ARTICLE
https://doi.org/10.26599/TAI.2022.9150000
CAAI Artificial Intelligence Research

用于高分辨率二分图像分割的双边参考算法

郑鹏[1257], 高德宏[3], 范登平 *[1], 刘丽[4], Jorma Laaksonen[5], 欧阳万里[6], Nicu Sebe[7]

ABSTRACT

我们引入了一种新颖的双边参考框架（BiRefNet），用于高分辨率二分图像分割（DIS）。该框架由两个基本组件组成：定位模块（LM）和我们提出的双边参考（BiRef）重建模块（RM）。LM 利用全局语义信息辅助目标定位。在 RM 中，我们利用 BiRef 进行重建过程，其中图像的层次化块作为源参考，梯度图作为目标参考。这些组件协同工作生成最终的预测图。此外，我们引入了辅助梯度监督，以增强对细节区域的关注。此外，我们还概述了针对 DIS 的实用训练策略，以提高图像质量和训练过程。为了验证我们方法的广泛适用性，我们在四个任务上进行了广泛的实验，证明 BiRefNet 在所有基准测试中表现出色，超越了任务特定的最新方法。我们的代码可见于：https://github.com/ZhengPeng7/BiRefNet。

KEYWORDS

二分图像分割，伪装目标检测，显著目标检测，双边参考，高分辨率分割

图 1　我们提出的 BiRefNet 与最新的最先进方法（例如，IS-Net [1] 和 UDUN [2]）在高分辨率二分图像分割（DIS）结果的视觉比较。细节部分进行了放大显示，以便更好地展示分割效果。

1　引言

随着高分辨率图像获取技术的进步，图像分割技术已经从传统的粗略定位发展到实现高精度的目标分割。这项任务，无论是涉及显著性目标检测 [6] 还是隐藏目标检测 [7,8]，都被称为高分辨率二分图像分割（DIS） [9]，并在工业界中引起了广泛关注和应用，例如三星、Adobe 和迪士尼等公司。

对于新的 DIS 任务，近期的研究考虑了中间监督 [9]、频率先验 [10] 和统一-分割-统一 [11] 等策略，并取得了不错的成果。本质上，它们要么在特征层级进行监督分割 [9,11]，要么引入额外的先验 [10] 以增强特征提取。然而，这些策略仍不足以捕捉非常细微的特征（见 Fig. 1）。根据我们的观察，通过对原始图像进行导数运算获得梯度特

1 南开大学，天津 300350, 中国 * 通讯作者 (dengpfan@gmail.com).
2 阿里巴巴，杭州 311121, 中国
3 西北工业大学，西安 710072, 中国
4 国防科技大学，长沙 410073, 中国
5 阿尔托大学，埃斯波 FI-02150，芬兰
6 上海人工智能实验室，上海 200232, 中国
7 特伦托大学，特伦托 I-38122，意大利

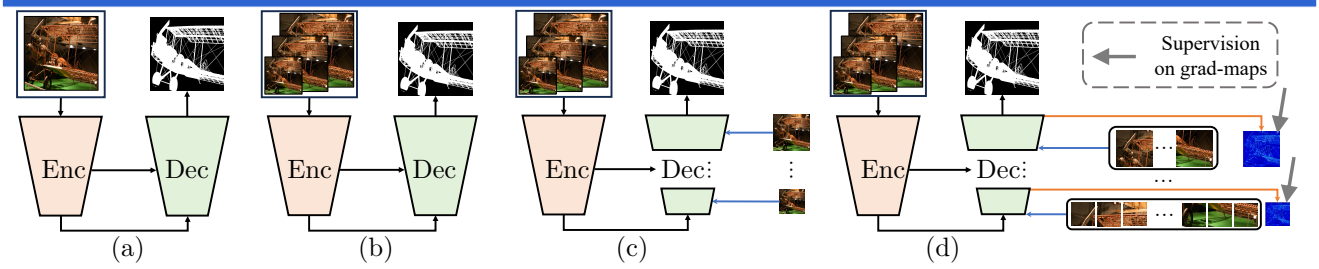

图 2 我们提出的 BiRefNet 与其他现有高分辨率分割任务方法的比较。 (a) **常见框架** [1]; (b) **图像金字塔作为输入** [2,3]; (c) **缩放后的图像作为内部参考** [4,5]; (d) BiRefNet：**原始尺度图像的块作为内部参考，梯度先验作为外部参考**。Enc = **编码器**, Dec = **解码器**。

征，可以很好地反映图像对象中的细微和非显著特征。此外，当某些位置在颜色和纹理上与背景高度相似时，梯度特征可能过于微弱。针对这种情况，我们进一步引入真实值（GT）特征作为侧向监督，使框架能够学习这些位置的特征。我们将图像参考的结合以及梯度和 GT 参考的引入称为*双边参考*。

我们提出了一种新颖的渐进双边参考网络 BiRefNet，用于处理高分辨率 DIS 任务，其中包括独立的定位和重建模块。对于定位模块，我们从视觉 Transformer 主干中提取层次化特征，这些特征在深层中被合并和压缩，以在低分辨率下获得粗略预测。对于重建模块，我们进一步设计了内部和外部参考作为双边参考（BiRef），其中源图像和梯度图在不同阶段输入解码器。与将原始图像调整大小以确保与每个阶段解码特征一致的方法不同 [4,5]，我们保持原始分辨率以保留内部参考的完整细节特征，并适应性地将它们裁剪成块，以与解码特征兼容。此外，我们还研究并总结了针对高分辨率（HR）数据的实用训练策略，包括长时间训练和区域级别的损失，以更好地分割细节部分，并采用多阶段监督以加速对它们的学习。

以下是我们的主要贡献总结：

1. 我们提出了一种名为**双边参考网络**（BiRefNet）的方法，作为执行高质量二分图像分割的简单但强大的基线模型。
2. 我们设计了一个**双边参考模块**，包括具有源图像引导的内部参考和具有梯度监督的外部参考。该模块在重建预测的高分辨率结果方面表现出显著的效果。
3. 我们探索并总结了各种**针对 DIS 定制的实用策略**，以便轻松提升性能、预测质量和收敛加速。
4. 我们的提出的 BiRefNet 在不仅在 DIS 任务上展现出卓越表现，还在 HRSOD 和 COD 任务上均取得了**领先水平**，分别提升了 6.8%、2.0% 和 5.6% 的平均 S_m 值 [12]。

2 相关工作

2.1 高分辨率下的类别无关式图像分割

高分辨率无类别图像分割一直是几十年来典型的计算机视觉目标，许多相关任务被提出并引起了广泛关注，如二分图像分割（DIS） [9]、高分辨率显著目标检测（HRSOD） [13] 和隐藏目标检测（COD） [8]。为了提供标准的 HRSOD 基准，提出了几个典型的 HRSOD 数据集（如 HRSOD [13]、UHRSD [14]、HRS10K [15]）和大量方法 [4,13,14,16]。Zeng 等人 [13] 在他们的网络中采用了多尺度输入的全局-局部融合。Xie 等人 [14] 使用跨模型移植模块处理来自多个主干（轻量级 [17] 和重型 [18]）的不同尺度图像。金字塔混合也在 [4] 中用于降低计算成本。由于周围干扰物的相似外观，隐藏目标很难定位 [19]。因此，图像先验如频率 [20]、边界 [21]、梯度 [22] 等，被用作辅助指导来训练 COD 模型。此外，研究发现更高的分辨率有助于目标检测 [22–24]。为了生成更精确和细节丰富的分割结果，Yin 等人 [25] 采用了带有掩码分离注意力的渐进细化。Li 等人 [5] 在细化过程中结合了不同尺度的原始图像。

高分辨率 DIS 是一项新提出的任务，更多地关注高分辨率图像中目标物体的复杂纤细结构，使其更具挑战性。Qin 等人 [9] 提出了 DIS5K 数据集和具有中间监督的 IS-Net，以缓解细微区域的丢失。此外，Zhou 等人 [10] 在其 DIS 网络中嵌入了频率先验，以捕捉更多细节。Pei 等人 [11] 将标签解耦策略 [26] 应用于 DIS 任务，在物体边界区域实现了具有竞争力的分割性能。Yu 等人 [27] 使用高分辨率图像的块以更节省内存的方式加速训练。与以前使用压缩/调整大小图像来增强高分辨率分割的模型不同，我们利用完整的高分辨率图像作为补充信息，以获得更高分辨率的更好预测。

2.2 分割中的渐进式细化

在图像抠图任务中，三分图（trimaps）被用作预定位技术，以获得更精确的分割结果 [28,29]。在 Fig. 2中，我们

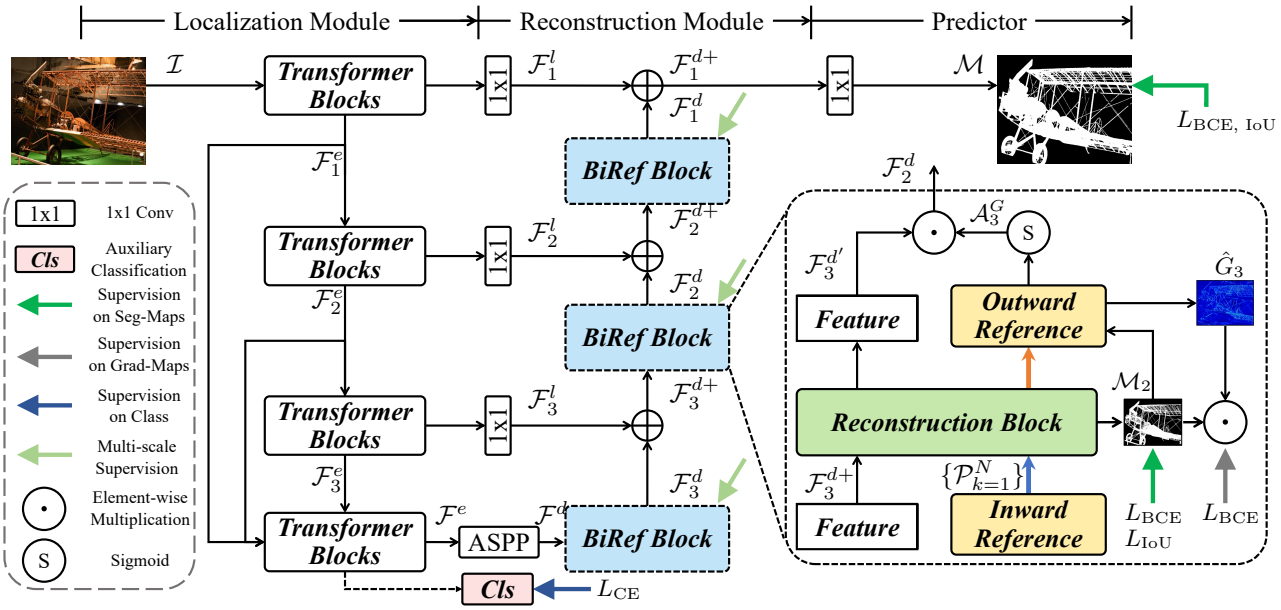

图 3 所提出的双边参考网络（BiRefNet）的流程图。BiRefNet **主要由定位模块（LM）和具有双边参考（BiRef）块的重建模块（RM）组成。详细信息请参见** Sec. 3.1。

展示了相关网络的不同方法并比较了它们的差异 [1, 9–11, 30]。许多基于渐进细化策略的方法被提出。Yu 等人 [31] 使用预测的低分辨率（LR）alpha matte 作为高分辨率（HR）地图的指导。在 BASNet [32] 中，初始结果通过额外的细化网络进行修正。CRM [33] 通过与细化目标连续对齐特征图来聚合详细特征。在 ICNet [3] 中，原始图像也被缩小并在不同阶段添加到解码器输出中进行细化。ICEG [34] 中的生成器和检测器通过交互迭代进化，以获得更好的隐藏物体检测（COD）分割结果。除了图像和真实标签（GT），现有方法还使用辅助信息。例如，Tang 等人 [35] 在边界上裁剪块以进一步细化。在用于图像超分辨率的 LapSRN 网络 [36, 37] 中，也生成了拉普拉斯金字塔以帮助高分辨率图像重建。尽管这些方法成功地采用了细化策略以获得更好的结果，但模型并未被引导专注于某些特定区域，这在 DIS 任务中是一个问题。因此，我们在外部参考中引入了梯度监督，以引导对具有更丰富细节区域敏感的特征。

3 方法

3.1 概览

如 Fig. 2(d) 所示，我们提出的 BiRefNet 不同于以往的 DIS 方法。一方面，我们的 BiRefNet 明确地将高分辨率数据上的 DIS 任务分解为两个模块，即定位模块（LM）和重建模块（RM）。另一方面，我们没有直接将源图像 [2] 或先验信息 [10] 添加到输入中，而是在 RM 中采用我们提出的双边参考（BiRef），充分利用原始尺度的源图像和梯度先验。我们 BiRefNet 的完整框架如 Fig. 3所示。

3.2 定位模块

对于一个批量的高分辨率图像输入 $\mathcal{I} \in \mathbb{R}^{N\times 3\times H\times W}$，使用变换器编码器 [18] 提取不同阶段的特征，即 $\mathcal{F}_1^e, \mathcal{F}_2^e, \mathcal{F}_3^e, \mathcal{F}^e$，它们的分辨率分别为 $\{[\frac{H}{k}, \frac{W}{k}], k = 4, 8, 16, 32\}$。前三个阶段的特征 $\{\mathcal{F}_i^e\}_{i=1}^3$ 通过侧向连接（1×1 卷积层）传递给相应的解码器阶段。同时，它们在最后一个编码器块中被堆叠和串联以生成 $\mathcal{F}^e$。

编码器输出特征 $\mathcal{F}^e$ 然后被送入一个分类模块，其中 $\mathcal{F}^e$ 经过全局平均池化层和全连接层进行分类，以获取更好的语义表示用于定位。高分辨率特征在瓶颈处被压缩。为了扩大感受野以覆盖大物体的特征并同时专注于高精度的局部特征 [38]，这对涉及高分辨率任务非常重要，我们在这里使用 ASPP 模块 [39] 进行多上下文融合。$\mathcal{F}^e$ 被压缩成 $\mathcal{F}^d$，以便传递到重建模块。

3.3 重构模块

感受野（RF）的设置一直是高分辨率（HR）分割的一个挑战。小感受野导致上下文信息不足，无法在大背景中准确定位目标，而大感受野则常常导致在细节区域特征提取不足。为实现平衡，我们在每个双边参考块（BiRef 块）中提出了重建块（RB）来替代普通的残差块。在 RB 中，我们采用了具有层次感受野的可变形卷积 [40]（即 1×1、3×3、7×7）和自适应平均池化层，以提取不同尺度感受野的特征。这些由不同感受野提取的特征被拼接为 $\mathcal{F}_i^\theta$，然后经过 1×1 卷积层和批归一化层，生成重建模块的输出特征 $\mathcal{F}_i^{d'}$。在重建模块中，压缩后的特征 $\mathcal{F}^d$ 被输入到 BiRef 块中，生成特征 $\mathcal{F}_3^d$。利用 $\mathcal{F}_3^l$，第一个 BiRef 块预测粗略图，然后通

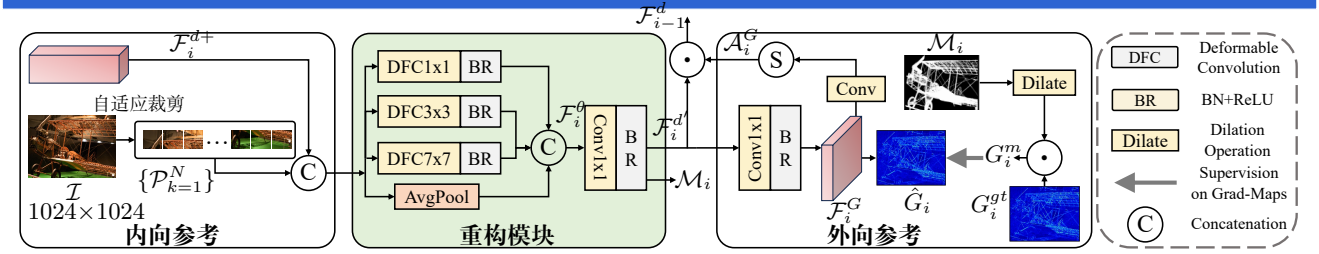

图 4 提出的双边参考块的流程。原始尺度的源图像与解码器特征结合作为内部参考，输入重建块中，采用具有层次感受野的可变形卷积。聚合的特征然后用于预测外部参考中的梯度图。梯度感知特征然后被转换为注意力图，作用于原始特征。

过后续的 BiRef 块重建成更高分辨率版本。参考 [41]，每个 BiRef 块的输出特征 $\mathcal{F}_i^d$ 在每个阶段与其侧向特征 $\mathcal{F}_i^l$ 相加，即 $\{\mathcal{F}_i^{d+} = \text{Upsample}\uparrow(\mathcal{F}_i^d + \mathcal{F}_i^l), i = 3, 2, 1\}$。同时，所有 BiRef 块通过多阶段监督生成中间预测 $\{\mathcal{M}_i\}_{i=3}^1$，分辨率依次递增。最后，最后的解码特征 $\mathcal{F}_1^{d+}$ 经过 1×1 卷积层，得到最终的预测图 $\mathcal{M} \in \mathbb{R}^{N\times 1\times H\times W}$。

3.4 双边参考

在高分辨率二值图像分割（DIS）中，高分辨率（HR）训练图像对于深度模型学习细节和实现高精度分割至关重要。然而，大多数分割模型遵循以往的工作 [1, 41]，采用编码器-解码器结构设计网络架构，分别进行下采样和上采样。此外，由于输入图像尺寸较大，聚焦目标物体变得更加困难。为了解决这两个主要问题，我们提出了*双边参考*，由内部参考（InRef）和外部参考（OutRef）组成，如 Fig. 4所示。内部参考和外部参考分别起到补充高分辨率信息和吸引注意力到细节密集区域的作用。

在 InRef 中，具有原始高分辨率的图像 $\mathcal{I}$ 被裁剪为与相应解码器阶段的输出特征一致大小的补丁 $\{\mathcal{P}_{k=1}^N\}$。这些补丁与原始特征 $\mathcal{F}_i^{d+}$ 叠加后输入重建模块（RM）。现有使用类似技术的方法，要么仅在最后的解码阶段添加 $\mathcal{I}$ [11]，要么将 $\mathcal{I}$ 调整尺寸，使其适用于低分辨率的原始特征。我们的内部参考通过自适应裁剪避免了这两个问题，并在每个阶段提供必要的高分辨率信息。

在 OutRef 中，我们使用梯度标签来吸引更多注意力到具有丰富梯度信息的区域，这对于细小结构的分割至关重要。首先，我们提取输入图像的梯度图作为 G_i^{gt}。同时，$\mathcal{F}_i^\theta$ 被用来生成特征 $\mathcal{F}_i^G$ 以产生预测的梯度图 $\hat{G}_i$。通过这种梯度监督，$\mathcal{F}_i^G$ 对梯度变得敏感。它通过一个卷积层和一个 sigmoid 层，用于生成梯度参考注意力 $\mathcal{A}_i^G$，然后与 $\mathcal{F}_i^{d'}$ 相乘，生成 BiRef 块的输出 $\mathcal{F}_{i-1}^d$。

考虑到背景可能包含大量梯度信息的非目标噪声，我们采用了一种掩蔽策略来减轻非目标区域的影响。我们对中间预测结果 $\mathcal{M}_i$ 进行形态学操作，并使用膨胀后的 $\mathcal{M}_i$ 作为掩膜。该掩膜用于与梯度图 G_i^{gt} 相乘，以生成 G_i^m，从而去除掩膜区域外的梯度。

3.5 目标函数

在高分辨率分割任务中，仅使用像素级监督（BCE 损失）通常会导致高分辨率数据中细节结构信息的劣化。受 [32] 中使用混合损失取得优秀结果的启发，我们将 BCE、IoU、SSIM 和 CE 损失结合使用，分别在像素、区域、边界和语义级别进行监督。

最终的目标函数是上述损失的加权组合，可以公式化为：

$$\begin{aligned} L &= L_{\text{像素}} + L_{\text{区域}} + L_{\text{边界}} + L_{\text{语义}} \\ &= \lambda_1 L_{\text{BCE}} + \lambda_2 L_{\text{IoU}} + \lambda_3 L_{\text{SSIM}} + \lambda_4 L_{\text{CE}}, \end{aligned} \tag{1}$$

其中，$\lambda_1, \lambda_2, \lambda_3$ 和 λ_4 分别设置为 30, 0.5, 10 和 5，以在训练开始时保持所有损失在相同的量级。最终的目标函数包括二元交叉熵（BCE）损失、交并比（IoU）损失、结构相似性指数（SSIM）损失和交叉熵（CE）损失。损失函数的完整定义如下所示。

- **二元交叉熵损失：**像素感知监督，用于生成二值图的像素级监督：

$$L_{BCE} = -\sum_{(i,j)} [G(i,j)\log(M(i,j)) + (1-G(i,j))\log(1-M(i,j))], \tag{2}$$

  其中 $G(i,j)$ 和 $\mathcal{M}(i,j)$ 分别表示像素 (i,j) 处的真实值和二值化预测图的值。

- **交并比损失：**区域感知监督，用于增强二值图预测：

$$L_{IoU} = 1 - \frac{\sum_{r=1}^{H}\sum_{c=1}^{W} M(i,j)G(i,j)}{\sum_{r=1}^{H}\sum_{c=1}^{W} [M(i,j) + G(i,j) - M(i,j)G(i,j)]}. \tag{3}$$

- **结构相似性指数损失：**在边界感知监督中，用于提高边界部分的准确性。给定 GT 地图 G 和预测地图 $\mathcal{M}$，$\mathrm{y} = \{y_j : j = 1, ..., N^2\}$ 和 $\mathrm{x} = \{x_j : j = 1, ..., N^2\}$ 分别表示从 G 和 $\mathcal{M}$ 中提取的两个对应的 $N \times N$ 区域的像素值。$SSIM(\mathrm{x}, \mathrm{y})$ 定义如下：

$$L_{SSIM} = 1 - \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{4}$$

其中，μ_x、μ_y 分别是 x 和 y 的均值，σ_x、σ_y 分别是它们的标准差，σ_{xy} 是它们的协方差。C 是为了避免除以零。

- **交叉熵损失：**语义感知监督，用于学习更好的语义表示：

$$L_{CE} = -\sum_{c=1}^{N} y_{o,c} \log(p_{o,c}), \tag{5}$$

这里，N 是类别数，$y_{o,c}$ 表示类别 c 是否是观测 o 的正确分类，$p_{o,c}$ 表示预测的观测 o 属于类别 c 的概率。

3.6 根据 DIS 任务量身定制的训练策略

由于在高分辨率数据上训练模型的高成本，我们已经探索了针对高分辨率分割任务的训练技巧，以提高性能并降低训练成本。

首先，我们发现我们的模型在 DIS5K 数据集上相对快速地收敛于目标定位和粗略结构的分割（使用 F-measure [42]、S-measure [12] 等指标，例如在 200 个 epochs 内）。然而，在分割精细部分的性能仍然在经过非常长时间的训练后继续提升（例如在 400 个 epochs 内），这可以通过诸如 F_β^ω 和 HCE_γ 等指标来反映。其次，虽然长时间训练可以轻松达到在结构和边缘上的优秀结果，但它消耗了太多计算资源；我们发现多阶段监督能显著加速学习，特别是在细节分割方面，并使模型在仅使用 30% 的训练 epochs 情况下达到类似的性能。第三，我们还发现，仅使用区域级别的损失进行微调可以轻松改善预测结果的二值化，并且提升那些与实际使用更为接近的度量分数（例如 F_β^ω、E_ϕ^m、HCE）。最后，我们在骨干网络上使用了上下文特征融合和图像金字塔输入，这些是常用的技巧，用于处理深度模型中的高分辨率图像。在实验中，这两种对骨干网络的修改在 DIS 和类似的高分辨率分割任务中都取得了整体改进。

表 1 多阶段监督方法在加速训练和减少训练轮数方面的定量剔除研究。

设置		DIS-VD					
MSS	Epoch	$F_\beta^x \uparrow$	$F_\beta^\omega \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$E_\phi^m \uparrow$	$HCE_\gamma \downarrow$
	200	.875	.848	.041	.886	.914	1207
	400	.897	.863	.036	.905	.937	1039
✓	200	.892	.858	.037	.901	.932	1043

如表. 1所示，我们展示了训练轮数和多阶段监督的有效性。结果显示，我们的 BiRefNet 在训练 200 个轮次后可以达到相对良好的结果。持续训练至 400 个轮次可以在衡量结构信息的指标（例如 F_β^x，S_m）中略微提升，同时在衡量细节信息的指标（例如 HCE_γ）中带来更大的改善。

尽管简单的长时间训练可以达到更好的结果，但改善幅度相对较小，考虑到在高分辨率数据上的高计算成本。我们研究了多阶段监督（MSS），这是二值分割工作中广泛使用的训练策略 [9, 43]。与这些工作中用于提高精度的 MSS 不同，我们的 MSS 在加速训练收敛方面发挥作用。如表. 1中的结果所示，我们的 BiRefNet 使用 MSS 训练 200 个轮次可以达到与使用 400 个轮次训练相似的性能。MSS 成功地将训练时间减少了一半，并可以用于进一步的高分辨率分割任务，以实现更高效的训练。

4 实验

4.1 数据集

训练集。对于 DIS 任务，我们按照 [9–11] 的方法，使用 DIS5K-TR 作为实验中的训练集。对于 HRSOD 任务，我们按照 [14] 的方法，设置不同的 HRSOD、UHRSD 和 DUTS 组合作为训练集。对于 COD 任务，我们按照 [8, 23] 的方法，使用 CAMO-TR 和 COD10K-TR 中的隐蔽样本作为训练集。

测试集。为了对我们的 BiRefNet 进行全面评估，我们在 DIS5K 的所有测试集（DIS-TE1、DIS-TE2、DIS-TE3 和 DIS-TE4）上进行了测试。此外，我们还在 HRSOD 测试集（DAVIS-S [13]、HRSOD-TE [13] 和 UHRSD-TE [14]）以及 COD 测试集（CAMO-TE [44]、COD10K-TE [19] 和 NC4K [45]）上对 BiRefNet 进行了评估。低分辨率的 SOD 测试集（DUTS-TE [46] 和 DUT-OMRON [47]）也被用于补充实验。

4.2 评估协议

为了进行全面评估，我们采用了广泛使用的指标，即 S-measure [12]（S_m）、最大/平均/加权 F-measure [42]（$F_\beta^x / F_\beta^m / F_\beta^\omega$）、最大/平均 E-measure [48]（$E_\xi^x / E_\xi^m$）、平均绝对误差 (MAE) 以及松弛 HCE [9]（HCE_γ）来评估性能。以下是这些指标的详细描述。

- S-measure [12]（结构度量，S_α）是一种衡量显著图与其对应 GT 图之间结构相似度的指标。S_α 的评估速度很快，无需二值化。S_α-measure 的计算公式如下：

$$S_\alpha = \alpha \cdot S_o + (1 - \alpha) \cdot S_r, \tag{6}$$

其中，S_o 和 S_r 分别表示对象感知和区域感知的结构相似度，α 默认设置为 0.5，参考 Fan et al. 在 [12] 中的建议。

- F-measure [42]（F_β）用于评估精度和召回率的加权调和平均值。显著性图的输出以不同的阈值进行二值化，以获得一组二值显著性预测。预测的显著性

图和 GT 图进行比较，以获得精度和召回率值。F-measure 的计算公式如下：

$$F_\beta = \frac{(1+\beta^2)\cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall}, \quad (7)$$

其中 β^2 设置为 0.3，以强调精度优于召回率，遵循 [49]。使用整个数据集的最佳阈值获得的最大 F-measure 得分记为 F_β^x。

- E-measure [48]（增强对齐度量，E_ξ）旨在作为一种感知度量，用于评估预测图和 GT 图在局部和全局上的相似性。E-measure 定义为：

$$E_\xi = \frac{1}{WH}\sum_{x=1}^{W}\sum_{y=1}^{H}\phi_\xi(x,y), \quad (8)$$

where ϕ_ξ 表示增强对齐矩阵。类似于 F-measure，我们同样采用最大 E-measure（E_ξ^x）和平均 E-measure（E_ξ^m）作为我们的评估指标。

- MAE（平均绝对误差，ϵ）是一个简单的像素级评估指标，用于测量非二值化预测结果 $\mathcal{M}$ 和 GT 图像 G 之间的绝对差异。它的定义如下：

$$\epsilon = \frac{1}{WH}\sum_{x=1}^{W}\sum_{y=1}^{H}|\hat{Y}(x,y) - \mathrm{G}(x,y)|. \quad (9)$$

- HCE_γ [9]（人工纠正代价，HCE）是一种新提出的评估指标，旨在评估在实际应用中满足特定精度要求所需的人力修正工作量。具体而言，HCE 通过大约需要的鼠标点击次数来量化。在实际应用中，可以容忍较小的预测错误。因此，引入了松弛 HCE（HCE_γ），其中 γ 表示容忍度。在实验中，我们使用 HCE_5（带有 $\gamma = 5$ 的松弛 HCE）以保持与原始论文 [9] 的一致性，该论文详细描述了 HCE_γ。

4.3 实现细节

抱歉，我理解了。以下是您提供的信息的翻译：

所有图像均调整大小为 1024×1024 用于训练和测试。生成的分割地图经过重新调整大小（即双线性插值），以便与相应的 GT 地图的原始大小进行评估。训练过程中唯一使用的数据增强技术是水平翻转。类别数 C 设置为 219，与 DIS-TR 中的定义一致。我们提出的 BiRefNet 分别对 DIS/HRSOD/COD 任务进行 600/150/150 个 epoch 的训练，使用 Adam 优化器 [50]。在最后的 20 个 epoch 中，模型使用 IoU 损失进行微调。初始学习率分别设置为 10^{-4}（DIS 任务）和 10^{-5}（HRSOD 和 COD 任务）。我们使用 PyTorch [51] 在八个 NVIDIA A100 GPU 上进行训练，每个 GPU 的批量大小为 $N = 4$。

4.4 消融研究

我们研究了引入到我们模型中的每个组成部分（即 RM 和 BiRef）以及实用策略（即 CFF、IPT 和 RLFT），并进行了它们对改善 DIS 结果贡献的调查。关于每个模块和策略的定量结果分别显示在表. 2和3中。

表 2 对提出的 BiRefNet 中的各个组成部分进行的定量消融研究，包括重建模块（RM）、内向参考（InRef）、外向参考（OutRef）以及它们的组合。

模块			DIS-VD					
RM	InRef	OutRef	$F_\beta^x\uparrow$	$F_\beta^\omega\uparrow$	$\mathcal{M}\downarrow$	$S_m\uparrow$	$E_\phi^m\uparrow$	$HCE_\gamma\downarrow$
			.837	.785	.056	.845	.887	1204
✓			.855	.831	.048	.865	.895	1167
	✓		.848	.825	.050	.857	.903	1152
✓	✓		.869	.834	.041	.886	.912	1093
✓		✓	.863	.831	.042	.891	.918	1106
	✓	✓	.861	.839	.044	.881	.911	1114
✓	✓	✓	.889	.851	.038	.900	.924	1065

Baseline. We provide a simple but strong encoder-decoder network as the baseline for the DIS task. To capture better hierarchical features on various scales, we chose the Swin transformer large [18] as our default backbone network. Then, to obtain a better semantic representation in the DIS task, we divided the images in DIS-TR into 219 classes according to their label names and added an auxiliary classification head at the end of the encoder. In the decoder of the baseline network, each decoder block is made up of two residual blocks [17]. All stages of the encoder and decoder are connected with an 1×1 convolution, except the deepest stage, where an ASPP [39] block is used for connectivity. With this setup, our baseline network has outperformed existing DIS models in most metrics, as shown in 表. 2 and 表. 4.

基准模型。我们提供了一个简单但强大的编码器-解码器网络作为 DIS 任务的基准线。为了在不同尺度上捕捉更好的层次特征，我们选择了 Swin Transformer Large 作为默认的骨干网络 [18]。然后，为了在 DIS 任务中获得更好的语义表示，我们根据其标签名称将 DIS-TR 中的图像划分为 219 个类，并在编码器的末端添加了一个辅助分类头。在基准网络的解码器中，每个解码器块由两个残差块 [17] 组成。编码器和解码器的所有阶段都通过 1×1 卷积连接，除了最深的阶段，这里使用了 ASPP [39] 块进行连接。通过这个设置，我们的基准网络在大多数指标上都优于现有的 DIS 模型，如表. 2和表. 4所示。

表 3 训练高分辨率分割的实用策略效果评估，包括上下文特征融合（CFF）、图像金字塔输入（IPT）、区域损失微调（RLFT）及其组合。这些结果由我们的最终模型得出。

模块			DIS-VD					
CFF	IPT	RLFT	$F_\beta^x \uparrow$	$F_\beta^\omega \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$E_\phi^m \uparrow$	$HCE_\gamma \downarrow$
			.889	.851	.038	.900	.924	1065
✓			.893	.856	.038	.904	.928	1054
	✓		.895	.857	.037	.904	.927	1051
		✓	.890	.861	.036	.899	.932	1043
✓	✓	✓	.897	.863	.036	.905	.937	1039

重建模块。如 表. 2 所示，我们的模型在提出的重建模块（RM）的帮助下总体上有所改进。RM 在高分辨率特征上提供了多尺度感受野，用于捕捉局部细节和整体语义信息。它在几乎不增加额外计算成本的情况下，相对提升了约 2.2% 的 F_β^x 指标。

双向参考。我们分别研究了双向参考中的内向参考 (InRef，使用源图像) 和外向参考 (OutRef，使用梯度标签) 的有效性。InRef 全局补充了无损的高分辨率信息，而 OutRef 则更加关注细节部分，以在这些区域实现更高的精度。如表. 2所示，它们共同使得我们的模型 BiRefNet 相对提高了 2.9% 的 F_β^x。RM 和 BiRef 的组合使得相对提高了 6.2% 的 F_β^x。

训练策略。如表. 3所示，所提出的策略从不同角度提升了性能。CCF 和 IPT 提高了整体性能，而 RLFT 则特别提高了边缘细节的精度，这在诸如 F_β^ω 和 HCE_γ 等指标中得到了体现。

4.5 SOTA 比较

为了验证我们方法的普适性，我们在四个任务上进行了广泛的实验，即高分辨率二分图像分割（DIS）、高分辨率显著物体检测（HRSOD）、隐蔽物体检测（COD）和显著物体检测（SOD）。我们将我们提出的模型（BiRefNet）与所有最新的任务特定模型在现有基准数据集上进行了比较 [8, 9, 13, 14, 44–47]。

定量结果。如表. 4所示，这是提出的 BiRefNet 与之前最先进方法的定量比较。我们的 BiRefNet 在广泛使用的评估指标上表现优异，超过了所有先前的方法。DIS-TE1∼DIS-TE4 的复杂性按升序排列。结构相似度指标（例如 S_α, E_ϕ^x）更关注全局信息。像素级指标（如 MAE (M)）强调细节的精确性。基于平均值的指标（例如 E_ϕ^m, F_ϕ^m）更符合实际应用中阈值化地图的要求。正如表. 4所示，我们的 BiRefNet 不仅在全局形状的准确性上优于先前的方法，而且在像素级的细节上也表现出色。特别值得注意的是，在更贴近实际应用的指标上，我们的结果更好。

Additionally, our BiRefNet outperforms existing task-specific models on the HRSOD and COD tasks. As shown in 表. 5, BiRefNet achieved much higher accuracy on both high-resolution and low-resolution SOD benchmarks. Compared with the previous SOTA method [14], our BiRefNet achieved an average improvement of 2.0% S_m. Furthermore, as shown in 表. 6, in the COD task, BiRefNet also shows a much better performance compared to the previous SOTA models, with an average improvement of 5.6% S_m on the three widely used COD benchmarks. These results show the remarkable generalization ability of our BiRefNet to similar HR tasks. 此外，我们的 BiRefNet 在 HRSOD 和 COD 任务上也优于现有的任务特定模型。如表. 5所示，BiRefNet 在高分辨率和低分辨率显著物体检测基准上都取得了更高的准确性。与之前的最先进方法 [14] 相比，我们的 BiRefNet 在 S_m 指标上平均提高了 2.0%。此外，如表. 6所示，在 COD 任务中，BiRefNet 相比之前的最先进模型也表现出色，在三个广泛使用的 COD 基准上，S_m 指标平均提高了 5.6%。这些结果显示了我们 BiRefNet 在类似高分辨率任务上的显著泛化能力。

此外，我们的 BiRefNet 在 HRSOD 和 COD 任务上也优于现有的任务特定模型。如表. 5所示，BiRefNet 在高分辨率和低分辨率 SOD 基准测试中均取得了更高的准确性。与之前的 SOTA 方法 [14] 相比，我们的 BiRefNet 在 S_m 指标上平均提高了 2.0%。此外，如表. 6所示，在 COD 任务中，BiRefNet 相比于之前的 SOTA 模型也表现更好，在三个广泛使用的 COD 基准测试中 S_m 指标平均提高了 5.6%。这些结果显示了我们 BiRefNet 在类似高分辨率任务上的显著泛化能力。

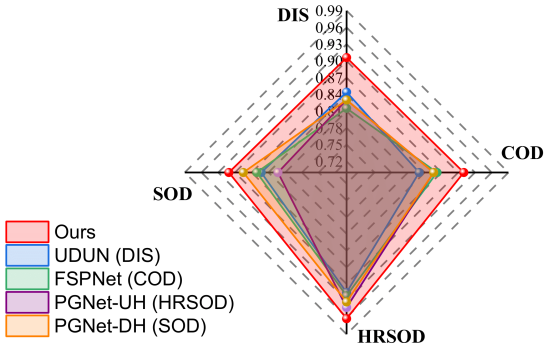

图 7 提出的 BiRefNet 与最佳任务特定模型的定量比较。这里使用 S-measure [12] 进行比较。UDUN [11]、FSPNet [23]、PGNet-UH [14] 和 PGNet-DH [14] 分别是目前在 DIS、COD、HRSOD 和 SOD 任务上的最佳模型。

定性结果。如 Fig. 5所示，这是由最具竞争力的现有 DIS 模型和提出的 BiRefNet 生成的分割图。结果表明，我们提供了所有测试集和一个验证集的样本。从两个方面来看，BiRefNet 优于之前的 DIS 方法，即目标物体的位置和物体细节的更精确分割。例如，在 DIS-TE4 和 DIS-TE2 的样本中，有邻近的干扰物吸引了其他模型的注意，产生了误报。相反，我们的 BiRefNet 消除了干扰物并准确分割了目标。在 DIS-TE3 和 DIS-VD 的样本中，BiRefNet

表 4 DIS5K 上我们的 BiRefNet 与现有方法的定量比较。“↑” (“↓”) 表示更高（更低）更好。我们使用来自 [11] 的结果，其中所有方法均采用 1024×1024 的输入。

方法	DIS-TE1 (500)						DIS-TE2 (500)						DIS-TE3 (500)					
	$F_\beta^x\uparrow$	$F_\beta^\omega\uparrow$	$\mathcal{M}\downarrow$	$S_m\uparrow$	$E_\phi^m\uparrow$	$HCE_\gamma\downarrow$	$F_\beta^x\uparrow$	$F_\beta^\omega\uparrow$	$\mathcal{M}\downarrow$	$S_m\uparrow$	$E_\phi^m\uparrow$	$HCE_\gamma\downarrow$	$F_\beta^x\uparrow$	$F_\beta^\omega\uparrow$	$\mathcal{M}\downarrow$	$S_m\uparrow$	$E_\phi^m\uparrow$	$HCE_\gamma\downarrow$
$BASNet_{19}$ [32]	.663	.577	.105	.741	.756	155	.738	.653	.096	.781	.808	341	.790	.714	.080	.816	.848	681
U^2Net_{20} [52]	.701	.601	.085	.762	.783	165	.768	.676	.083	.798	.825	367	.813	.721	.073	.823	.856	738
$HRNet_{20}$ [53]	.668	.579	.088	.742	.797	262	.747	.664	.087	.784	.840	555	.784	.700	.080	.805	.869	1049
$PGNet_{22}$ [14]	.754	.680	.067	.800	.848	162	.807	.743	.065	.833	.880	375	.843	.785	.056	.844	.911	797
$IS\text{-}Net_{22}$ [9]	.740	.662	.074	.787	.820	149	.799	.728	.070	.823	.858	340	.830	.758	.064	.836	.883	687
$FP\text{-}DIS_{23}$ [10]	.784	.713	.060	.821	.860	160	.827	.767	.059	.845	.893	373	.868	.811	.049	.871	.922	780
$UDUN_{23}$ [11]	.784	.720	.059	.817	.864	140	.829	.768	.058	.843	.886	325	.865	.809	.050	.865	.917	658
BiRefNet	.860	.819	.037	.885	.911	106	.894	.857	.036	.900	.930	266	.925	.893	.028	.919	.955	569
$BiRefNet_{SwinB}$	.857	.819	.038	.884	.912	110	.890	.854	.037	.898	.930	275	.919	.886	.030	.915	.953	597
$BiRefNet_{SwinT}$	.823	.774	.048	.855	.887	117	.862	.821	.046	.877	.912	290	.899	.860	.036	.897	.942	627
$BiRefNet_{PVTv2b2}$	.839	.796	.042	.870	.903	111	.881	.842	.040	.888	.925	280	.903	.866	.036	.901	.941	614

方法	DIS-TE4 (500)						DIS-TE (1-4) (2,000)						DIS-VD (470)					
	$F_\beta^x\uparrow$	$F_\beta^\omega\uparrow$	$\mathcal{M}\downarrow$	$S_m\uparrow$	$E_\phi^m\uparrow$	$HCE_\gamma\downarrow$	$F_\beta^x\uparrow$	$F_\beta^\omega\uparrow$	$\mathcal{M}\downarrow$	$S_m\uparrow$	$E_\phi^m\uparrow$	$HCE_\gamma\downarrow$	$F_\beta^x\uparrow$	$F_\beta^\omega\uparrow$	$\mathcal{M}\downarrow$	$S_m\uparrow$	$E_\phi^m\uparrow$	$HCE_\gamma\downarrow$
$BASNet_{19}$ [32]	.785	.713	.087	.806	.844	2852	.744	.664	.092	.786	.814	1007	.737	.656	.094	.781	.809	1132
U^2Net_{20} [52]	.800	.707	.085	.814	.837	2898	.771	.676	.082	.799	.825	1042	.753	.656	.089	.785	.809	1139
$HRNet_{20}$ [53]	.772	.687	.092	.792	.854	3864	.743	.658	.087	.781	.840	1432	.726	.641	.095	.767	.824	1560
$PGNet_{22}$ [14]	.831	.774	.065	.841	.899	3361	.809	.746	.063	.830	.885	1173	.798	.733	.067	.824	.879	1326
$IS\text{-}Net_{22}$ [9]	.827	.753	.072	.830	.870	2888	.799	.726	.070	.819	.858	1016	.791	.717	.074	.813	.856	1116
$FP\text{-}DIS_{23}$ [10]	.846	.788	.061	.852	.906	3347	.831	.770	.047	.847	.895	1165	.823	.763	.062	.843	.891	1309
$UDUN_{23}$ [11]	.846	.792	.059	.849	.901	2785	.831	.772	.057	.844	.892	977	.823	.763	.059	.838	.892	1097
BiRefNet	.904	.864	.039	.900	.939	2723	.896	.858	.035	.901	.934	916	.891	.854	.038	.898	.931	989
$BiRefNet_{SwinB}$	.899	.860	.040	.895	.938	2836	.891	.855	.036	.898	.933	954	.881	.844	.039	.890	.925	1029
$BiRefNet_{SwinT}$	.880	.834	.049	.878	.925	2888	.866	.822	.045	.877	.916	980	.862	.819	.045	.874	.917	1070
$BiRefNet_{PVTv2b2}$	.890	.846	.045	.886	.929	2871	.878	.838	.041	.886	.925	969	.868	.827	.044	.880	.919	1073

在精确分割细节丰富的区域方面表现出色。与之前的方法相比，我们的 BiRefNet 能够清晰地分割细长的形状和曲线边缘。

我们还提供了 COD 任务的定性比较。如 Fig. 6所示，显示了具有不同挑战的难样本。例如，在被遮挡的青蛙那一行，青蛙的区域被覆盖它的树枝分隔开，而我们的 BiRefNet 能够准确分割出几乎与 GT 图一致的散落碎片。相比之下，其他方法的结果中很难找到所有碎片，更不用说提供精确的分割图了。对于微小和细长的物体，BiRefNet 表现出更好的目标识别能力。我们的 BiRefNet 在找到多个隐蔽物体方面也显示出优越性。

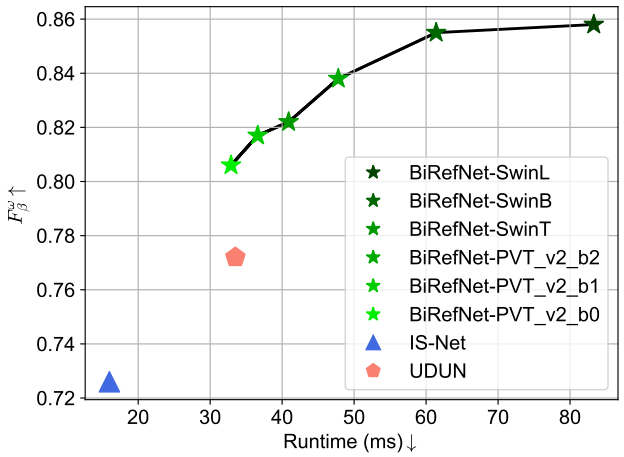

图 8 BiRefNet 与现有 DIS 方法在效率、规模、复杂性和性能方面的比较。

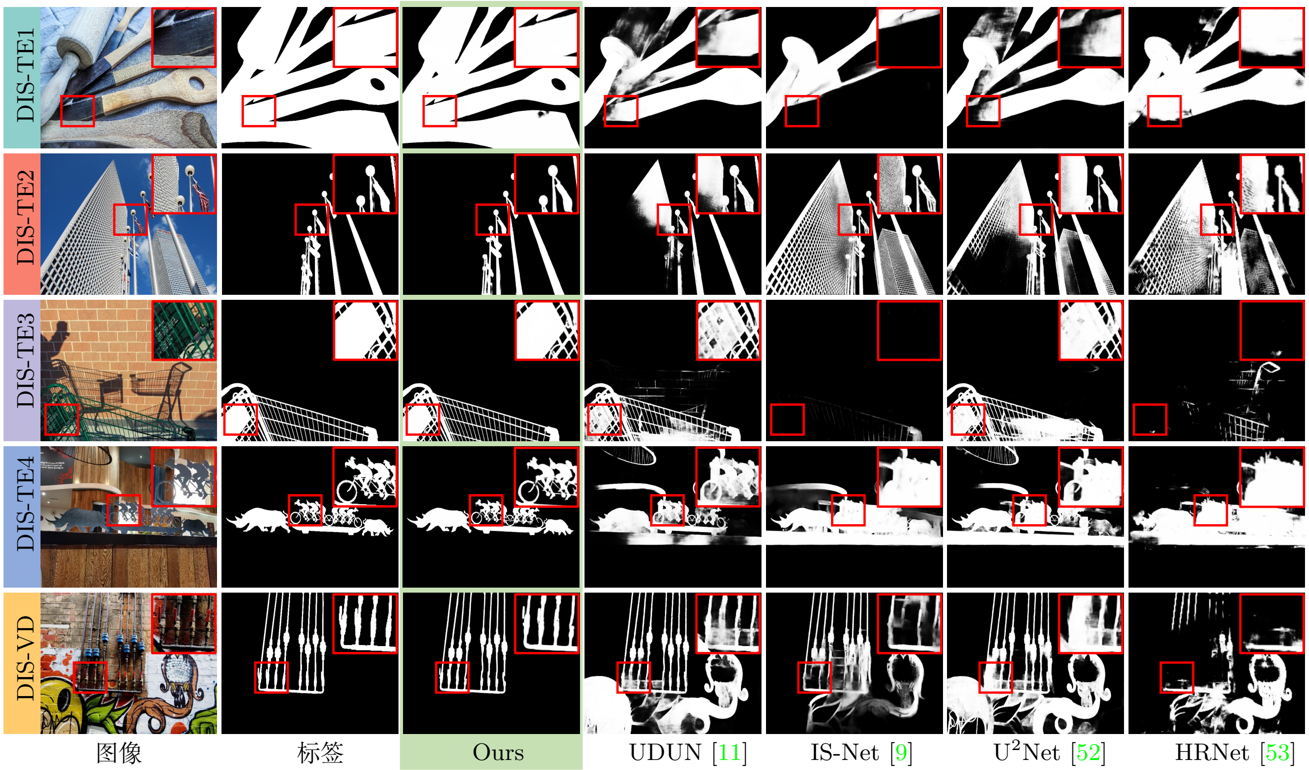

图 5 所提出的 BiRefNet 在 DIS5K 数据集上与先前方法的定性比较。先前方法的结果来源于文献 [11]，其中所有模型均使用 1024×1024 大小的图像进行训练。放大以获得更好的视图。

表 5 我们的 BiRefNet 与高分辨率和低分辨率显著物体检测数据集中最先进方法的定量比较。这里 TR 表示训练集。为了进行公平比较，我们使用不同的训练集组合来训练我们的 BiRefNet，其中 1、2 和 3 分别代表 DUTS [46]、HRSOD [13] 和 UHRSD [14]。

测试集		高分辨率基准												低分辨率基准							
		DAVIS-S (92)				HRSOD-TE (400)				UHRSD-TE (988)				DUTS-TE (5,019)				DUT-OMRON(5,168)			
方法	TR	S_m ↑	F_β^x ↑	E_ϕ^m ↑	$\mathcal{M}$ ↓	S_m ↑	F_β^x ↑	E_ϕ^m ↑	$\mathcal{M}$ ↓	S_m ↑	F_β^x ↑	E_ϕ^m ↑	$\mathcal{M}$ ↓	S_m ↑	F_β^x ↑	E_ϕ^m ↑	$\mathcal{M}$ ↓	S_m ↑	F_β^x ↑	E_ϕ^m ↑	$\mathcal{M}$ ↓
LDF_{20} [26]	1	.922	.911	.947	.019	.904	.904	.919	.032	.888	.913	.891	.047	.892	.898	.910	.034	.838	.820	.873	.051
$HRSOD_{19}$ [13]	1,2	.876	.899	.955	.026	.896	.905	.934	.030	-	-	-	-	.824	.835	.885	.050	.762	.743	.831	.065
DHQ_{21} [16]	1,2	.920	.938	.947	.012	.920	.922	.947	.022	.900	.911	.905	.039	.894	.900	.919	.031	.836	.820	.873	.045
$PGNet_{22}$ [14]	1	.935	.936	.947	.015	.930	.931	.944	.021	.912	.931	.904	.037	.911	.917	.922	.027	.855	.835	.887	.045
$PGNet_{22}$ [14]	1,2	.948	.950	.975	.012	.935	.937	.946	.020	.912	.935	.905	.036	.912	.919	.925	.028	.858	.835	.887	.046
$PGNet_{22}$ [14]	2,3	.954	.957	.979	.010	.938	.945	.946	.020	.935	.949	.916	.026	.859	.871	.897	.038	.786	.772	.884	.058
BiRefNet	1	.967	.966	.984	.008	.957	.958	.972	.014	.931	.933	.943	.030	.939	.937	.958	.019	.868	.813	.878	.040
BiRefNet	1,2	.973	.976	.990	.006	.962	.963	.976	.011	.937	.942	.951	.024	.938	.935	.960	.018	.868	.818	.882	.040
BiRefNet	1,3	.975	.977	.989	.006	.959	.958	.972	.014	.952	.960	.965	.019	.942	.942	.961	.018	.881	.837	.896	.036
BiRefNet	2,3	.976	.980	.990	.006	.956	.953	.967	.016	.952	.958	.964	.019	.933	.928	.954	.020	.864	.810	.879	.040
BiRefNet	1,2,3	.975	.979	.989	.006	.962	.961	.973	.013	.957	.963	.969	.016	.944	.943	.962	.018	.882	.839	.896	.038

效率和复杂性比较。我们为 BiRefNet 配备了不同的骨干网络，以获得不同规模的模型。进一步测试它们的运行时间、参数数量、MACs 和性能，以提供与其他方法的全面比较。首先，我们在表. 7中提供了定量比较。最大的 BiRefNet 的 FPS 可以超过 10，这在大多数实际应用中是可以接受的。我们还使用 PyTorch 2.0 [51] 对 $BiRefNet_{SwinL}$ 进行编译，得到的版本（$BiRefNet_{SwinL\_cp}$）使推理速度加快了 13%。此外，我们在 Fig. 8中绘制了每个模型的性能和运行时间，以更清晰地展示。BiRefNet 与不同骨干网络在 DIS-TEs 和 DIS-

表 6 与最近方法的比较。正如所见，BiRefNet 的表现远优于先前的方法。

方法	CAMO (250)						COD10K (2,026)						NC4K (4,121)					
	$S_m \uparrow$	$F_\beta^\omega \uparrow$	$F_\beta^m \uparrow$	$E_\phi^m \uparrow$	$E_\phi^x \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$F_\beta^\omega \uparrow$	$F_\beta^m \uparrow$	$E_\phi^m \uparrow$	$E_\phi^x \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$F_\beta^\omega \uparrow$	$F_\beta^m \uparrow$	$E_\phi^m \uparrow$	$E_\phi^x \uparrow$	$\mathcal{M} \downarrow$
$SINet_{20}$ [19]	.751	.606	.675	.771	.831	.100	.771	.551	.634	.806	.868	.051	.808	.723	.769	.871	.883	.058
$BGNet_{22}$ [21]	.812	.749	.789	.870	.882	.073	.831	.722	.753	.901	.911	.033	.851	.788	.820	.907	.916	.044
$SegMaR_{22}$ [54]	.815	.753	.795	.874	.884	.071	.833	.724	.757	.899	.906	.034	.841	.781	.820	.896	.907	.046
$ZoomNet_{22}$ [55]	.820	.752	.794	.878	.892	.066	.838	.729	.766	.888	.911	.029	.853	.784	.818	.896	.912	.043
$SINetv2_{22}$ [8]	.820	.743	.782	.882	.895	.070	.815	.680	.718	.887	.906	.037	.847	.770	.805	.903	.914	.048
$FEDER_{23}$ [56]	.802	.738	.781	.867	.873	.071	.822	.716	.751	.900	.905	.032	.847	.789	.824	.907	.915	.044
$HitNet_{23}$ [24]	.849	.809	.831	.906	.910	.055	.871	.806	.823	.935	.938	.023	.875	.834	.853	.926	.929	.037
$FSPNet_{23}$ [23]	.856	.799	.830	.899	.928	.050	.851	.735	.769	.895	.930	.026	.879	.816	.843	.915	.937	.035
BiRefNet	.904	.890	.904	.954	.959	.030	.913	.874	.888	.960	.967	.014	.914	.894	.909	.953	.960	.023

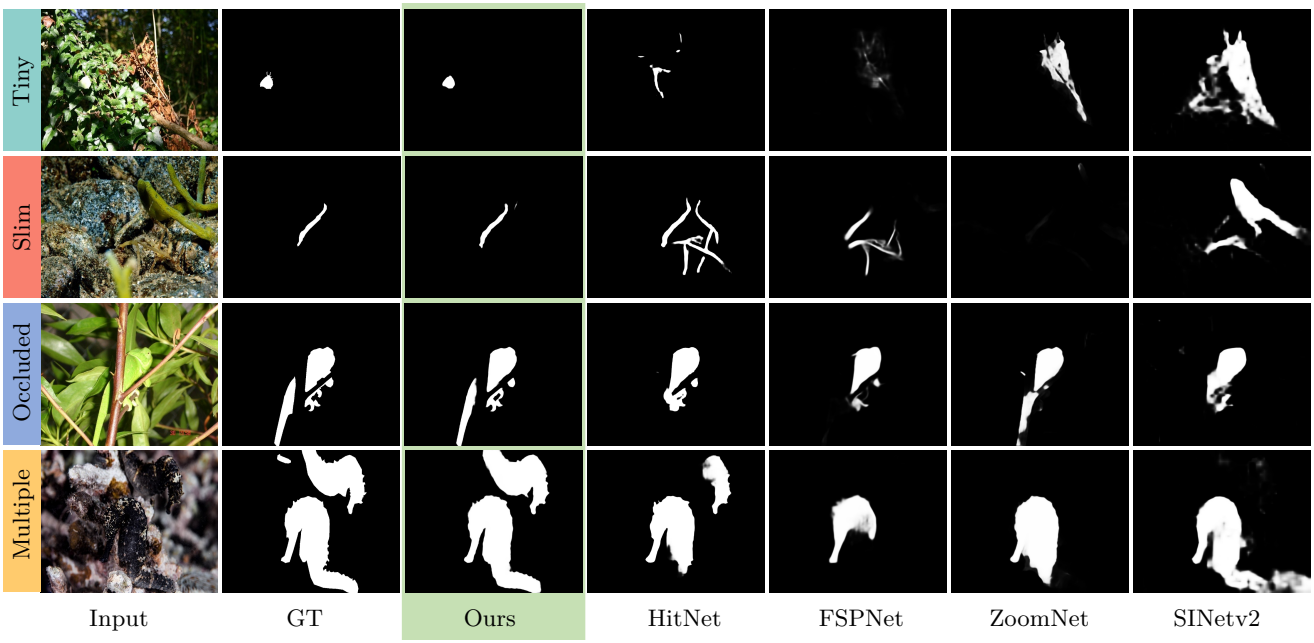

图 6 在 COD10K 基准测试上，提出的 BiRefNet 与其他竞争方法的视觉比较。这里提供了具有不同挑战性的样本，以展示 BiRefNet 在不同视角上的优越性。

VD 上进行评估，并与现有 DIS 方法进行比较。不同方法用不同颜色和标记表示。所有测试均在单个 NVIDIA A100 GPU 和 AMD EPYC 7J13 CPU 上进行。

5 潜在应用

We envisage that generated fine maps have the potential to be utilized in various practical applications.

潜在应用 #1 裂缝检测。墙壁的质量对建筑的健康至关重要 [57]。然而，通常基于常用数据集（例如 COCO [58]）训练的分割模型只能分割常规的前景对象。而我们提出的 BiRefNet 在 DIS5K 数据集上训练，对细节更加敏感，能够分割具有更高形状复杂性的目标。如 Fig. 9 （A）所示，我们的 BiRefNet 能够准确地发现墙壁上的裂缝，并帮助确定何时需要进行修复。

潜在应用 #2 高精度对象提取。前景对象提取和背景去除近年来已成为流行的应用。然而，常见的方法在目标对象具有高形状复杂性时（例如细节丰富的目标）往往难以生成高质量的结果 [32,52]，或者需要手动引导（例如涂鸦、点选或粗略遮罩）才能进行更精确的分割 [28,59]。我们提出的 BiRefNet 在 DIS5K 数据集上训练，能够生成更高分辨率的结果，并且可以在不使用遮罩的情况下分割到头发级别的细线，如 Fig. 9 (B) 所示。基于这样精细化的结果，未来可能有许多成功的下游应用。

表 7 不同 DIS 方法在性能、效率和模型复杂性方面的比较。

模型	运行时间（毫秒）	# 参数数量 (MB)	MACs (G)	DIS-TEs (HCE, F_β^ω)
BiRefNet$_{SwinL}$	83.3	215	1143	916, .858
BiRefNet$_{SwinL_cp}$	78.3	215	1143	916, .858
BiRefNet$_{SwinB}$	61.4	101	561	954, .855
BiRefNet$_{SwinT}$	40.9	39	231	980, .822
BiRefNet$_{PVTv2b2}$	47.8	35	195	969, .838
BiRefNet$_{PVTv2b1}$	36.6	23	147	978, .817
BiRefNet$_{PVTv2b0}$	32.9	11	89	1013, .806
IS-Net	16.0	44	160	1016, .726
UDUN$_{Res50}$	33.5	25	142	977, .772

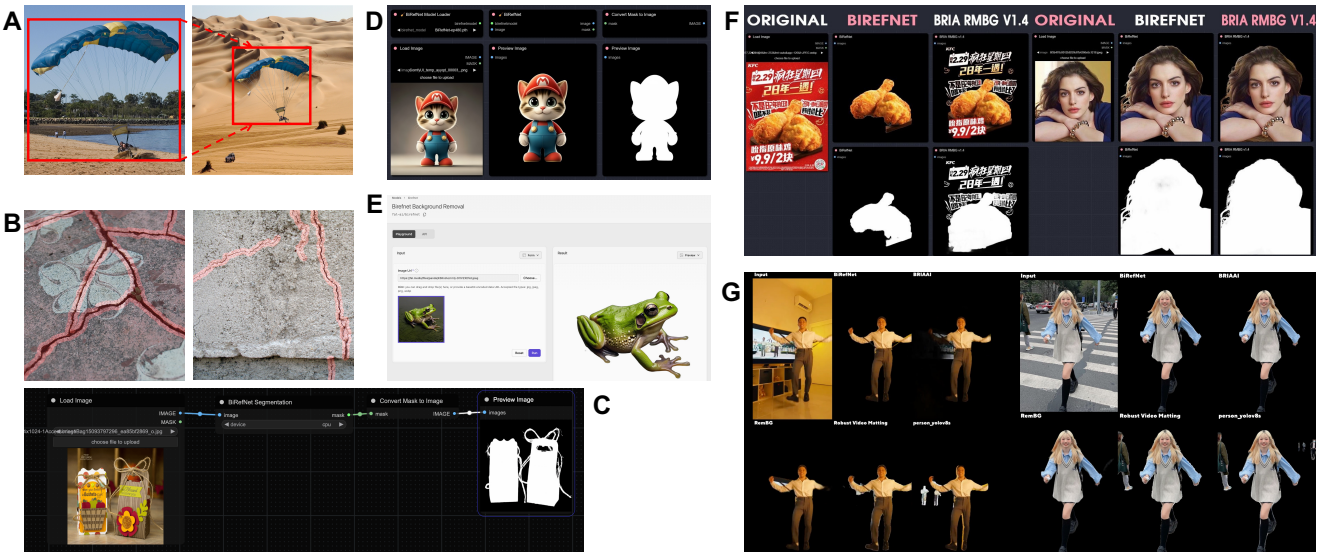

图 9 潜在应用及基于 BiRefNet 的已选择第三方应用，并在社交媒体上进行视觉比较。(A) 潜在应用 #1：建筑裂缝检测，用于维护建筑物健康。(B) 潜在应用 #2：在高分辨率自然图像中进行高精度对象提取。(C) 由 viperyl 开发的项目首先将我们的 BiRefNet 打包为 ComfyUI 节点，使这一 SOTA 模型更易于所有人使用。(D) ZHO 还提供了一个基于 ComfyUI 的项目，进一步改进了我们的 BiRefNet 的用户界面，特别是针对视频数据。(E) Fal.AI 在线封装了我们的 BiRefNet，提供了更多有用的 UI 选项和 API 调用模型的功能。(F) ZHO 提供了我们的 BiRefNet 与之前的 SOTA 方法 BRIA RMGB v1.4 的视觉比较，后者在他们的私有训练数据集上进行了额外训练。(G) Toyxyz 进行了我们的 BiRefNet 与之前竞争对手人像抠图方法（例如 BRIAAI、RemBG、Robust Video Matting 和 Person YOLOv8s）的视觉比较，涵盖了视频和图像。

6 第三方创造

自从我们的项目于 2024 年 3 月 7 日发布以来，它已经引起了社区中许多研究人员和开发者的广泛关注，并得到了他们的自发推广。此外，也有许多出色的第三方应用基于我们的 BiRefNet 开发。由于相关工作的快速增长，我们只列出了一些典型的应用。

#1 **实际应用**。因为我们的 BiRefNet 表现出色，越来越多的社区开发者创建了第三方应用[1][2]。如 Fig. 9 (C 和 D) 所示，一些开发者已将我们的 BiRefNet 集成到 ComfyUI 作为节点，这在前景分割抠图方面对后续的稳定扩散模型处理有很大帮助。为了更好的在线访问体验，Fal.AI 已经

1 https://github.com/comfyanonymous/ComfyUI
2 https://github.com/ZHO-ZHO-ZHO/ComfyUI-BiRefNet-ZHO

建立了一个在线演示，运行在 A6000 GPU 上[1]，如 Fig. 9 (E) 所示。除了常见的结果预测外，这个在线应用还提供了 API 服务，方便通过 HTTP 请求进行使用。

#2 **社交媒体**。

最近，我们的 BiRefNet 引起了社区的关注。许多推文已经在 X 平台（前身为 Twitter）上发布[2]。ZHO 提供了我们的 BiRefNet 与其他方法的视觉比较，如 Fig. 9 (F) 所示。在他们的测试中，我们的 BiRefNet 在与之前的 SOTA 方法 BRIA RMGB v1.4 的竞争中取得了有竞争力的结果[3]。需要注意的是，我们的 BiRefNet 是在开放源代码数据集 DIS5K [9] 的训练集上（根据 MIT 许可证）训练的，而其他方法则是在他们精心选择的私有数据集上训练，不能用于商业用途。如 Fig. 9 (G) 所示，Toyxyz 在 Twitter 上提供了更多关于我们的 BiRefNet 与先前优秀的前景人像抠图方法之间的视频和图像数据比较[4]。除了这些帖子，PurzBeats 还使用我们的 BiRefNet 制作了动画并上传了相关视频[5]。在 YouTube 上，“AI is in wonderland” 发布了一段关于如何在 ComfyUI 中使用我们的 BiRefNet 的日语视频教程[6]。

7 总结

这项工作提出了一个配备双边参考的 BiRefNet 框架，能够在同一框架内进行二分图像分割、高分辨率显著对象检测以及隐蔽对象检测。通过广泛的实验，我们发现，对未缩放的源图像并专注于信息丰富区域，对生成高分辨率图像中的精细和详细区域至关重要。为此，我们提出了双边参考来填补细部的缺失信息（内部参考），并引导模型更多关注富有细节的区域（外部参考）。这显著提升了模型捕捉微小像素特征的能力。为了减轻高分辨率数据训练的高成本，我们还提供了多种实用技巧，以提供更高质量的预测和更快的收敛速度。在 13 个基准测试上的竞争结果显示，我们的 BiRefNet 具有出色的性能和强大的泛化能力。我们还展示了 BiRefNet 技术在许多实际应用中的可转移性和使用情况。我们希望所提出的框架能够促进学术界各种任务的统一模型发展，并且我们的模型能够激励开发者社区创造更多优秀的作品。

致谢

本工作得到阿里巴巴创新研究项目（No.17371955）和南开大学基本科研业务费（南开大学，070-63243150）的支持。我们感谢程明明教授在本项目中的大力支持，他是上述阿里巴巴研究项目的项目经理。

参考文献

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer Assisted Interventions, 2015.

[2] William I Grosky and Ramesh Jain. A pyramid-based approach to segmentation applied to region matching. IEEE Transactions on Pattern Analysis and Machine Intelligence, 8(5):639–650, 1986.

[3] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. In European Conference on Computer Vision Workshop, 2018.

[4] Taehun Kim, Kunhee Kim, Joonyeong Lee, Dongmin Cha, Jiho Lee, and Daijin Kim. Revisiting image pyramid structure for high resolution salient object detection. In Asian Conference on Computer Vision, 2022.

[5] Xiaofei Li, Jiaxin Yang, Shuohao Li, Jun Lei, Jun Zhang, and Dong Chen. Locate, refine and restore: A progressive enhancement network for camouflaged object detection. In International Joint Conference on Artificial Intelligence, 2023.

[6] Deng-Ping Fan, Jing Zhang, Gang Xu, Ming-Ming Cheng, and Ling Shao. Salient objects in clutter. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(2):2344–2366, 2023.

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

[8] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. IEEE Transactions on

1 https://fal.ai/models/birefnet
2 https://twitter.com/search?q=birefnet&src=typed_query
3 https://twitter.com/ZHOZHO672070/status/1771026516388041038
4 https://twitter.com/toyxyz3/status/1771413245267746952
5 https://twitter.com/i/status/1772323682934775896
6 https://www.youtube.com/watch?v=o2_nMDUYk6s

Pattern Analysis and Machine Intelligence, 44(10):6024–6042, 2022.

[9] Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, et al. Highly accurate dichotomous image segmentation. In European Conference on Computer Vision Workshop, 2022.

[10] Yan Zhou, Bo Dong, Yuanfeng Wu, Wentao Zhu, Geng Chen, and Yanning Zhang. Dichotomous image segmentation with frequency priors. In International Joint Conference on Artificial Intelligence, 2023.

[11] Jialun Pei, Zhangjun Zhou, Yueming Jin, He Tang, and Heng Pheng-Ann. Unite-divide-unite: Joint boosting trunk and structure for high-accuracy dichotomous image segmentation. In ACM International Conference on Multimedia, 2023.

[12] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In IEEE / CVF International Conference on Computer Vision, 2017.

[13] Yi Zeng, Pingping Zhang, Jianming Zhang, Zhe Lin, and Huchuan Lu. Towards high-resolution salient object detection. In IEEE / CVF Computer Vision and Pattern Recognition Conference, 2019.

[14] Chenxi Xie, Changqun Xia, Mingcan Ma, Zhirui Zhao, Xiaowu Chen, and Jia Li. Pyramid grafting network for one-stage high resolution saliency detection. In IEEE / CVF Computer Vision and Pattern Recognition Conference, 2022.

[15] Xinhao Deng, Pingping Zhang, Wei Liu, and Huchuan Lu. Recurrent multi-scale transformer for high-resolution salient object detection. In ACM International Conference on Multimedia, 2023.

[16] Lv Tang, Bo Li, Yijie Zhong, Shouhong Ding, and Mofei Song. Disentangled high quality salient object detection. In IEEE / CVF Computer Vision and Pattern Recognition Conference, 2021.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In IEEE / CVF Computer Vision and Pattern Recognition Conference, 2016.

[18] 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 IEEE / CVF International Conference on Computer Vision, 2021.

[19] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In IEEE / CVF Computer Vision and Pattern Recognition Conference, 2020.

[20] Yijie Zhong, Bo Li, Lv Tang, Senyun Kuang, Shuang Wu, and Shouhong Ding. Detecting camouflaged object in frequency domain. In IEEE / CVF Computer Vision and Pattern Recognition Conference, 2022.

[21] Yujia Sun, Shuo Wang, Chenglizhao Chen, and Tian-Zhu Xiang. Boundary-guided camouflaged object detection. In International Joint Conference on Artificial Intelligence, 2022.

[22] 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. Machine Intelligence Research, 20(1):92–108, 2023.

[23] 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 IEEE / CVF Computer Vision and Pattern Recognition Conference, 2023.

[24] Xiaobin Hu, Shuo Wang, Xuebin Qin, Hang Dai, Wenqi Ren, Donghao Luo, Ying Tai, and Ling Shao. High-resolution iterative feedback network for camouflaged object detection. In AAAI Conference on Artificial Intelligence, 2023.

[25] 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, 2022.

[26] Jun Wei, Shuhui Wang, Zhe Wu, Chi Su, Qingming Huang, and Qi Tian. Label decoupling framework for salient object detection. In IEEE / CVF Computer Vision and Pattern Recognition Conference, 2020.

[27] Qian Yu, Xiaoqi Zhao, Youwei Pang, Lihe Zhang, and Huchuan Lu. Multi-view aggregation network for dichotomous image segmentation. arXiv preprint arXiv:2404.07445, 2024.

[28] Jizhizi Li, Jing Zhang, Stephen J Maybank, and Dacheng Tao. Bridging composite and real: towards end-to-end deep image matting. International Journal of Computer Vision, 130(2):246–266, 2022.

[29] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In IEEE / CVF Computer Vision and Pattern Recognition Conference, 2017.

[30] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In IEEE / CVF Computer Vision and Pattern Recognition Conference, 2017.

[31] Qihang Yu, Jianming Zhang, He Zhang, Yilin Wang, Zhe Lin, Ning Xu, Yutong Bai, and Alan Yuille. Mask guided matting via progressive refinement network. In IEEE / CVF Computer Vision and Pattern Recognition Conference, 2021.

[32] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In IEEE / CVF Computer Vision and Pattern Recognition Conference, 2019.

[33] Tiancheng Shen, Yuechen Zhang, Lu Qi, Jason Kuen, Xingyu Xie, Jianlong Wu, Zhe Lin, and Jiaya Jia. High quality segmentation for ultra high-resolution images. In IEEE / CVF Computer Vision and Pattern Recognition Conference, 2022.

[34] Chunming He, Kai Li, Yachao Zhang, Yulun Zhang, Chenyu You, Zhenhua Guo, Xiu Li, Martin Danelljan, and Fisher Yu. Strategic preys make acute predators: Enhancing camouflaged object detectors by generating camouflaged objects. In International Conference on Learning Representations, 2023.

[35] Chufeng Tang, Hang Chen, Xiao Li, Jianmin Li, Zhaoxiang Zhang, and Xiaolin Hu. Look closer to segment better: Boundary patch refinement for instance segmentation. In IEEE / CVF Computer Vision and Pattern Recognition Conference, 2021.

[36] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In IEEE / CVF Computer Vision and Pattern Recognition Conference, 2017.

[37] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Fast and accurate image super-resolution with deep laplacian pyramid networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(11):2599–2613, 2018.

[38] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In IEEE / CVF Computer Vision and Pattern Recognition Conference, pages 3684–3692, 2018.

[39] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In European Conference on Computer Vision Workshop, 2018.

[40] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In IEEE / CVF International Conference on Computer Vision, 2017.

[41] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In IEEE / CVF Computer Vision and Pattern Recognition Conference, 2017.

[42] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In IEEE / CVF Computer Vision and Pattern Recognition Conference, 2009.

[43] Zhao Zhang, Wenda Jin, Jun Xu, and Ming-Ming Cheng. Gradient-induced co-saliency detection. In European Conference on Computer Vision Workshop, 2020.

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

[45] 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 / CVF Computer Vision and Pattern Recognition Conference, 2021.

[46] 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 IEEE / CVF Computer Vision and Pattern Recognition Conference, 2017.

[47] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In IEEE / CVF Computer Vision and Pattern Recognition Conference, pages 3166–3173, 2013.

[48] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In International Joint Conference on Artificial Intelligence, 2018.

[49] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. IEEE Transactions on Image Process., 24:5706–5722, 2015.

[50] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In International Conference on Learning Representations, 2015.

[51] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems, 2019.

[52] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. Pattern Recognition, 106:107404, 2020.

[53] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(10):3349–3364, 2020.

[54] 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 / CVF Computer Vision and Pattern Recognition Conference, 2022.

[55] 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 / CVF Computer Vision and Pattern Recognition Conference, 2022.

[56] 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 IEEE / CVF Computer Vision and Pattern Recognition Conference, 2023.

[57] Qin Zou, Zheng Zhang, Qingquan Li, Xianbiao Qi, Qian Wang, and Song Wang. Deepcrack: Learning hierarchical convolutional features for crack detection. IEEE Transactions on Image Process., 28(3):1498–1512, 2018.

[58] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In European Conference on Computer Vision Workshop, 2014.

[59] Linhui Dai, Xiang Song, Xiaohong Liu, Chengqi Li, Zhihao Shi, Martin Brooks, and Jun Chen. Enabling trimap-free image matting with a frequency-guided saliency-aware network via joint learning. IEEE Transactions on Multimedia, 25:4868–4879, 2022.