $\tilde{S} \in \mathbb{R}^{H\times W}$。如图 3 所示，首先将输入图像 I 输入到一个冻结权重的深度网络，生成无源深度 $D_{sf} \in \mathbb{R}^{H\times W}$（第 3.1 节）。然后，将模拟的多模态图像一起输入深度凸显网络，计算中间的凸显深度 $D_{po} \in \mathbb{R}^{H\times W}$（第 3.2 节）。这个中间表示以及 RGB 图像 I 之后会被送入分割网络，并转换为接触表面 $D_c \in \mathbb{R}^{H\times W}$ 和语义预测 $\tilde{S} \in \mathbb{R}^{H\times W}$（第 3.3 节）。一方面，语义预测直接受到真实掩膜 GT（记为 G）的监督，这类似于传统的分割监督。另一方面，本文进一步探索接触表面，通过提出的物体分离模块（第 3.4 节）将物体从背景中凸显出来。这将几何线索转换为伪语义信息，并提供另一层次的监督。

3.1. 任意来源的深度网络

在实际应用中，GT 深度并不总是可用的。因此，本文以现成的方式生成任意来源的深度 D_{sf}，以模拟多模态输入。本文选择了最先进的 DPT 模型 [48] 并冻结其权重作为本文的深度预测网络。这种选择是基于它的泛化能力 [49]。为了通过增强局部细节获得最高质量的深度图，本文将增强方法 [43] 与 DPT 结合使用。尽管基于深度学习的方法取得了可信的结果，但由于领域差距，获得的任意来源的深度并不总能提供高质量的几何线索。因此，本文利用几何先验和语义先验来共同微调任意来源的深度。

3.2. 目标凸出网络

网络结构：本文构建了一个深度凸显网络来对任意源深度进行细化/平滑。深度凸显网络采用编码器-解码器设计，使用跳跃连接，如图 4 所示。本文将 RGB 图像和无源深度简单地在输入端融合，形成一个 4 通道输入，然后将其输入深度凸显网络。编码器提取语义线索并生成五个尺度的输出。本文的解码器由 Conv2D、BN、ReLU 和上采样层组成。类似于 U-Net [50]，本文通过简单的加法构建了一个连接。

结构保持：为了监督本文的凸显网络，本文首先利用

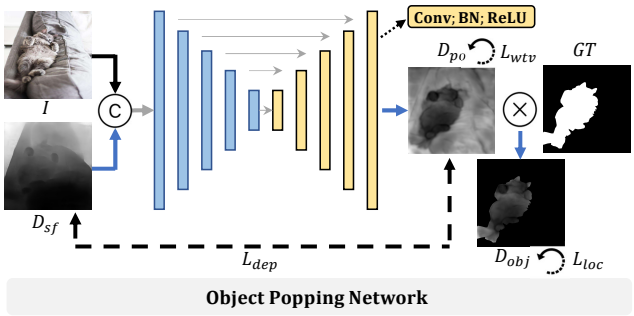

图 4. 本文的**目标凸出网络**将 RGB-D 输入映射到凸出深度。该网络分别使用结构保留、局部深度平滑和深度边缘锐化损失 $\mathcal{L}_{dep}$ 、$\mathcal{L}_{loc}$ 和 $\mathcal{L}_{wtv}$ 进行监督。

生成的伪深度辅助深度细化。这里本文仅使用结构相似性，因为本文的目标是检测、保持和提取物体结构。本文使用以下 SSIM 损失 [13] 来衡量结构相似性。

$$\mathcal{L}_{dep} = SSIM(D_{po}, D_{sf}). \tag{1}$$

局部深度平滑：除了伪深度监督外，本文还结合了语义和几何先验提出了两种约束深度的损失方法。本文假设物体的结构应该与背景区分开，即在物体区域内应该是平滑的，在边界像素处应该是锐利的。因此，本文提出了利用弱语义线索和几何先验的方法。具体来说，本文首先引入了局部平滑损失。局部性由地面真实语义 G 定义。在技术上，本文通过元素级乘法将背景像素屏蔽，以通过 $D_{obj} = D_{po} \otimes G$ 抑制非活动区域。设 ∇_x 和 ∇_y 为 Sobel 操作。那么，本文的局部损失 $\mathcal{L}_{loc}$ 表示为：

$$\begin{aligned} \overrightarrow{n(p)} &= (-\nabla_x(D_{obj}(p));\ -\nabla_y(D_{obj}(p));\ 1); \\ \mathcal{L}_{loc} &= \sum_p \sum_{q\in\mathcal{N}(p)} 1 - cosine(\overrightarrow{n(p)};\ \overrightarrow{n(q)}), \end{aligned} \tag{2}$$

其中 $overrightarrow$ 表示法线，p 表示目标区域内的像素，$\mathcal{N}(p)$ 表示相邻像素，$cosine$ 表示两个向量之间的余弦相似度。这样，本文的局部损失只作用于对象区域，使得对象结构在目标区域内保持一致。应用局部平滑度损失可以减少对象层面的深度噪声。

深度边缘锐化：除了局部平滑损失，本文还使用了边缘锐化。边缘锐化损失被构造为加权总变差。为此，本文首先计算了任意像素 p 处的边缘感知权重

$w(p)$，如下所示：

$$w(p)=\begin{cases} w_0, & \text{if } \nabla_x(G(p))^2+\nabla_y(G(p))^2\neq 0, \\ w_0+\gamma, & \text{otherwise,} \end{cases} \tag{3}$$

其中 w_0 是预先定义的非零权重，γ 是边界像素的附加权重。在本文的设置中，本文选择 w_0 作为边界像素的归一化（按图像大小）计数，并设置 $\gamma=0.5$。本文采用平方形式，使得大梯度起到更重要的作用。本文的加权总变差损失如下所示：

$$\mathcal{L}_{wtv}=\sum_{p}\sum_{q\in\mathcal{N}(p)} w(p)\cdot||D_{po}(p)-D_{po}(q)||_2. \tag{4}$$

本文的加权总方差损失与传统的边缘损失不同。更具体地说，与常用的图像梯度不同，本文的加权函数依赖于语义边界。本文之所以使用语义边界而非图像梯度，是因为本文希望在具有挑战性的条件下进行物体检测，例如伪装物体。在这种情况下，图像梯度可能会导致权重误导。乍一看，本文的损失函数似乎与语义引导的深度估计方法 [4, 33, 70] 类似。然而，[70] 使用的是 GT 深度，而 [4, 33] 使用的是多帧监督。遗憾的是，在本文的设置中，这样的监督是不可能的。同时需要注意的是，本文利用单视角任意来源的深度，同时仅使用真实语义进行监督。此外，尽管不同域之间存在不同，本文还是成功实现了深度知识的转移。现有的基于深度和语义之间正则化的算法也被证明对本文的算法有效。本文希望强调，本文利用凸显先验的网络架构既不是显而易见的，同时又是通用的和易于使用的。

用于监督目标凸出网络的总损失函数 $\mathcal{L}_{pop}$ 由以下公式给出：

$$\mathcal{L}_{pop}=\mathcal{L}_{dep}+\lambda_1\cdot\mathcal{L}_{loc}+\lambda_2\cdot\mathcal{L}_{wtv}, \tag{5}$$

where λ_1 and λ_2 are the hyperparameters.

3.3. 接触面的分割

平滑性和边缘损失促进了对象结构的均匀化，使其与背景区分开。现在，本文的目标是进一步增加物体和背景之间的距离，使得物体结构能够凸显出来。具体来说，本文使用一个 RGB-D 分割网络，如图 5 所示。本文的分割网络的主要组成部分是一个三流 RGB-D 网络，并采用一些融合设计。在本文的设置中，本文选择 [89] 作为基线，因为它是当前 RGB-D 显著性检测领域的最强方法之一。本文添加了一个接触面头，用于学习接触表面的深度 D_c，其分辨率与输入深度 D_{po} 相同。接触面头由 ConvLayer（Conv2D、BN、ReLU）和一个 Conv2D 组成。它首先对特征图进行解码，然后将其转换为一个 1-D 的深度图。

3.4. 分割网络

在前面讨论的接触表面的基础上，本节中本文的目标是将物体与背景分离开来。在这一点上，本文假设接触表面前面的像素属于物体，其余像素属于背景。这个假设能够将 3D 知识明确地转换为 2D 语义。

假设接触表面的预测深度为 $D_c\in\mathbb{R}^{H\times W}$。本文使用凸显深度 D_{po} 和表面深度 D_c 得到伪语义 S_s，如下所示：

$$S_s=sigmoid(\sigma\cdot(D_{po}-D_c)), \tag{6}$$

其中 σ 是一个标量值，用于控制 sigmoid 函数的斜率。在本文的实验中使用 $\sigma=10$ 来执行软阈值，模仿二进制输出所需的硬阈值。这种软阈值有利于训练所需的梯度反向传播。最后，本文用二元交叉熵（BCE）最小化伪语义 S_s 和 GT 语义 G 之间的差距：

$$\mathcal{L}_{sep}=BCE(S_s,G). \tag{7}$$

3.5. 总损失函数

本文的两个可训练模块，即目标凸出和分割网络（图 3，都是以端到端的方式进行训练的。因此，总体损失函数由三部分组成：深度凸出损失（depth popping loss）$\mathcal{L}_{pop}$、物体分离损失（object separation loss）$\mathcal{L}_{sep}$ 以及来自 RGB-D 基线网络的传统语义损失（conventional semantic loss）$\mathcal{L}_{sem}$。用于训练的总损失 $\mathcal{L}_{total}$ 由以下公式给出：

$$\mathcal{L}_{total}=\mathcal{L}_{pop}+\alpha_1\cdot\mathcal{L}_{sep}+\alpha_2\cdot\mathcal{L}_{sem}, \tag{8}$$

其中 α_1 和 α_2 为超参数。

表 1. RGB-D SOD 数据集的定量比较。↑（↓）表示越高（越低）越好。本文使用平均绝对误差 (M)、最大 F-measure(F_m)、S-measure 度量 (S_m) 和最大 E-measure 度量 (E_m) 作为评价指标。G.D. 代表 GT 深度，打叉代表使用伪深度，打勾代表使用 GT 深度。**粗体**表示最佳性能。

G.D.	Public.	Dataset	NLPR [46]				NJUK [21]				STERE [44]				SIP [10]			
		Metric	$M\downarrow$	$F_m\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_m\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_m\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_m\uparrow$	$S_m\uparrow$	$E_m\uparrow$
任意来源深度训练的 RGB-D 模型的性能																		
✗	MM_{21} [82]	DFM-Net	.027	.909	.914	.944	.046	.903	.895	.927	.042	.906	.903	.934	.067	.873	.850	.891
✗	TIP_{22} [60]	DCMF	.027	.915	.921	.943	.044	.908	.903	.929	.041	.909	.907	.931	.067	.873	.853	.893
✗	$CVPR_{22}$ [19]	SegMAR	.024	.923	.920	.952	.036	.921	.909	.941	.037	.916	.907	.936	.052	.893	.872	.914
✗	$CVPR_{22}$ [45]	ZoomNet	.023	.916	.919	.944	.037	.926	.914	.940	.037	.918	.909	.938	.054	.891	.868	.909
✗	Ours	PopNet	.022	.925	.926	.956	.031	.931	.920	.949	.032	.922	.916	.947	.046	.911	.885	.926
使用 GT 深度训练的 RGB-D 模型的性能																		
✓	TIP_{21} [84]	BIANet	.032	.888	.900	.930	.056	.878	.867	.898	.048	.898	.895	.918	.091	.816	.802	.847
✓	TIP_{21} [31]	HAINet	.024	.920	.924	.956	.037	.924	.911	.940	.040	.917	.907	.938	.052	.907	.879	.917
✓	$TNNLS_{21}$ [10]	D3Net	.029	.904	.911	.942	.046	.909	.899	.927	.044	.902	.906	.925	.063	.880	.860	.897
✓	$ECCV_{22}$ [29]	SPSN	.023	.917	.923	.956	.032	.927	.918	.949	.035	.909	.906	.941	.043	.910	.891	.932
✓	Ours	PopNet	.019	.927	.932	.963	.030	.936	.924	.952	.033	.924	.917	.947	.040	.923	.897	.937

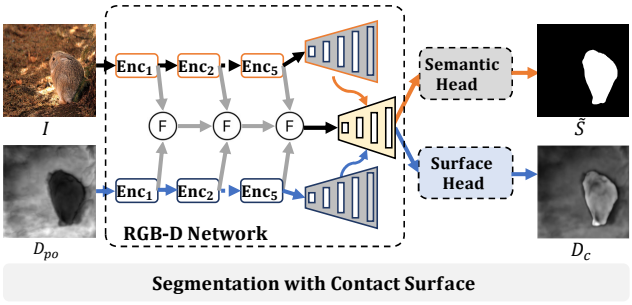

图 5. 本文的**分割网络**使用了一个基本的 RGB-D 三流网络。除了传统的语义头，本文还学习预测像素点的 **接触面**，以便更好地将深度知识转换为语义。

备注： 本文的损失函数以互补的方式发挥作用。$\mathcal{L}_{pop}$ 的作用类似于图像平滑中的平滑滤波器。它通过弱语义标签的帮助，去除由于领域差异导致的深度响应中的噪声，同时保持物体结构。因此，平滑的背景变得不那么信息丰富，而物体区域变得均匀，使其易于检测。这些功能有助于将任意源的深度转换为凸显深度，达到本文期望的效果。这样的过程能将物体置于背景表面之上，即使它们与摄像机和其他干扰表面的距离可能较远。另一方面，$\mathcal{L}_{sep}$ 充分利用了“凸显”先验来从背景中分割出物体。在学习出的接触面的帮助下，该损失将前景和背景拉向相反的方向，从而增大了它们之间的距离。这种拉动导致了一个类似二值化的掩膜，有效地弥合了几何知识和语义之间的差距。最后，深度转换和学习的语义都与地面真值进行比较，用于监督训练。

4. 结果

4.1. 实验步骤

数据集准备：为了更好地说明本文的方法的通用性，本文在 SOD 和 COD 的各项基准上评估了本文的方法的有效性。本文选择了四个广泛使用的 4 个 RGB-D SOD 数据集：NLPR [46], NJUK [21], STERE [44], and SIP [10], 以及 4 个广泛使用的 COD 数据集：i.e., CAMO [28], CHAMELEON [55], COD10K [9], and NC4K [40]。对于 SOD 数据集，本文使用 GT 深度和任意来源的深度进行实验。本文遵循传统的训练准则 [11,17,66,89]，使用 700 幅 NLPR 图像和 1485 幅 NJUK 图像进行训练。其余图像用于测试。对于单模态 COD 数据集，本文同时比较了 RGB COD 模型和在相同任意来源的深度 D_{sf} 的 COD 数据集上重新训练的 RGB-D SOD 模型。本文遵循传统的训练/测试准则 [8,9,19,40,45]，使用来自 COD10K 的 3,040 幅图像和来自 CAMO 的 1,000 幅图像进行训练。其余用于测试。

评价指标：本文用四个公认的指标来评估性能：平均绝对误差 (M)、最大 F-measure(F_m)、S-measure(S_m) 和最大 E-measure(E_m)。所有分割掩码都是从官方资源中重新训练或下载的。为了进行公平的比较，本文使用标准化的评估方法 [89] 对预测语义进行评估。

实施细节：本文的模型是基于 Pytorch 和 V100 GPU

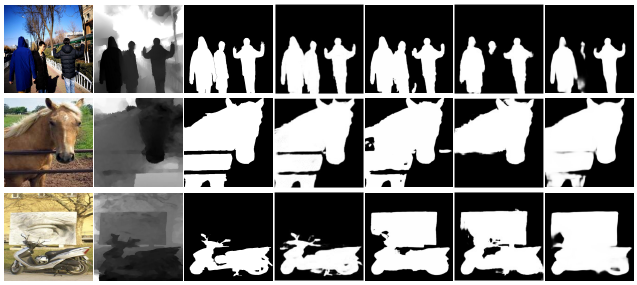

图 6. **显著场景的定性比较**. 在都使用真实深度的情况下，本文的模型表现的任何其他的方法都好。

实现的。本文使用 Adam 算法作为优化器。学习率初始化为 $1e{-}4$，每 60 个 epochs 除以 10。本文将 RGB 和深度的输入分辨率设置为 352×352。表3中提供了与更高分辨率的详细对比。在训练过程中，采用了随机翻转、旋转和边界剪切等传统的数据增强。在 200 个 epochs 中，RGB-SOD 任务的训练时间约为 12 小时，COD 任务的训练时间约为 24 小时。

4.2. 对比

与 RGB-D SOD 模型的比较: 本文在表1中列出了使用本文的任意来源的深度或 GT 深度的 SOD 基准性能。可以看出，与许多采用 GT 深度的 RGB-D 模型相比，本文的任意来源深度模型取得了非常有竞争力的性能。本文采用 GT 深度的方法也优于 SOTA 方法。显著性检测的例子可以在图 6中找到。关于多对象的更多讨论请参见表6。请注意，本文还在 SOD 数据集上仅使用 RGB 图像重新训练了 SOTA 单模态 COD 模型 SegMAR [19] 和 ZoomNet [45] 。本文的结果表明，本文的任意来源深度模型优于这些模型，这表明本文的方法可以在不同的任务中更好地通用，并获得更好的性能。

与 COD 模型的比较: 本文在表2中列出了最具竞争力的 SOTA 方法的性能，包括特定任务的 COD 模型以及经过重新训练的任意来源的深度的 RGB-D SOD 模型。为了进行公平比较，本文对所有 RGB-D 方法、SegMAR [19], and ZoomNet [45] 进行了重新训练，并在与本文相同分辨率的图像上进行了端到端训练。一些纯 RGB 方法的性能甚至优于许多 RGB-D 方法，这主要是由于其针对特定任务的 COD 设计以及缺乏 RGB-D 方法所需的地面实况深度。值得注意的是，传统纯 RGB COD 方法在 SOD 任务中表现不佳。请参考表1 和 3。可以观察到这些方法在跨任务上的普适性较差。

更高的分辨率: 以往的研究表明，图像分辨率可能会影响模型的性能 [41,45,71,80]。例如，当前的 SOTA COD 方法 ZoomNet [45] 的主尺度为 384^2，意味着其最高分辨率为 $(384 \times 1.5)^2 = 576^2$，因为它在 $0.5\times, 1\times$, and $1.5\times$ 尺度上运行。为了进行公平的比较，本文用与相同的分辨率 (352^2 or 512^2) 重新训练模型。在表3-ZoomNet 中显示，其结果正如预期的那样有所恶化。与这些方法相比，本文的方法在准确性和效率之间进行了很好的权衡。关于 SOD 和 COD 基准的更多比较，请参见补充材料。

定性比较: 图 7展示了本文的网络在具有挑战性的情况下的输出结果。可以看出，在处理被薄物体遮挡的物体时（$1^{st} - 4^{th}$ 行），本文的方法能够准确推理出更接近 GT 的分割掩模。本文在处理多个物体时也取得了更好的性能（最后一行）。

4.3. 消融实验

损失: 在本节中，本文将进行实验来分析所提出的损失的有效性。表 4提供了不同损耗组合的定量结果。可以看出，所提出的每种损失都表现良好，与基线相比性能有所提高。关于超参数的更多讨论和消融研究可参见补充材料

作为插件的目标凸出网络: 本文的目标凸出网络可以很容易地适应不同的编码器和不同的现有 RGB-D 模型。例如，使用 ResNet-18 [14] 编码器和基于卷积的解码器，本文的凸出模型只需花费大约 12.7M 的额外学习参数或 48.7MB 的模型大小。本文在表5中显示，与基线相比，本文的方法能够以不到 10%GFlops 的额外成本有效提高性能。

减少训练数据下的凸出: 在这里，本文对使用任意来源深度的优势进行分析。本文通过减少训练数据来进行不同的实验。如图8(左) 所示，当本文的 PopNet 和 RGB 基线模型都使用所有数据进行训练时，本文的 PopNet 在 COD10K 数据集上的最大 F-measure 和 S-measure 分别提高了 5.6% 和 3.9%。如果本文

表 2. COD 数据集的定量比较。Pseudo 代表 RGB-D 方法使用了伪深度。

Pseudo	Public.	Dataset	CAMO [28]				CHAMELEON [55]				COD10K [9]				NC4K [40]			
		Metric	$M\downarrow$	$F_m\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_m\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_m\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_m\uparrow$	$S_m\uparrow$	$E_m\uparrow$
RGB COD **模型的性能**																		
✗	*CVPR*$_{20}$ [9]	SINet	.099	.762	.751	.790	.044	.845	.868	.908	.051	.708	.771	.832	.058	.804	.808	.873
✗	*CVPR*$_{21}$ [40]	SLSR	.080	.791	.787	.843	.030	.866	.889	.938	.037	.756	.804	.854	.048	.836	.839	.898
✗	*CVPR*$_{21}$ [76]	MGL-R	.088	.791	.775	.820	.031	.868	.893	.932	.035	.767	.813	.874	.053	.828	.832	.876
✗	*CVPR*$_{21}$ [42]	PFNet	.085	.793	.782	.845	.033	.859	.882	.927	.040	.747	.800	.880	.053	.820	.829	.891
✗	*CVPR*$_{21}$ [30]	UJSC	.072	.812	.800	.861	.030	.874	.891	.948	.035	.761	.808	.886	.047	.838	.841	.900
✗	*IJCAI*$_{21}$ [56]	C2FNet	.079	.802	.796	.856	.032	.871	.888	.936	.036	.764	.813	.894	.049	.831	.838	.898
✗	*ICCV*$_{21}$ [72]	UGTR	.086	.800	.783	.829	.031	.862	.887	.926	.036	.769	.816	.873	.052	.831	.839	.884
✗	*CVPR*$_{22}$ [19]	SegMAR	.080	.799	.794	.857	.032	.871	.887	.935	.039	.750	.799	.876	.050	.828	.836	.893
✗	*CVPR*$_{22}$ [45]	ZoomNet	.074	.818	.801	.858	.033	.829	.859	.915	.034	.771	.808	.872	.045	.841	.843	.893
使用任意来源的深度重新训练的 RGB-D **模型的性能**																		
✓	*MM*$_{21}$ [77]	CDINet	.100	.638	.732	.766	.036	.787	.879	.903	.044	.610	.778	.821	.067	.697	.793	.830
✓	*CVPR*$_{21}$ [17]	DCF	.089	.724	.749	.834	.037	.821	.850	.923	.040	.685	.766	.864	.061	.765	.791	.878
✓	*TIP*$_{21}$ [31]	HAINet	.084	.782	.760	.829	.028	.876	.876	.942	.049	.735	.781	.865	.057	.809	.804	.872
✓	*ICCV*$_{21}$ [78]	CMINet	.087	.798	.782	.827	.032	.881	.891	.930	.039	.768	.811	.868	.053	.832	.839	.888
✓	*ICCV*$_{21}$ [89]	SPNet	.083	.807	.783	.831	.033	.872	.888	.930	.037	.776	.808	.869	.054	.828	.825	.874
✓	*TIP*$_{22}$ [60]	DCMF	.115	.737	.728	.757	.059	.807	.830	.853	.063	.679	.748	.776	.077	.782	.794	.820
✓	*ECCV*$_{22}$ [29]	SPSN	.084	.782	.773	.829	.032	.866	.887	.932	.042	.727	.789	.854	.059	.803	.813	.867
✓	Ours	PopNet	.073	.821	.806	.869	.022	.893	.910	.962	.031	.789	.827	.897	.043	.852	.852	.908

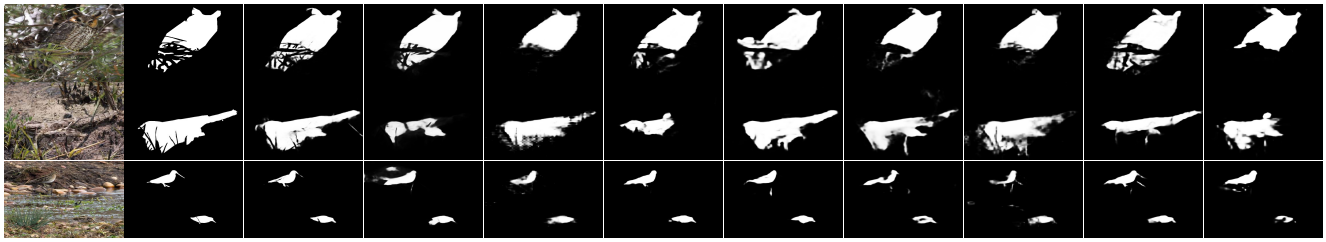

图 7. **伪装定性比较**. 与其他方法相比，本文的方法能够更好地保留对象结构，尤其是在处理有遮挡的对象时（($1^{st}-4^{th}$ 行）。本文的方法在处理多个物体时表现良好（最后一行）。更好地放大。

表 3. 在 SOD 基准上对不同分辨率进行端到端比较。与 SOTA COD 模型相比，本文的任意来源的深度方法具有更好的通用性。

Model	Size	Flops (G)	NJUK [21]				SIP [10]			
			$M\downarrow$	$F_m\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_m\uparrow$	$S_m\uparrow$	$E_m\uparrow$
SegMAR [19]	352^2	67.3	.036	.921	.909	.941	.052	.893	.872	.914
ZoomNet [45]	352^2	167.8	.037	.926	.914	.940	.054	.891	.868	.909
Ours	352^2	228.8	.031	.931	.920	.949	.046	.911	.885	.926
SegMAR [19]	512^2	142.4	.035	.927	.914	.943	.050	.899	.878	.917
ZoomNet [45]	512^2	353.4	.036	.926	.915	.942	.052	.895	.873	.910
Ours	512^2	484.0	.031	.933	.922	.951	.044	.911	.890	.927

表 4. 所提出损失的消融研究.

$\mathcal{L}_{dep}$	$\mathcal{L}_{loc}$	$\mathcal{L}_{wtv}$	$\mathcal{L}_{sep}$	SIP [10]				NC4K [40]			
				$M\downarrow$	$F_m\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_m\uparrow$	$S_m\uparrow$	$E_m\uparrow$
-	-	-	-	.048	.903	.884	.922	.052	.832	.832	.893
✓	-	-	-	.046	.907	.889	.925	.051	.833	.839	.895
-	✓	-	-	.045	.908	.893	.929	.048	.837	.844	.898
-	-	✓	-	.046	.906	.891	.927	.050	.833	.841	.894
-	-	-	✓	.043	.914	.893	.933	.048	.840	.848	.900
✓	✓	-	-	.044	.911	.893	.928	.049	.837	.844	.897
✓	-	✓	-	.046	.909	.893	.927	.046	.840	.845	.898
✓	-	✓	✓	.040	.918	.897	.935	.045	.848	.849	.904
✓	✓	-	✓	.042	.916	.894	.931	.044	.850	.850	.906
✓	✓	✓	✓	.040	.923	.897	.937	.043	.852	.852	.908

的 PopNet 仅使用 25% 的数据进行训练，与本文使用所有数据训练的基线相比，它仍然能够实现具有竞争力的性能。在 NC4K 数据集上也可以观察到类似的现象，如图8（右）所示。总之，本文的方法可以有效地探索几何先验，并显著减少所需的训练数据量。

与基线相比的增益: 如图所示，通过深度线索，本文在 3125 幅图像和 4121 幅图像（~75% 的情况下）中提升了性能。表 6还显示，本文的网络在处理单个或多个对象时的表现优于基线。当任意来源的深度

表 5. 不同 RGB-D 基线的通用性和成本。

Dataset	Flops	Param	SIP [10]				NC4K [40]			
Metric	(G)	(M)	$M\downarrow$	$F_m\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_m\uparrow$	$S_m\uparrow$	$E_m\uparrow$
HAINet [31]	363.2	59.8	.053	.899	.874	.919	.057	.809	.804	.872
+ Ours	373.7	72.5	.051	.910	.886	.923	.055	.814	.811	.878
SPNet [89]	149.0	150.4	.044	.911	.887	.914	.054	.828	.825	.874
+ Ours	159.5	163.1	.042	.917	.894	.932	.044	.851	.851	.905

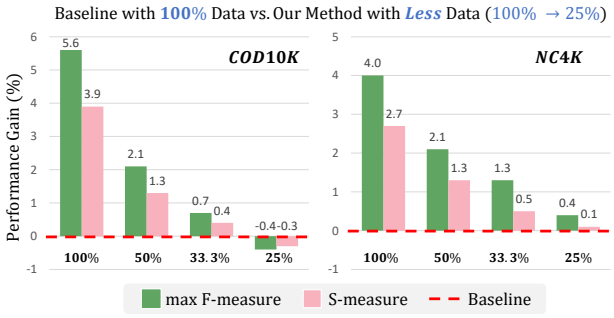

图 8. 本文的方法在减少训练数据方面的优势。当本文的网络仅使用 25% 的数据进行训练时，其性能与基线相比仍然具有竞争力，与使用所有数据训练的基线相比，在 NC4K 上的绝对性能提升为最大 F-measure+0.4%，S-measure+0.1%。

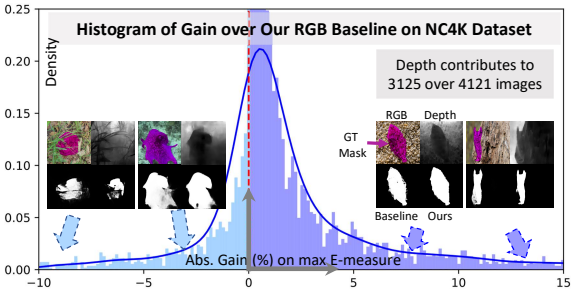

图 9. 增益直方图。请放大查看详情。

表 6. 在 NC4K [40] 上的多目标性能（大小为 512^2）。

Obj. Nbr. (%)	Single (92%)		Two (6%)		More (2%)		Overall	
Metric	$M\downarrow$	$F_m\uparrow$	$M\downarrow$	$F_m\uparrow$	$M\downarrow$	$F_m\uparrow$	$M\downarrow$	$F_m\uparrow$
RGB Baseline	.054	.828	.067	.811	.091	.738	.056	.825
+ D_{sf}	.050	.842	.063	.821	.093	.746	.051	.839
Ours	.040	.864	.051	.847	.079	.767	.042	.861

为无线索时，本文的方法可能会失败。这种情况主要发生在物体被很好地隐藏起来，从而骗过了深度预测网络。然而，这种情况也是非常有挑战的，甚至对人类来说也是如此。

RGB-D **方法比纯** RGB **方法更好吗？**在有 GT 深度的情况下，RGB-D 方法确实优于纯 RGB 方法。然而，只有少数 RGB-D 方法可以从任意来源的深度中获益。例如，如表 7所示，当使用任意来源的深度训练时，DASNet [85] 与 RGB 基线相比性能较差。同

表 7. 使用 RGB-D_{sf} 模型与仅使用 RGB 基线的性能对比. D_{sf} 代表任意来源深度.

Dataset	COD10K [9]				NC4K [40]			
Metric	$M\downarrow$	$F_m\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_m\uparrow$	$S_m\uparrow$	$E_m\uparrow$
DASNet [85]	.041	.643	.793	.864	.055	.747	.830	.879
+ D_{sf}	.041	.642	.796	.858	.055	.743	.830	.874
SPNet [89]	.040	.743	.801	.867	.052	.846	.833	.883
+ D_{sf}	.037	.776	.808	.869	.054	.828	.825	.874

样，即使是 SOTA RGB-D 模型之一的 SPNet [89]，在 NC4K 数据集上仅使用 RGB 输入时的性能也优于使用额外的任意来源的深度输入时的性能。此外，当提供任意来源的深度时，现有的 RGB-D 方法在 COD 上的表现都不如表现最好的纯 RGB 方法（e.g., ZoomNet [45]）。在 SOD 上也有同样的观察结果。这可能是由于领域差距和融合设计等原因造成的。本文还发现，将性能最好的纯 RGB 方法扩展到 RGB-D 情况并非易事。本文希望强调，提出的 PopNet 在任意来源深度图或 GT 深度图情况下的表现优于所有现有的纯 RGB 和 RGB-D 方法。

5. 结论

本文成功地展示了通过仅使用源模型的跨领域跨任务深度到语义知识传递的一个案例。在本文中，本文利用给定源模型提供的目标的无源深度。所提出的方法通过物体的凸显先验来学习从深度到语义的知识传递。本文通过设计一种新颖的网络架构来帮助本文的网络使用这种先验。设计的网络通过从提供的深度图中将物体凸显出来再进行推理。然后，通过学习的接触面将物体与背景分离。本文展示了通过目标语义成功监督物体凸显和接触表面的联合学习。在 SOD 和 COD 基准测试上进行的详尽实验显示了深度知识成功地传递到目标领域，表现出了提高的性能和泛化能力。

参考文献

[1] Peshal Agarwal, Danda Pani Paudel, Jan-Nico Zaech, and Luc Van Gool. Unsupervised robust domain adaptation without source data. In WACV, 2022. 3

[2] Amir Atapour-Abarghouei and Toby P Breckon. Real-time monocular depth estimation using synthetic data

with domain adaptation via image style transfer. In IEEE CVPR, 2018. 1

[3] Adriano Cardace, Luca De Luigi, Pierluigi Zama Ramirez, Samuele Salti, and Luigi Di Stefano. Plugging self-supervised monocular depth into unsupervised domain adaptation for semantic segmentation. In WACV, 2022. 1

[4] Po-Yi Chen, Alexander H Liu, Yen-Cheng Liu, and Yu-Chiang Frank Wang. Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019. 5

[5] Innes C Cuthill, Martin Stevens, Jenna Sheppard, Tracey Maddocks, C Alejandro Párraga, and Tom S Troscianko. Disruptive coloration and background pattern matching. Nature, 434(7029):72–74, 2005. 3

[6] Jiahua Dong, Zhen Fang, Anjin Liu, Gan Sun, and Tongliang Liu. Confident anchor-induced multi-source free domain adaptation. NeurIPS, 2021. 1

[7] Aliya El Nagar, Daniel Osorio, Sarah Zylinski, and Steven M Sait. Visual perception and camouflage response to 3d backgrounds and cast shadows in the european cuttlefish, sepia officinalis. JEB, 224(11), 2021. 3

[8] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. IEEE TPAMI, 44(10):6024–6042, 2022. 6

[9] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In IEEE CVPR, 2020. 3, 6, 8, 9

[10] Deng-Ping Fan, Zheng Lin, Zhao Zhang, Menglong Zhu, and Ming-Ming Cheng. Rethinking RGB-D salient object detection: Models, datasets, and large-scale benchmarks. IEEE TNNLS, 32(5):2075–2089, 2021. 3, 6, 8, 9

[11] Deng-Ping Fan, Yingjie Zhai, Ali Borji, Jufeng Yang, and Ling Shao. BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network. In ECCV, 2020. 6

[12] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In IEEE CVPR, 2019. 1

[13] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In IEEE ICCV, 2019. 1, 3, 4

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In IEEE CVPR, 2016. 7

[15] Ian P Howard. Perceiving in depth, volume 1: basic mechanisms. Oxford University Press, 2012. 1

[16] 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, 2023. 3

[17] Wei Ji, Jingjing Li, Shuang Yu, Miao Zhang, Yongri Piao, Shunyu Yao, Qi Bi, Kai Ma, Yefeng Zheng, Huchuan Lu, et al. Calibrated RGB-D salient object detection. In IEEE CVPR, 2021. 6, 8

[18] Wei Ji, Jingjing Li, Miao Zhang, Yongri Piao, and Huchuan Lu. Accurate RGB-D salient object detection via collaborative learning. In ECCV, 2020. 3

[19] 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 CVPR, 2022. 3, 6, 7, 8

[20] Wen-Da Jin, Jun Xu, Qi Han, Yi Zhang, and Ming-Ming Cheng. Cdnet: Complementary depth network for rgb-d salient object detection. IEEE TIP, 30:3376–3390, 2021. 3

[21] Ran Ju, Ling Ge, Wenjing Geng, Tongwei Ren, and Gangshan Wu. Depth saliency based on anisotropic center-surround difference. In IEEE ICIP, 2014. 6, 8

[22] Hongwen Kang, Alexei A Efros, Martial Hebert, and Takeo Kanade. Image composition for object pop-out. In IEEE ICCVW, 2009. 2

[23] Youngeun Kim, Donghyeon Cho, Kyeongtak Han, Priyadarshini Panda, and Sungeun Hong. Domain adaptation without source data. IEEE TAI, 2(6):508–518, 2021. 1, 3

[24] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In IEEE CVPR, 2021. 1

[25] Jogendra Nath Kundu, Akshay R Kulkarni, Suvaansh Bhambri, Deepesh Mehta, Shreyas Anand Kulkarni, Varun Jampani, and Venkatesh Babu Radhakrishnan. Balancing discriminability and transferability for source-free domain adaptation. In ICML, 2022. 2

[26] Jogendra Nath Kundu, Naveen Venkat, R Venkatesh Babu, et al. Universal source-free domain adaptation. In IEEE CVPR, 2020. 1, 2, 3

[27] Vinod K Kurmi, Venkatesh K Subramanian, and Vinay P Namboodiri. Domain impression: A source data free domain adaptation method. In WACV, 2021. 2, 3

[28] Trung-Nghia Le, Tam V Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. CVIU, 184:45–56, 2019. 6, 8

[29] Minhyeok Lee, Chaewon Park, Suhwan Cho, and Sangyoun Lee. Spsn: Superpixel prototype sampling network for rgb-d salient object detection. In ECCV, 2022. 6, 8

[30] Aixuan Li, Jing Zhang, Yunqiu Lv, Bowen Liu, Tong Zhang, and Yuchao Dai. Uncertainty-aware joint salient object and camouflaged object detection. In IEEE CVPR, 2021. 3, 8

[31] Gongyang Li, Zhi Liu, Minyu Chen, Zhen Bai, Weisi Lin, and Haibin Ling. Hierarchical alternate interaction network for rgb-d salient object detection. IEEE TIP, 30:3528–3542, 2021. 3, 6, 8, 9

[32] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In IEEE CVPR, 2020. 3

[33] Rui Li, Danna Xue, Shaolin Su, Xiantuo He, Qing Mao, Yu Zhu, Jinqiu Sun, and Yanning Zhang. Learning depth via leveraging semantics: Self-supervised monocular depth estimation with both implicit and explicit semantic guidance. Pattern Recognition (PR), page 109297, 2023. 5

[34] Xinhao Li, Jingjing Li, Lei Zhu, Guoqing Wang, and Zi Huang. Imbalanced source-free domain adaptation. In ACM MM, 2021. 2

[35] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In ICML, 2020. 3

[36] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. In IEEE CVPR, 2019. 3

[37] Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. In IEEE CVPR, 2021. 1, 3

[38] Yukang Lu, Dingyao Min, Keren Fu, and Qijun Zhao. Depth-cooperated trimodal network for video salient object detection. In ICIP. IEEE, 2022. 3

[39] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. ACM TOG, 39(4):1–13, 2020. 1

[40] 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 CVPR, 2021. 6, 8, 9

[41] Daniel McKee, Zitong Zhan, Bing Shuai, Davide Modolo, Joseph Tighe, and Svetlana Lazebnik. Transfer of representations to video label propagation: Implementation factors matter. arXiv preprint arXiv:2203.05553, 2022. 7

[42] Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In IEEE CVPR, 2021. 8

[43] S Mahdi H Miangoleh, Sebastian Dille, Long Mai, Sylvain Paris, and Yagiz Aksoy. Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. In IEEE CVPR, 2021. 1, 4

[44] Yuzhen Niu, Yujie Geng, Xueqing Li, and Feng Liu. Leveraging stereopsis for saliency analysis. In IEEE CVPR, 2012. 6

[45] 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 CVPR, 2022. 3, 6, 7, 8, 9

[46] Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. RGBD salient object detection: a benchmark and algorithms. In ECCV, 2014. 6

[47] Yongri Piao, Zhengkun Rong, Miao Zhang, Weisong Ren, and Huchuan Lu. A2dele: Adaptive and attentive depth distiller for efficient RGB-D salient object detection. In IEEE CVPR, 2020. 3

[48] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In IEEE ICCV, 2021. 1, 4

[49] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE TPAMI, 44(3):1623–1637, 2022. 1, 3, 4

[50] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In MICCAI, 2015. 4

[51] Suman Saha, Anton Obukhov, Danda Pani Paudel, Menelaos Kanakis, Yuhua Chen, Stamatios Georgoulis, and Luc Van Gool. Learning to relate depth and semantics for unsupervised domain adaptation. In IEEE CVPR, 2021. 1

[52] Roshni Sahoo, Divya Shanmugam, and John Guttag. Unsupervised domain adaptation in the absence of source data. arXiv preprint arXiv:2007.10233, 2020. 3

[53] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In IEEE CVPR, 2016. 1

[54] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In ECCV, 2012. 1

[55] Przemysław Skurowski, Hassan Abdulameer, J Błaszczyk, Tomasz Depta, Adam Kornacki, and P Kozieł. Animal camouflage analysis: Chameleon database. Unpublished manuscript, 2(6):7, 2018. 6, 8

[56] Yujia Sun, Geng Chen, Tao Zhou, Yi Zhang, and Nian Liu. Context-aware cross-level fusion network for camouflaged object detection. In IJCAI, 2021. 3, 8

[57] Yujia Sun, Shuo Wang, Chenglizhao Chen, and Tian-Zhu Xiang. Boundary-guided camouflaged object detection. In IJCAI, 2022. 3

[58] Alessio Tonioni, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Unsupervised domain adaptation for depth prediction from images. IEEE TPAMI, 42(10):2396–2409, 2019. 1

[59] Anne Treisman. Preattentive processing in vision. ICVGIP, 31(2):156–177, 1985. 2

[60] Fengyun Wang, Jinshan Pan, Shoukun Xu, and Jinhui Tang. Learning discriminative cross-modality features for RGB-D saliency detection. IEEE TIP, 31:1285–1297, 2022. 6, 8

[61] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In IEEE CVPR, 2022. 3

[62] Cho-Ying Wu, Jialiang Wang, Michael Hall, Ulrich Neumann, and Shuochen Su. Toward practical monocular indoor depth estimation. In IEEE CVPR, 2022. 1

[63] Yu-Huan Wu, Yun Liu, Jun Xu, Jia-Wang Bian, Yu-Chao Gu, and Ming-Ming Cheng. Mobilesal: Extremely efficient rgb-d salient object detection. IEEE TPAMI, 44(12):10261–10269, 2022. 3

[64] Zongwei Wu, Guillaume Allibert, Christophe Stolz, and Cédric Demonceaux. Depth-adapted CNN for RGB-D cameras. In ACCV, 2020. 1

[65] Zongwei Wu, Guillaume Allibert, Christophe Stolz, Chao Ma, and Cédric Demonceaux. Modality-guided subnetwork for salient object detection. In 3DV, 2021. 3

[66] Zongwei Wu, Shriarulmozhivarman Gobichettipalayam, Brahim Tamadazte, Guillaume Allibert, Danda Pani Paudel, and Cédric Demonceaux. Robust RGB-D fusion for saliency detection. 3DV, 2022. 3, 6

[67] 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. In IEEE ICCV, 2023. 1

[68] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In IEEE CVPR, 2019. 3

[69] Haifeng Xia, Handong Zhao, and Zhengming Ding. Adaptive adversarial network for source-free domain adaptation. In IEEE ICCV, 2021. 1, 2, 3

[70] Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. Structure-guided ranking loss for single image depth prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 5

[71] 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. 7

[72] 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 IEEE ICCV, 2021. 3, 8

[73] Shiqi Yang, Joost van de Weijer, Luis Herranz, Shangling Jui, et al. Exploiting the intrinsic neighborhood structure for source-free domain adaptation. NeurIPS, 2021. 2

[74] Shiqi Yang, Yaxing Wang, Joost van de Weijer, and Luis Herranz. Unsupervised domain adaptation without source data by casting a bait. arXiv preprint arXiv:2010.12427, 2020. 3

[75] Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, and Shangling Jui. Generalized source-free domain adaptation. In IEEE ICCV, 2021. 1, 2, 3

[76] Qiang Zhai, Xin Li, Fan Yang, Chenglizhao Chen, Hong Cheng, and Deng-Ping Fan. Mutual graph learning for camouflaged object detection. In IEEE CVPR, 2021. 8

[77] Chen Zhang, Runmin Cong, Qinwei Lin, Lin Ma, Feng Li, Yao Zhao, and Sam Kwong. Cross-modality discrepant interaction network for RGB-D salient object detection. In ACM MM, 2021. 3, 8

[78] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Xin Yu, Yiran Zhong, Nick Barnes, and Ling Shao. RGB-D saliency detection via cascaded mutual information minimization. In IEEE ICCV, 2021. 8

[79] Jing Zhang, Yunqiu Lv, Mochu Xiang, Aixuan Li, Yuchao Dai, and Yiran Zhong. Depth-guided camouflaged object detection. CoRR, abs/2106.13217, 2021. 3

[80] Miao Zhang, Shuang Xu, Yongri Piao, Dongxiang Shi, Shusen Lin, and Huchuan Lu. Preynet: Preying on camouflaged objects. In ACM MM, 2022. 3, 7

[81] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Baocai Yin. Learning uncertain convolutional features for accurate saliency detection. In IEEE ICCV, 2017. 3

[82] Wenbo Zhang, Ge-Peng Ji, Zhuo Wang, Keren Fu, and Qijun Zhao. Depth quality-inspired feature manipulation for efficient RGB-D salient object detection. In ACM MM, 2021. 6

[83] Zhoutong Zhang, Forrester Cole, Richard Tucker, William T Freeman, and Tali Dekel. Consistent depth of moving objects in video. ACM TOG, 40(4):1–12, 2021. 1

[84] Zhao Zhang, Zheng Lin, Jun Xu, Wen-Da Jin, Shao-Ping Lu, and Deng-Ping Fan. Bilateral attention network for rgb-d salient object detection. IEEE TIP, 30:1949–1961, 2021. 6

[85] Jiawei Zhao, Yifan Zhao, Jia Li, and Xiaowu Chen. Is depth really necessary for salient object detection? In ACM MM, 2020. 3, 9

[86] Shanshan Zhao, Huan Fu, Mingming Gong, and Dacheng Tao. Geometry-aware symmetric domain adaptation for monocular depth estimation. In IEEE CVPR, 2019. 1

[87] Yijie Zhong, Bo Li, Lv Tang, Senyun Kuang, Shuang Wu, and Shouhong Ding. Detecting camouflaged object in frequency domain. In IEEE CVPR, 2022. 3

[88] Tao Zhou, Deng-Ping Fan, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Rgb-d salient object detection: A survey. CVMJ, pages 1–33, 2021. 1

[89] Tao Zhou, Huazhu Fu, Geng Chen, Yi Zhou, Deng-Ping Fan, and Ling Shao. Specificity-preserving RGB-D saliency detection. In IEEE ICCV, 2021. 3, 5, 6, 8, 9

[90] 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. 3

[91] Jinchao Zhu, Xiaoyu Zhang, Shuo Zhang, and Junnan Liu. Inferring camouflaged objects by texture-aware interactive guidance network. In AAAI, 2021. 3