本文发表于 Computational Visual Media, 2026. https://doi.org/10.26599/CVM.2025.9450512

基于级联视觉语言模型的开放词汇伪装目标分割

赵凯[1], 袁武帮[1], 王正[1], 李冠壹[1], 朱晓强[1,✉] 范登平[2] 曾丹[1]

[1] 上海大学 [2] 南开大学

kz@kaizhao.net, xqzhu@shu.edu.cn

摘要

开放词汇伪装目标分割（Open-Vocabulary Camouflaged Object Segmentation, OVCOS）旨在对任意类别的伪装目标进行分割与分类。由于视觉歧义强、测试类别未见等原因，该任务具有独特挑战。现有方法大多采用两阶段范式：先分割目标，再利用视觉语言模型（Vision-Language Models, VLMs）对分割区域进行分类。然而，这类方法存在两方面问题：其一，VLM 通常在整图上预训练，而推理时却对裁剪区域进行识别，因而产生明显的域间差异；其二，它们通常依赖针对边界清晰目标优化的通用分割模型，而这类模型对伪装目标并不友好。在缺乏显式引导的情况下，通用分割模型往往忽略细微边界，从而造成分割不精确。为解决上述问题，本文提出一种由 VLM 引导的级联框架。在分割阶段，本文利用 Segment Anything Model（SAM），并借助 VLM 进行引导。框架将 VLM 提取的特征作为显式提示注入 SAM，从而有效引导模型关注伪装区域，显著提升定位精度。在分类阶段，本文不再采用硬裁剪方式，以避免由此引入的域差异；相反，将分割结果通过 alpha 通道作为软空间先验，以在保留整幅图像上下文的同时提供精确的空间引导，从而实现更加准确且具上下文感知能力的伪装目标分类。分割与分类两个阶段共享同一个 VLM，以保证效率与语义一致性。在 OVCOS 基准以及传统伪装目标分割基准上的大量实验表明，本文方法显著优于现有方法，验证了利用丰富 VLM 语义同时服务于伪装目标分割和分类的有效性。代码与模型已开源，详见 https://github.com/intcomp/camouflaged-vlm。

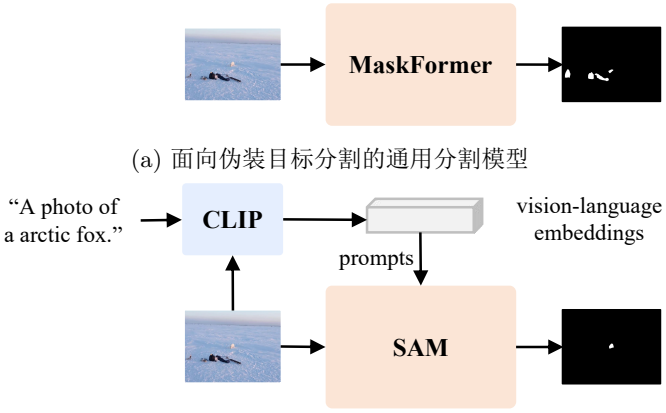
关键词: 开放词汇分割；伪装目标检测; 视觉语言模型；CLIP；SAM

1 引言

开放词汇伪装目标分割（OVCOS）是一项具有挑战性的任务，要求对伪装目标进行分割与分类，这些目标可能属于训练过程中未见过的类别 [39]。与传统的语义分割 [6, 45, 4] 相比，OVCOS 面临更多挑战，因为它需要在视觉上具有歧义的场景中识别新类别，伪装导致低对比度、边界不清晰以及目标与背景之间的高度相似性。这些挑战在医学图像分析 [13] 和农业监测 [30] 等实际应用中尤为重要。

几种现有的开放词汇分割方法 [9, 25, 3, 44] 利用视觉语言模型（VLM），例如 CLIP [41]，直接对整个输入图像中的每个像素进行分类，从而提升语义泛化能力。这类方法采用一阶段框架。然而，VLM 是预训练用于图像级理解的，这造成了粒度不匹配，阻碍了有效的视觉-语义对齐并限制了语义迁移，通常导致次优的性能 [49]。

为弥补这一差距，近期工作 [39, 49, 11, 12, 28] 首先进行类别无关的分割，然后使用 VLM 对分割出的区域进行分类。这一流程构成了两阶段框架。分割与分类的解耦部分缓解了粒度不匹配问题 [49]。然而，在分割阶段（见 图 1a），许多现有方法通常依赖通用的分割架构 [49, 11, 12, 28, 6] 来定位目标区域。这些通用分割模型主要针对轮廓清晰的目标进行定制，难以有效泛化到伪装场景中——在这些场景下，目标微妙、模糊且视觉上嵌入于复杂背景之中。预训练目标与伪装分割需求之间缺乏一致性，导致了定位不精确。此外，大多数现有方法没有引入明确的边缘感知机制，而该机制对于精确描绘具有弱边界或模糊边界的目标至关重要。



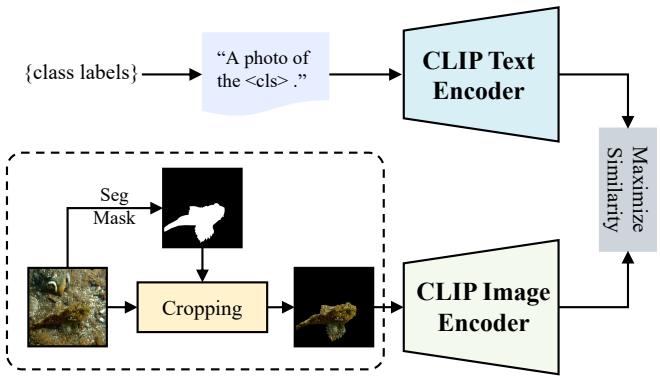
(a) 面向伪装目标分割的通用分割模型

(b) 视觉语言提示引导的伪装目标分割

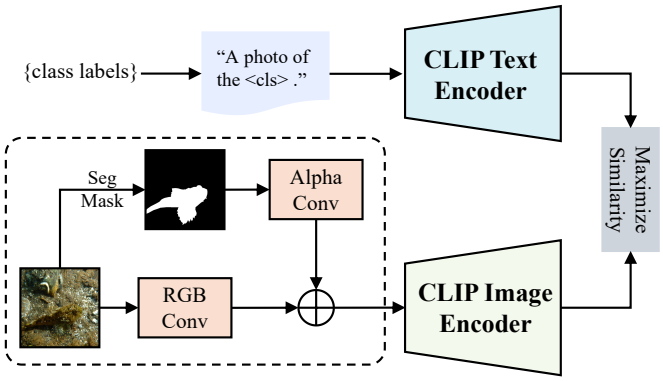
图 1: 两阶段开放词汇伪装目标分割中的不同分割范式。(a) 通用分割模型（如 MaskFormer）通常在输入图像上直接操作，没有目标特定的引导，主要设计用于分割显著的前景目标。(b) 本文的分割模型利用 CLIP 的视觉语言嵌入作为提示，引导 SAM 模型将注意力聚焦于伪装区域。

近年来，先进的基础模型，如 Segment Anything Model（SAM） [23]，展现出了对多种分割任务的卓越泛化能力 [56, 50]，这主要得益于其能够执行提示引导的分割。通过使用提示来指定目标区域，SAM 可以将其注意力自适应地聚焦于用户定义的区域，这使其在伪装目标分割等专门任务中表现得尤为有效。为了解决通用分割架构在处理边界较弱或模糊的伪装目标时存在的局限性，本文提出

✉ 朱晓强是本文的通讯作者



(a) 基于掩码裁剪的硬引导



(b) 本文的软空间引导

图 2: 掩码引导分类策略的对比。(a) 掩码裁剪策略利用分割掩码对输入图像进行裁剪，然后将其送入 CLIP 图像编码器。(b) 本文方法将分割掩码与原始图像进行融合，在保留完整图像上下文的同时实现区域感知分类。

了一种面向伪装目标分割的自适应 SAM 架构。见 图 1b。本文将源自 CLIP 的视觉嵌入和文本嵌入作为提示信息引入 SAM 的掩码解码器中，从而提供任务特定的语义引导，增强模型对伪装目标区域的聚焦能力。此外，本文通过引入条件多路注意力机制和边缘感知细化模块对掩码解码器进行增强，以提升边界定位精度，从而有效应对伪装目标轮廓模糊、不清晰的特性。

在分类阶段，大多数现有方法通常对分割得到的区域进行裁剪后再进行分类 [39, 49, 11]（见 图 2a），但由于 CLIP 是在完整图像上进行预训练的，这种做法会引入域间差异。为缓解这一域间差异，本文采用了一种区域感知分类策略，以由分割掩码生成的软空间先验替代硬裁剪，并通过图像的 alpha 通道加以应用。该方法在保留完整图像上下文信息的同时，提供了显式的空间引导。预测得到的分割掩码作为软空间先验，通过一个轻量级融合模块与输入图像进行融合，随后共同送入 CLIP [41] 图像编码器进行处理。硬空间引导与 软空间引导之间的对比见 图 2。此外，本文借鉴 [21] 提出的思想，采用多模态提示策略对 CLIP 进行微调，对视觉提示和文本提示进行联合优化。这增强了语义对齐能力和任务特定适应性，使模型能够在不破坏全局语义信息的前提下实现区域感知分类。

基于上述关键组成，本文提出了级联式开放词汇伪装目标理解网络（Cascaded Open-vocabulary Camouflaged Understanding network, COCUS），这是一种面向 OVCOS 任务的新型两阶段框架，显式地将整个过程解耦为分割和分类两个阶段。在第一阶段（分割阶段）中，本文使用 CLIP [41] 提取视觉特征和文本特征。这些特征作为提示输入到 SAM [23] 中以执行分割。这种基于提示的引导方式使 SAM 能够更加精确地聚焦于伪装目标区域，从而提升其在视觉模糊场景中的定位能力。在第二阶段（分类阶段）中，分割结果作为空间引导，用于优化其与原始图像的融合方式，使 CLIP 能够更加聚焦于目标区域并执行开放词汇分类。通过将分割与分类过程解耦，所提出的方法结合基于提示的引导分割与区域感知分类，实现了对伪装目标更加准确的语义理解。

在 OVCamo [39] 基准上的大量实验表明，所提出的框架在 OVCOS 任务中具有显著的有效性。与强基线方法 OVCoser [39] 相比，本文在所有主要评价指标上均取得了稳定提升，在这一具有挑战性的基准上建立了新的最优性能。此外，改进后的 SAM [23] 在传统 COS 任务上同样表现出较强性能，这验证了 CLIP 提示引导和边缘感知分割在标准闭集场景中的有效性。除此之外，本文方法还展现出较强的跨领域泛化能力，在医学和农业数据集上也取得了具有竞争力的结果，进一步体现了其鲁棒性和实际应用价值。综上，本文的主要贡献如下：

- 本文提出了一种面向 OVCOS 的新型两阶段框架，显式地将分割与分类过程解耦。
- 该方法采用基于提示引导的分割模型生成掩码，并将其作为分类阶段的软空间引导，在保留完整图像上下文信息的同时提升分类效果。
- 在 OVCamo 基准上的大量实验表明，本文方法取得了当前最优性能。此外，改进后的 SAM 在传统 COS 任务上表现出较强的泛化能力，验证了所提出框架在开放集和闭集伪装目标分割场景中的有效性。

本文其余部分的组织结构如下。 章 2 回顾了开放词汇分割与伪装目标理解领域的最新研究进展。 章 3 介绍了所提出的框架，详细阐述了其级联式设计、CLIP 微调流程以及改进后的 SAM 分割模型。 章 4 给出了实现细节，包括训练设置和网络结构配置，并进一步展示了全面的实验结果与消融研究。

2 相关工作

2.1 视觉语言模型

视觉语言模型（Vision-Language Models, VLMs）是一类神经网络架构，通过将图像和文本输入嵌入到共享语义空间中，学习视觉与文本的联合表征。该领域中的代表性模型之一是 CLIP [41]，其通过在大规模网络数据上进行对比学习，联合学习图像与文本表示，并在开放词汇目标检测 [54, 16, 55, 27] 和开放词汇分割 [9, 49, 11, 12, 28, 48, 53, 31, 47] 等任务中展现出较强的泛化能力。然而，基础 CLIP 在下游任务中如果缺乏任务特定的适配，往往表现欠佳。为了解决这一问题，研究者提出了多种微调方法。Alpha-CLIP [42] 引入空间自适应注意力，以

提升模型对语义相关图像区域的关注能力。CoOp [60] 和 Co-CoOp [59] 分别通过优化文本提示来提升少样本性能和泛化能力。Visual Prompt Tuning [1] 则通过向视觉分支注入细粒度提示，进一步增强模型的适应性。为克服单模态调优的局限，近期工作 [21, 22, 51] 开始采用多模态策略。FGVP [51] 学习 patch 级视觉提示，以提升跨任务的对齐能力。MaPLe [21] 则在视觉编码器和文本编码器中联合调优提示，在保持 CLIP 通用性的同时增强其任务特定适应能力。本文采用了与 MaPLe 类似的多模态提示调优框架对 CLIP 进行微调，以增强其在 OVCOS 任务中的语义对齐能力。

2.2 伪装目标分割

伪装目标分割（Camouflaged Object Segmentation, COS）已成为计算机视觉中的一个重要研究方向，其目标是分割那些在视觉上与周围环境高度融合的目标。与显著性目标检测 [2, 38, 19, 29, 26] 和语义分割 [18, 58] 等传统任务不同，COS 由于目标与背景之间对比度低、边界模糊以及背景相似性高而更具挑战性。该任务在医学图像分析 [13] 和农业监测 [30] 等领域具有重要应用价值。COS 通常被建模为一个类别无关任务，重点是在复杂视觉场景中分割伪装区域。现有研究 [13, 14, 32, 35, 37, 20, 17] 已在多个经典数据集 [13, 32, 24] 上取得了较强性能。近年来，也有一些基于 SAM 的方法 [23, 5, 33] 被适配到 COS 任务中，这些方法通过提示调优和结构改进来提升复杂场景下的分割性能。

2.3 开放词汇伪装目标分割

开放词汇伪装目标分割（Open-Vocabulary Camouflaged Object Segmentation, OVCOS）是开放词汇分割中的一个专门子任务，其目标是对属于任意文本类别的伪装目标进行分割与识别。开放词汇分割旨在通过在共享嵌入空间中对齐视觉表示与文本表示，实现对未见类别或新类别的像素级分割。早期方法 [57] 利用语义层次结构和概念图来建立词语概念与语义关系之间的联系。随着 CLIP [41] 等视觉语言模型的兴起，近年来的开放词汇分割方法逐渐转向利用预训练视觉语言模型直接建立视觉区域与文本查询之间的联系。这些方法主要分为单阶段和两阶段两类。单阶段方法如 MaskCLIP [12] 在无需额外训练的情况下将 CLIP 适配到分割任务中；SAN [48] 通过适配器增强特征表示；CAT-Seg [9] 引入图像嵌入与文本嵌入之间的代价聚合；FC-CLIP [53] 则采用层次化特征融合。然而，由于 CLIP 本身主要学习的是图像级表示，这类方法往往存在视觉—文本对齐不足的问题。两阶段方法则通过将分割与分类解耦来缓解这一问题。例如，SimSeg [49] 采用级联式设计，利用 MaskFormer [6] 生成类别无关掩码，再借助 CLIP 进行分类；OVSeg [28] 则通过在多样且含噪数据上微调 CLIP 来提升其泛化能力。在 [46] 中，研究者采用文本到图像扩散模型生成掩码。尽管这些两阶段框架在通用目标上效果较好，但在伪装场景下仍存在不足。OVCOS 尤其困难，因为低对比度视觉特征、模糊边界以及与背景高度相似的外观，都会导致分割与分类性能下降。OVCoser [39] 是首个针对该任务提出的方法，其通过两阶段流水线将专门的伪装目标分割模型与基于 CLIP 的分类器结合起来。然而，该方法在分类阶段依赖裁剪后的输入，且未能在分割过程中充分利用视觉语言模型的语义信息。

3 方法

3.1 问题定义

开放词汇伪装目标分割旨在对训练过程中未见过的新类别伪装目标进行分割与分类。形式化地，令 $\mathcal{C}_{\text{seen}}$ 表示训练阶段可用的类别集合，$\mathcal{C}_{\text{unseen}}$ 表示推理阶段的目标类别集合，且两者互不相交，即 $\mathcal{C}_{\text{seen}} \cap \mathcal{C}_{\text{unseen}} = \emptyset$。给定输入图像 I 和新类别标签 $\mathcal{C}_{\text{unseen}}$，模型需要生成一个分割掩码 M 以突出显示伪装目标，并预测其对应的类别标签 $\hat{y} \in \mathcal{C}_{\text{unseen}}$。

为完成这一任务，本文采用先分割后分类的策略。在第一阶段中，类别无关的分割模型在视觉语义和文本语义的引导下对伪装区域进行定位。在第二阶段中，视觉语言模型通过将分割区域的视觉表示与新类别标签的文本嵌入进行比较来执行开放词汇分类，从而支持开放集场景下的目标识别。

3.2 总体概述

图 3展示了本文提出的面向 OVCOS 的两阶段框架。在推理过程中，第一阶段生成类别无关的伪装目标分割掩码，第二阶段则基于分割区域执行开放词汇分类。两个阶段均采用同一个 CLIP 模型。本文的 CLIP 模型接收一个三元组 $\{I \in \mathbb{R}^{H\times W\times 3}, M, \text{text}\}$ 作为输入，其中 I 和 M 分别表示图像和掩码，text 为输入图像的文本描述，其格式为 'a photo of <*something*>'。CLIP 模型输出视觉嵌入和文本嵌入，即 E_v 和 E_t，它们在第一阶段中作为提示用于引导分割，在第二阶段中则用于基于相似度的开放词汇分类。值得注意的是，为了保证两个阶段输入格式的一致性，第一阶段使用全 1 掩码作为输入，而在第二阶段中使用预测得到的分割掩码作为输入。

在第一阶段中，如 图 3(左) 所示，本文在文本嵌入和视觉嵌入的引导下执行分割。输入包括一幅 RGB 图像 $I \in \mathbb{R}^{H\times W\times 3}$ 和一组类别标签 $\mathcal{C} = \{c_1, \ldots, c_N\}$，其中 N 表示候选类别的数量。这些输入经过 CLIP [41] 模型处理后，得到针对伪装目标理解优化的文本嵌入 E_t 和视觉嵌入 E_v。这些嵌入作为提示，与图像 I 一同输入改进后的 SAM 模型，以引导预测类别无关的伪装目标分割掩码 $M \in [0,1]^{H\times W\times 1}$，从而有效定位伪装目标。

在第二阶段中（见 图 3(右)），本文在分割结果的引导下执行开放词汇分类。输入包括与第一阶段相同的 RGB 图像 I 和类别标签 $\mathcal{C}_{\text{unseen}}$，同时将预测得到的分割掩码 M 作为额外输入送入 CLIP 模型，为分类过程提供空间引导。这些输入与第一阶段类似，共同送入 CLIP 模型进行处理，此时模型能够更加精确地聚焦于已定位的目标区域。随后，模型输出预测类别标签 $\hat{y} \in \mathcal{C}_{\text{unseen}}$，以识别伪装目标所属的类别。令 $E_t^N \in \mathbb{R}^{N\times d}$，以及 $E_v \in \mathbb{R}^{1\times d}$ 分别表示文本嵌入和视觉嵌入，其中 $d = 768$ 为特征维度。

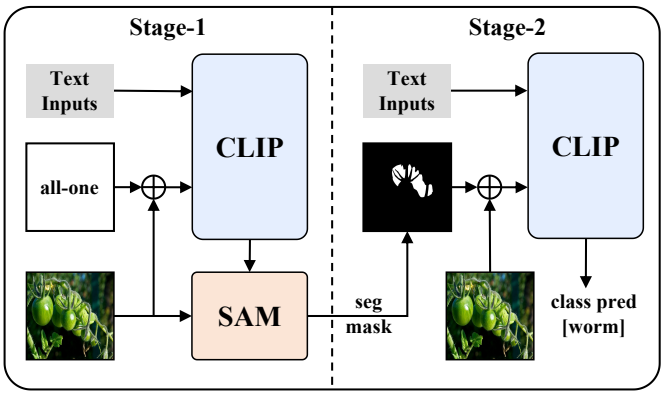


图 3: 级联式分割与分类框架概述。在阶段 1 中，改进后的 SAM 模型以文本嵌入和视觉嵌入作为提示，生成类别无关的伪装目标分割掩码。在阶段 2 中，利用生成的分割掩码实现区域感知的开放词汇分类。

首先计算相似度分数 $S \in \mathbb{R}^N$：

$$S = E_t^N \cdot (E_v)^T. \tag{1}$$

在训练过程中，首先通过优化语言分支和视觉分支中的可学习提示，对 CLIP [41] 模型进行微调，以增强其对伪装目标的敏感性，同时冻结所有编码器参数。图 4 展示了本文 CLIP 的微调流程。微调完成后，将 CLIP 模型冻结为特征提取器，并利用来自 CLIP 的视觉—文本特征作为提示来训练 SAM [23]。CLIP 微调过程的细节见章 3.3，SAM 的结构见 章 3.4。

3.3 CLIP 微调流程

本文采用多模态提示策略对 CLIP 模型进行微调，以增强其捕获细微语义线索的能力，从而更好地服务于伪装目标分割，如 图 4 所示。本文所使用的 CLIP 变体是对 Alpha-CLIP [42] 的改进版本。以往 CLIP [41] 中的提示学习策略通常仅作用于视觉模态或文本模态。仅基于语言的提示调优方法 [49, 60, 59] 只在语言分支中优化可学习提示，而仅基于视觉的策略 [12, 28, 51] 则只在视觉分支中注入提示。本文遵循 [21]，采用多模态提示策略，对文本提示和视觉提示进行联合优化，以增强多模态对齐能力，并更好地适应任务特定目标。

具体而言，本文在语言分支中附加可学习的文本提示 P_t，并生成相应的视觉提示 P_v，后者通过一个 MLP 注入器基于文本提示进行条件生成。文本提示和视觉提示 P_t 与 P_v 如 图 4(center) 所示。在微调过程中，仅更新文本提示和注入器的参数，而 CLIP 模型的其余部分均保持冻结。该轻量化策略能够实现高效适配，并提升跨模态语义对齐能力。

下面对 CLIP 模型的微调流程进行说明。微调流程首先从语言分支开始，其中基础类别标签 $\mathcal{C}_{\text{seen}}$ 通过提示模板 “A photo of the <*class*> camouflaged in the background.” 进行格式化，并结合可学习的文本提示 P_t。随后，这些输入被送入冻结的 CLIP 文本编码器，生成文本嵌入 $E_t^N \in \mathbb{R}^{N\times 768}$，其中 N 表示基础类别的数量。

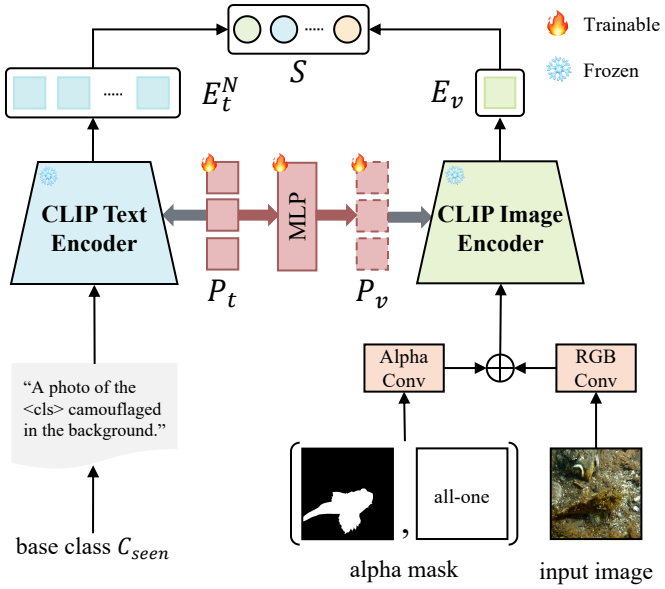


图 4: CLIP 微调流程。语言分支利用伪装特定的提示模板和可学习的文本提示 P_t 对基础类别标签 C_{seen} 进行编码，以获得文本嵌入 E_t^N。视觉分支对输入图像与 alpha 掩码的特征进行融合，并结合通过 MLP 注入的视觉提示 P_v，然后将其送入冻结的 CLIP 图像编码器，以获得视觉嵌入 E_v。通过在共享空间中对齐 E_t^N 和 E_v 来计算相似度分数 S。

在视觉分支中，输入 RGB 图像 $I \in \mathbb{R}^{H\times W\times 3}$ 会结合一个辅助 alpha 掩码 $A \in \mathbb{R}^{H\times W\times 1}$ 进行增强。alpha 掩码 A 以相同概率随机选取为全 1 掩码 A_J 或真实分割掩码 A_{gt}。这使得 CLIP 模型能够选择性地接收掩码作为输入；当没有显式掩码提供时，则默认使用全 1 矩阵，例如在第一阶段的分割过程中即是如此。

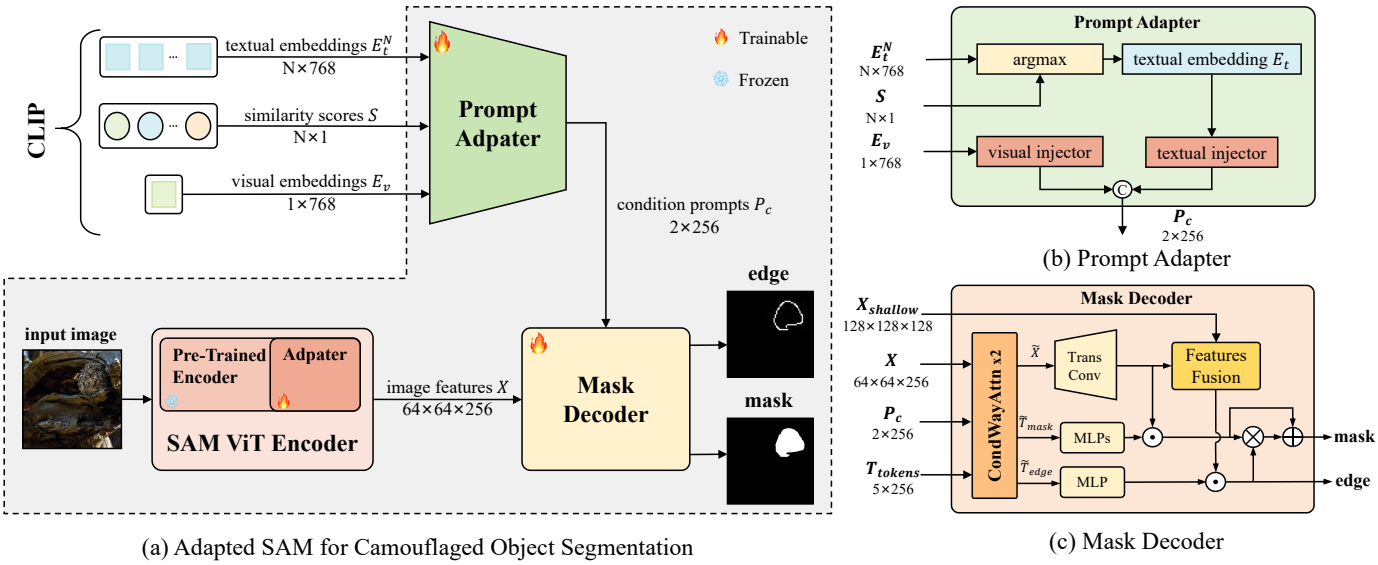
图像 I 和 alpha 掩码 A 分别通过专用卷积层进行处理，例如 图 4 中的 AlphaConv 和 RGBConv，以提取各自模态特有的特征，随后将这些特征融合形成视觉表示。该融合表示与由轻量级 MLP 注入器生成并注入的视觉提示 P_v 一同输入冻结的 CLIP 图像编码器，以获得视觉嵌入 $E_v \in \mathbb{R}^{1\times 768}$。

最后，文本嵌入和视觉嵌入按照 公式 (1) 中定义的方式计算相似度分数，并基于该分数相对于真实类别标签计算交叉熵损失。

3.4 改进的 SAM

本文在 SAM [23] 的基础上展开研究，以应对 COS 的独特挑战。尽管 SAM 在通用分割任务中表现优异，但对于伪装目标所固有的细微视觉线索和语义歧义，其处理能力仍然有限。为克服这些不足（见 图 5(a)），本文通过引入文本与视觉嵌入引导以及边缘感知增强机制，对 SAM 进行了改进，以提升分割性能。

具体而言，本文将微调后的 CLIP 模型与 SAM 相结合，以提供语义上下文信息。CLIP 模型生成文本嵌入 $E_t^N \in \mathbb{R}^{N\times 768}$、视觉嵌入 $E_v \in \mathbb{R}^{1\times 768}$ 以及相似度分数



(a) Adapted SAM for Camouflaged Object Segmentation

(b) Prompt Adapter

(c) Mask Decoder

图 5: 改进后的 SAM 框架概述。(a) 用于 *COS* 的改进 *SAM*：经过微调的 CLIP 输出文本嵌入 E_t^N、视觉嵌入 E_v 以及相似度分数 S，并通过提示适配器将其映射为条件提示 P_c。由 SAM 的 ViT 编码器提取的图像特征 X 经适配器模块进一步优化。掩码解码器融合 X 和 P_c，预测分割掩码 M 和边缘图 E，从而实现精确定位。(b) 提示适配器根据 S 选择最相关的文本嵌入，并通过轻量级 MLP 将 E_t 和 E_v 投影到统一的条件空间中，以引导解码器。(c) 改进的掩码解码器结合图像特征 X、条件提示 P_c 和输出 token T_{tokens}，生成精确的掩码和边缘图，从而提升伪装场景下的分割效果。

$S \in \mathbb{R}^{N\times 1}$。这些嵌入随后由提示适配器进一步处理，并投影为条件提示 $P_c \in \mathbb{R}^{2\times 256}$，从而在分割流程中提供高层语义引导。

与此同时，SAM 的 ViT 编码器从输入图像中提取图像特征 $X \in \mathbb{R}^{64\times 64\times 256}$。为了使 SAM 更适应伪装目标的特征，本文引入了轻量级适配器模块，在保持主干网络冻结的同时，对图像特征 X 进行优化。

最后，优化后的图像特征 X 与条件提示 P_c 在掩码解码器中进行融合，输出分割掩码 $M \in \mathbb{R}^{H\times W\times 1}$ 和边缘图 $E \in \mathbb{R}^{H\times W\times 1}$。通过在解码器中融合优化后的图像特征与条件提示，模型能够实现更加准确的目标定位和边界刻画。

如 图 5(b) 所示，提示适配器对来自微调后 CLIP 的文本嵌入和视觉嵌入进行优化，以生成用于分割引导的条件提示。给定文本嵌入 $E_t^N = \{e_t^1, \ldots, e_t^N\}$、视觉嵌入 E_v 以及相似度分数 $S = \{s_1, \ldots, s_N\}$，适配器首先选取与最高相似度分数对应的文本嵌入：

$$i^* = \arg\max_i s_i, \quad E_t = e_t^{i^*}. \tag{2}$$

选取的文本嵌入 E_t 和视觉嵌入 E_v 通过轻量级的基于 MLP 的注入器被投影到共享的 256 维条件空间中。由此得到的条件提示 $P_c \in \mathbb{R}^{2\times 256}$ 为分割掩码解码器提供高层语义和视觉引导，从而增强目标定位能力和边界预测精度。其形式化表示如下：

$$P_t = \text{MLP}_{\text{text}}(E_t), \quad P_v = \text{MLP}_{\text{vis}}(E_v), \tag{3}$$

$$P_c = [P_t, P_v] \in \mathbb{R}^{2\times 256}, \tag{4}$$

其中，$\text{MLP}_{\text{text}}(\cdot)$ 和 $\text{MLP}_{\text{vis}}(\cdot)$ 分别表示文本特征和视觉特征的投影函数。

本文对原始 SAM [23] 的掩码解码器进行了改进，以应对伪装目标分割中的特定挑战，具体方法是引入语义条件约束和边缘感知增强机制。改进后的解码器融合多层图像特征 X、条件提示 P_c 和输出 token T_{tokens}，其中包括掩码 token T_{mask} 和边缘 token T_{edge}，以实现对目标的精确定位和边界的准确细化（见 图 5(c)）。

首先，本文采用两个条件多路注意力模块（CondWayAttn × 2）来建模图像特征、条件提示和 token 之间的交互。每个模块都能够在这些组成部分之间实现稠密的双向信息流。具体而言，其包含图像到 token 和图像到条件的注意力，以融入视觉上下文；包含 token 到条件和 token 到图像的注意力，以使输出 token 与语义线索和空间线索对齐；还包括 token 自注意力和一个 MLP 层，用于捕获 token 内部依赖关系并执行特征变换。增强后的输出可表示为：

$$\tilde{X}, \tilde{T}_{\text{mask}}, \tilde{T}_{\text{edge}} = \text{CondWayAttn}(X, P_c, T_{\text{token}}). \tag{5}$$

随后，注意力增强后的特征 $\tilde{X}$ 先通过转置卷积进行上采样，以恢复空间分辨率。为恢复细节信息，这些特征再与浅层图像特征 X_{shallow} 通过如下融合模块进行结合：

$$\begin{aligned} X_{\text{fusion}} = \text{TConv}(\tilde{X}) + \\ \text{Conv}\big(\text{ReLU}\big(\text{Norm}\big(\text{Conv}(X_{\text{shallow}})\big)\big)\big). \end{aligned} \tag{6}$$

接着，注意力增强后的掩码 token 和边缘 token 分别通过任务特定的 MLP 进行映射。粗分割掩码通过掩码 token 与上采样特征逐元素相乘得到：

$$M_{\text{coarse}} = \text{MLPs}(\tilde{T}_{\text{mask}}) \odot \text{TConv}(\tilde{X}). \tag{7}$$

类似地，边缘图通过结合边缘 token 与融合特征图进行预测：

$$E = \mathrm{MLP}(\tilde{T}_{\mathrm{edge}}) \odot X_{\mathrm{fusion}}. \tag{8}$$

最终的细化掩码通过利用边缘图对粗掩码进行引导，并在此基础上进行残差相加得到：

$$M_{\mathrm{fine}} = M_{\mathrm{coarse}} + (M_{\mathrm{coarse}} \otimes E). \tag{9}$$

这种边缘引导的细化方式在保持区域一致性的同时提升了边界精度，从而能够有效应对低对比度和细微伪装结构带来的困难。该模块的有效性已在 章 4.3.5 的消融实验中得到验证。

对于损失函数，本文采用两种监督损失：用于分割的掩码损失和用于边界细化的边缘损失。

按照 [52] 的设置，将预测掩码 M 重新缩放至输入分辨率，并与真实掩码 G_m 进行比较。监督信号由二元交叉熵（BCE） [10] 和交并比（IoU）损失 [34] 共同构成：

$$\mathcal{L}_{\mathrm{mask}} = \mathcal{L}_{\mathrm{bce}}(M, G_m) + \mathcal{L}_{\mathrm{iou}}(M, G_m). \tag{10}$$

对于边缘估计，本文遵循 [39]，采用 Dice 损失 [36] 对预测边缘 E 与真实边缘 G_e 进行监督：

$$\mathcal{L}_{\mathrm{edge}} = \mathcal{L}_{\mathrm{dice}}(E, G_e). \tag{11}$$

总体目标函数定义为两项损失的无权重求和形式：

$$\mathcal{L} = \mathcal{L}_{\mathrm{mask}} + \mathcal{L}_{\mathrm{edge}}. \tag{12}$$

4 实验

4.1 实现细节

4.1.1 数据集

本文在两个任务上对所提出的方法进行了评估：伪装目标分割（COS）和开放词汇伪装目标分割（OVCOS）。对于 OVCOS 任务，所有实验均在 OVCamo [39] 数据集上进行，该数据集是专门为该任务设计的基准数据集。它包含 11,483 张来自多个公开数据集的图像，涵盖了嵌入复杂自然场景中的 75 类伪装目标。为了支持开放词汇评估，该数据集按照类别划分为两个互不重叠的子集：训练集 $\mathcal{D}_{\mathrm{train}}$ 包含 14 个已见类别中的 7,713 张图像，测试集 $\mathcal{D}_{\mathrm{test}}$ 包含 61 个未见类别中的 3,770 张图像，整体划分比例约为 7:3。

对于 COS 任务，本文在三个广泛使用的基准数据集上进行了评估：CAMO [24]、COD10K [13] 和 NC4K [32]。其中，来自 CAMO 和 COD10K 的共 4,040 张图像被用于训练。本文在这两个数据集的其余图像以及完整的 NC4K 数据集上进行了测试。所有数据集的详细统计信息，包括训练/测试划分，见 表 1。

表 1: OVCOS 和 COS 任务所使用数据集的统计概览。

数据集	任务	总数	训练	测试	类别数
OVCamo	OVCOS	11,483	7,713	3,770	75 (14/61)
CAMO	COS	1,250	1,000	250	–
COD10K	COS	5,066	3,040	2,026	–
NC4K	COS	4,121	–	4,121	–

4.1.2 评价指标

为了对 OVCOS 性能进行公平且全面的评估，本文采用了一组专门面向 OVCOS 的评价指标，这些指标是从伪装场景理解任务中提出的指标 [13, 7] 适配而来。具体而言，本文使用六项指标：类别结构度量 cS_m、类别加权 F-measure cF_β^w、类别平均绝对误差 cMAE、类别标准 F-measure cF_β、类别增强对齐度量 cE_m 和类别交并比 cIoU。这些指标是开放词汇分割领域中的标准评价指标 [9, 48, 53, 49, 28, 61]，能够从分类准确性和分割质量两个方面对模型性能进行平衡评估。

对于 COS 任务，本文遵循已有研究协议 [13]，采用四种常用评价指标：结构度量 S_α、增强对齐度量 E_ϕ、加权 F-measure F_β^ω 以及平均绝对误差 MAE。前 3 项指标用于评估预测结果与真实标注之间在结构和区域层面的相似性，其值越高表示性能越好；相反，MAE 用于衡量逐像素误差，其值越低表示精度越高。

4.1.3 训练细节

所有实验均在配备两块 NVIDIA RTX 3090Ti GPU 的工作站上完成，操作系统为 Ubuntu 20.04。本文框架基于 PyTorch 实现，并使用 CUDA 11.8 进行 GPU 加速。

在 CLIP 模型微调阶段，本文在预训练的 ViT-L/14 Alpha-CLIP 模型 [42] 上采用了多模态提示策略。该模型在 OVCamo [39] 数据集上训练 10 个 epoch，使用随机梯度下降（SGD）优化器，学习率设为 0.0035，batch size 设为 8，并在单块 GPU 上进行训练，遵循 [21] 中的设置。此外，输入的 alpha 掩码以相同概率随机选取为全 1 掩码或真实分割掩码，以平衡全局上下文编码与局部目标聚焦。

在改进后的 SAM 训练阶段，微调后的 CLIP 被集成到本文提出的改进 SAM 架构中，该架构基于 SAM [23] 的 ViT-H 版本。网络训练了 20 个 epoch，采用 Adam 优化器，初始学习率为 2×10^{-4}，并通过余弦退火策略进行衰减。训练在两块 GPU 上进行，batch size 设为 2，整个训练过程约耗时 24 小时。

4.2 与最先进方法的比较

本节从定性、定量以及效率等多个角度，将本文方法与 OVCOS 和 COS 两个任务上的最先进方法进行比较。

4.2.1 OVCOS 上的定性比较

本文首先在 OVCamo [39] 数据集上展示开放词汇伪装目标分割与分类的可视化结果，见 图 6。本文方法始终能够获得更优的分割质量，即使在低对比度、背景杂乱的场景中，也能准确勾勒伪装目标，并较好地保持目标形状与边界细节。相比其他方法，本文方法能够更好地保持目标

完整性并减少背景泄漏，展现出在伪装场景中的更强鲁棒性。

在分类方面，本文方法在不同样本上均能稳定预测出正确类别，优于以往常常对视觉模糊目标发生误分类的方法。分类准确性的提升主要得益于区域感知分类策略，该策略将分割掩码作为 alpha 掩码引入微调后的 CLIP 模型中。结合多模态提示和边缘感知解码，本文方法在开放词汇条件下同时实现了准确的定位与识别。

4.2.2 OVCOS 上的定量比较

为了全面评估所提出框架的有效性，本文将其与近期最先进的开放词汇分割方法进行了比较，包括 CAT-Seg [9]、SAN [48]、SimSeg [49]、OVSeg [28]、FC-CLIP [53]、ODISE [46] 以及基线方法 OVCoser [39]。为保证公平比较，所有模型均在 OVCamo [39] 数据集上进行训练或微调。在条件允许的情况下，本文均采用这些方法的大模型版本；唯一例外是 SimSeg，其仅发布了基础版本。正如 表 2 所示，本文方法在多个评价指标上持续优于所有对比方法。 表 2 汇总了 OVCamo [39] 数据集上的定量结果。尽管 SAN [48]、OVSeg [28] 和 CAT-Seg [9] 等开放词汇分割方法受益于大规模预训练，但由于缺乏任务特定的适配，它们在 OVCOS 任务上的性能仍然有限（例如，OVSeg 的 cS_m 为 0.164，cIoU 为 0.123）。基线方法 OVCoser [39] 通过将伪装目标分割与基于 CLIP 的分类相结合，提升了性能，取得了 0.579 的 cS_m 和 0.443 的 cIoU，但该方法未对视觉语言嵌入进行微调，也未将语义引导引入分割阶段。

不同于现有方法，本文框架利用微调后的 CLIP 和任务适配的 SAM，同时增强了分割与分类性能。本文方法取得了当前最优结果，相比基线 OVCoser [39]，在所有指标上均有显著提升：cS_m 提升 8.9%，cIoU 提升 12.5%，cF_β^w 提升 12.5%，cF_β 提升 11.1%，cE_m 提升 8.1%，同时 cMAE 降低了 7.1%。这些结果表明，本文提出的级联式设计和跨模态语义条件约束在应对 OVCOS 挑战方面具有显著优势。

4.2.3 COS 上的定量比较

如 表 3 所示，本文提出的改进 SAM 模型在三个广泛使用的 COS 基准数据集上取得了具有竞争力的性能，包括 CAMO [24]、COD10K [13] 和 NC4K [32]。与传统的非 SAM 方法 [13, 14, 32, 35, 37, 17, 20] 以及近期基于 SAM 的方法 [23, 5, 33] 相比，本文模型在所有数据集上均持续优于其他方法。

具体而言，改进后的 SAM 在 12 项评价指标中的 11 项上排名第一，在剩余 1 项上排名第二，体现出其在多样化伪装场景中的强泛化性与鲁棒性。本文方法在结构感知指标（S_α, E_ϕ）、区域感知精度（F_β^ω）和像素级精度（MAE）方面均取得了显著提升，尤其在 COD10K 和 NC4K 数据集上表现突出。这些结果表明，本文提出的边缘增强结构和提示引导分割能够有效捕获细粒度边界信息并保持语义一致性。

4.2.4 模型规模与运行时间

本文在 表 5 中比较了各模型的推理时间与显存占用，所有结果均以单块 NVIDIA RTX 3090 Ti（24GB）上单张图像的推理开销进行统计。本文方法具有较好的效率：SAM-ViT-B 仅需 240 ms 和 9.8 GB，即可在速度与精度之间取得较优平衡。更大的骨干网络（SAM-ViT-L/H）则以增加运行时间和显存开销为代价，进一步带来性能提升，从而能够在不同资源约束下实现灵活折中。

如 表 4 所示，即使是轻量级的 SAM-ViT-B 也已经优于基线 OVCoser [39]，而 SAM-ViT-H 则取得了最佳整体结果。这表明本文方法具有良好的可扩展性，能够同时适应对效率敏感和对精度要求较高的应用场景。

4.3 消融实验与相关研究

4.3.1 微调后 CLIP 的有效性

为了评估微调后 CLIP 的有效性，本文首先在多种裁剪策略下比较了不同视觉语言模型的性能。

表 6 给出了 CLIP-ConvNeXt-L [8]、CLIP-ViT-L/14 [41]、Alpha-CLIP [42] 以及本文微调后 CLIP 的结果。对于相同骨干网络，将 *full-image* 分类替换为 *hard* 裁剪会导致性能下降，这表明完整图像预训练与裁剪图像推理之间存在不匹配问题 [39]。采用软裁剪的微调后 CLIP 在所有设置下均取得了最优结果，这表明软先验能够有效缓解这种不匹配问题，同时进一步受益于微调过程中学习到的更强语义表征。

为了验证微调后 CLIP 的分类性能，本文在 OVCamo [39] 测试集上分别使用 全 *1* 掩码和 真实掩码作为软先验，对本文的 CLIP 进行了测试。 表 7 中的结果清楚表明，本文方法在 *all-one* 和 *gt* 两种设置下均提升了分类准确率。这些结果表明，任务特定的 CLIP 微调是有效的，并且在提供可靠掩码时能够进一步提升性能。

4.3.2 Alpha 掩码选择概率

本文进一步研究了在 CLIP 微调过程中，全 *1* 掩码与真实掩码之间的选择概率 P 对性能的影响。从 表 8 可以观察到，P 的取值对各项性能指标的影响较小。不同 P 取值下的结果整体较为稳定，其中 $P = 0.5$ 在所有指标上均取得了最佳性能。基于这些结果，本文在框架中采用 $P = 0.5$ 作为默认配置。

4.3.3 不同文本模板的影响

本文还进行了额外实验，以评估不同文本提示模板对分类性能的影响。具体而言，本文比较了一个通用提示模板：*'A photo of a <cls>.'*，以及本文提出的伪装特定提示模板：*'A photo of the <cls> camouflaged in the background.'*。如 表 9 所示，伪装特定模板始终优于通用模板。

4.3.4 两阶段流水线的优势

本文通过将两阶段的 分割 + 分类流水线与同时执行这两项任务的单阶段流水线进行比较，来验证两阶段设计

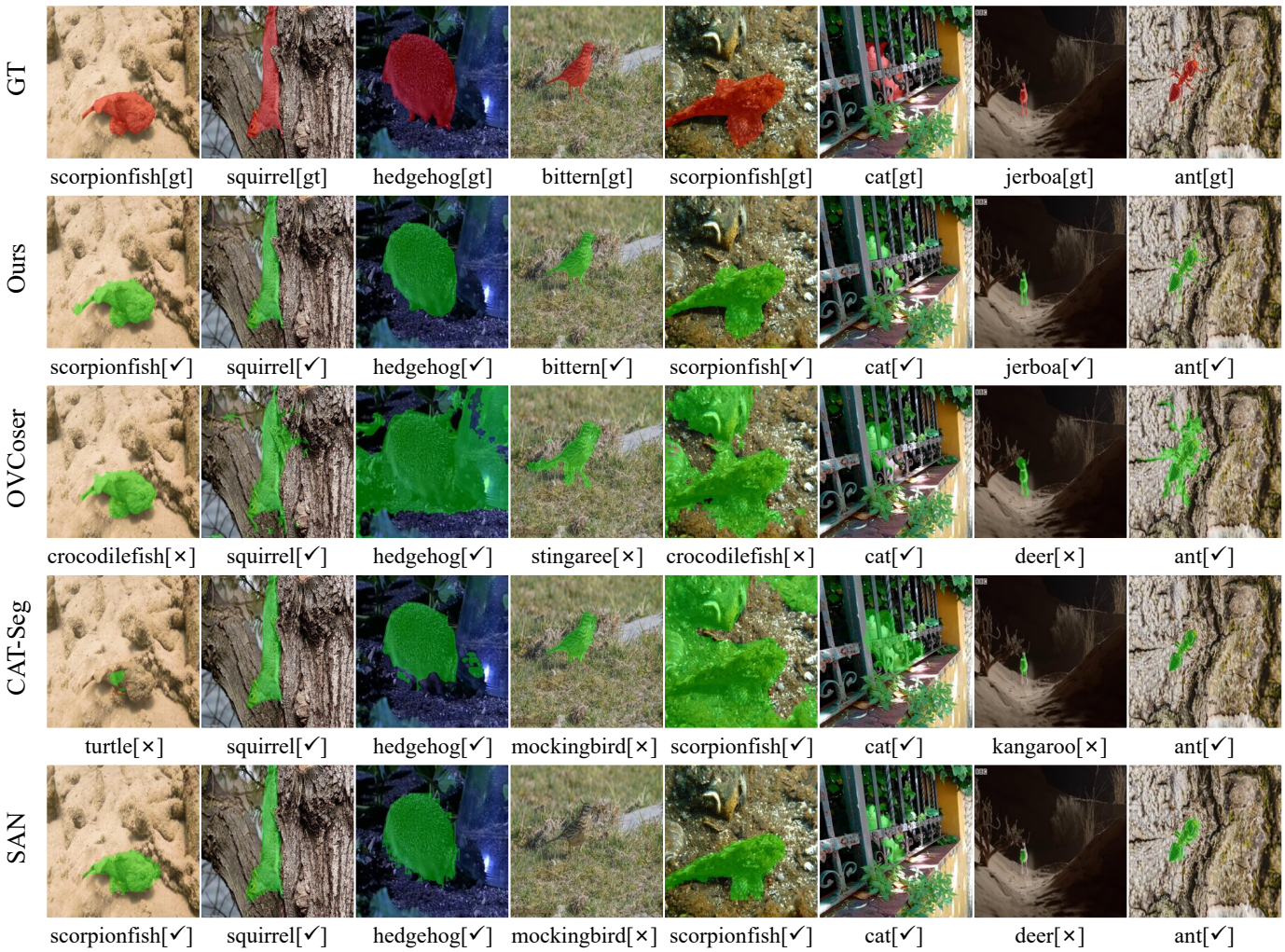


图 6: 本文方法与基于 CLIP 的基线方法在 OVCamo 上的定性比较。各列展示了输入图像、分割结果及预测标签。预测标签显示在每个分割结果下方，其中 [✓] 表示预测正确，[✗] 表示预测错误。

表 2: 本文方法与最先进的基于 CLIP 的 OVCOS 方法在 OVCamo 数据集上的比较。加粗值表示本文方法的结果，其取得了最佳整体性能；带下划线的值表示次优结果。

模型	VLM	训练集	微调	$cS_m \uparrow$	$cF_\beta^w \uparrow$	cMAE ↓	$cF_\beta \uparrow$	$cE_m \uparrow$	cIoU ↑
SimSeg	CLIP-ViT-B/16	COCO-Stuff	OVCamo	0.098	0.071	0.852	0.081	0.128	0.0
OVSeg	CLIP-ViT-L/14	COCO-Stuff	OVCamo	0.164	0.131	0.763	0.147	0.208	0.123
ODISE	CLIP-ViT-L/14	COCO-Stuff	OVCamo	0.182	0.125	0.691	0.219	0.309	0.189
SAN	CLIP-ViT-L/14	COCO-Stuff	OVCamo	0.321	0.216	0.550	0.236	0.331	0.204
FC-CLIP	CLIP-ConvNeXt-L	COCO-Stuff	OVCamo	0.124	0.074	0.798	0.088	0.162	0.072
CAT-Seg	CLIP-ViT-L/14	COCO-Stuff	OVCamo	0.185	0.094	0.702	0.110	0.185	0.088
OVCoser	CLIP-ConvNeXt-L	OVCamo	–	<u>0.579</u>	<u>0.490</u>	<u>0.336</u>	<u>0.520</u>	<u>0.616</u>	<u>0.443</u>
Ours	Our Fine-Tuned CLIP	OVCamo	–	**0.668**	**0.615**	**0.265**	**0.631**	**0.697**	**0.568**

的有效性。见 表 10。指标 cS_m、cF_β^w、$cMAE$、cF_β、cE_m 和 $cIoU$ 同时反映了分割质量和分类精度。与单阶段流水线相比，两阶段流水线在所有指标上均取得了更高性能。这些结果表明，尽管两阶段框架可能存在误差累积问题，但其整体收益依然大于潜在缺点。

表 3: CAMO、COD10K 和 NC4K 数据集上的 COS 性能比较。每项指标中的最佳结果以加粗标出，次优结果以下划线标出。

方法	CAMO				COD10K				NC4K			
	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	MAE $\downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	MAE $\downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	MAE $\downarrow$
SINet	0.751	0.771	0.606	0.100	0.771	0.806	0.551	0.051	0.808	0.883	0.768	0.058
RankNet	0.712	0.791	0.583	0.104	0.767	0.861	0.611	0.045	0.840	0.904	0.802	0.048
PFNet	0.782	0.852	0.695	0.085	0.800	0.868	0.660	0.040	0.829	0.887	0.784	0.053
SINetV2	0.820	0.882	0.743	0.070	0.815	0.887	0.680	0.037	0.847	0.903	0.770	0.048
ZoomNet	0.820	0.892	0.752	0.066	0.838	0.911	0.729	0.029	0.853	0.912	0.784	0.043
SegMaR	0.815	0.872	0.742	0.071	0.833	0.895	0.724	0.033	0.841	0.905	0.781	0.046
DGNet	0.839	0.901	0.769	0.057	0.822	0.896	0.693	0.033	0.857	0.911	0.784	0.042
SAM	0.684	0.687	0.606	0.132	0.783	0.798	0.701	0.050	0.767	0.776	0.696	0.078
SAM-Adapter	0.847	0.873	0.765	0.070	0.883	0.918	0.801	0.025	-	-	-	-
MedSAM	0.820	**0.904**	0.779	0.065	0.841	0.917	0.751	0.033	0.866	0.929	0.821	0.041
Ours	**0.865**	0.902	**0.789**	**0.057**	**0.905**	**0.947**	**0.845**	**0.019**	**0.904**	**0.933**	**0.852**	**0.031**

表 4: 不同骨干网络的性能比较，结果表明即使是轻量级的 SAM-ViT-B，相比 OVCoser 也能在保持较低计算成本的同时取得显著性能提升。

模型	骨干网络	$cS_m \uparrow$	$cF_\beta^w \uparrow$	cMAE $\downarrow$	$cF_\beta \uparrow$	$cE_m \uparrow$	cIoU $\uparrow$
OVCoser	CLIP-ConvNeXt-L	0.579	0.490	0.336	0.520	0.616	0.443
Ours	SAM-ViT-B	0.614	0.519	0.278	0.552	0.650	0.461
Ours	SAM-ViT-L	0.659	0.600	0.267	0.614	0.691	0.549
Ours	SAM-ViT-H	**0.668**	**0.615**	**0.265**	**0.631**	**0.697**	**0.568**

表 5: 不同模型的推理时间和显存占用比较。推理时间在单块 NVIDIA RTX 3090 Ti（24GB）上、batch size = 1 的设置下测得。

模型	骨干网络	时间 **(ms)**	显存 **(GB)**
SimSeg	ResNet101	350	6.5
OVSeg	Swin-B	980	15.2
ODISE	StableDiffusion	860	16.8
SAN	ViT Adapter	160	5.4
CAT-Seg	Swin-B	140	4.8
FC-CLIP	CLIP-CvNeXt-L	210	7.2
OVCoser	CLIP-CvNeXt-L	125	3.9
Ours	SAM-ViT-B	240	9.8
Ours	SAM-ViT-L	370	13.5
Ours	SAM-ViT-H	530	19.7

4.3.5 改进掩码解码器的影响

在 表 11 中，本文在 OVCamo [39] 数据集上进行了消融实验，以评估所提出的条件多路注意力（CMA）和边缘增强（EDE）模块在改进掩码解码器中的有效性。基线模型建立在带有轻量级适配器的 SAM [23] 之上，对应于未引入任何增强模块的原始 SAM 掩码解码器。尽管结构简单，该基线仍然取得了较强性能，其原因主要在于两点：(i) SAM 在大规模且多样化的数据上进行了预训练，具备强大的特征表示能力，能够提供鲁棒且具有泛化性的先验；(ii) 本文提出的两阶段设计使得即便是轻量级适配器也能有效利用这些先验。

在这一较强基线的基础上，本文方法在 cMAE 和 cS_m 上进一步取得了接近 10% 的性能提升。具体而言，加入 EDE 模块（$+$ *EDE*）后，轮廓敏感指标如 cF_β^w 和 cIoU 得到了明显提升，这表明显式边缘建模有助于提高边界精度。相比之下，引入 CMA 模块（$+$ *CMA*）则在语义感知指标如 cE_m 和 cF_β 上带来了更强提升，说明条件注意力能够有效增强文本—视觉特征融合。

当两个模块结合使用时，本文方法在所有指标上均取得了最佳性能，这表明语义条件约束与边缘感知细化具有互补优势。上述结果验证了这两种增强设计在提升复杂伪装目标场景分割质量方面的重要性。

4.3.6 分类对掩码质量的敏感性

第一阶段生成的分割掩码作为第二阶段分类的软先验。为了评估分类对掩码质量的敏感性，本文通过对正确预测的掩码施加形态学操作，模拟了存在误差的分割掩码。具体而言，本文对预测分割掩码进行膨胀和腐蚀操作，操作幅度由像素数量控制，从而生成扩张和收缩两类变体。这些模拟得到的错误掩码被用作第二阶段分类中的软先验。一个示例见 图 7，定量结果总结见 图 8。

在 图 7 中，膨胀后的掩码（上行）仍然保留了大部分伪装目标，因此能够实现正确分类（*scorpionfish*）。然而，腐蚀后的掩码（下行）排除了关键目标区域，从而导致误

表 6: 针对基于 VLM 的 OVCOS 中不同掩码裁剪策略的消融实验。*Full-image* 表示分类时使用完整图像而不进行裁剪。*Hard* 裁剪通过分割掩码直接去除周围上下文，而 *soft* 裁剪则将分割掩码与原始图像进行融合，在保留上下文线索的同时突出目标区域。

模型	**VLM**	裁剪方式	$cS_m\uparrow$	$cF_\beta^w\uparrow$	cMAE $\downarrow$	$cF_\beta\uparrow$	$cE_m\uparrow$	cIoU $\uparrow$
COCUS	CLIP-ConvNeXt-L	*Full-image*	0.573	0.524	0.365	0.544	0.607	0.495
COCUS	CLIP-ConvNeXt-L	*Hard*	0.567	0.518	0.375	0.534	0.591	0.481
COCUS	CLIP-ViT-L/14	*Full-image*	0.591	0.545	0.343	0.562	0.629	0.515
COCUS	CLIP-ViT-L/14	*Hard*	0.580	0.536	0.353	0.551	0.617	0.503
COCUS	Alpha-CLIP	*Soft*	0.639	0.589	0.299	0.603	0.668	0.545
COCUS	Our Fine-Tuned CLIP	*Soft*	**0.668**	**0.615**	**0.265**	**0.631**	**0.697**	**0.568**

表 7: 不同 CLIP 模型在 OVCamo 测试集上的分类性能。

模型	**Alpha**	**Top-1**↑	**Top-5**↑
CLIP-ConvNeXt-L	–	0.6944	0.8918
CLIP-ViT-L/14	–	0.7040	0.8915
Alpha-CLIP	all one	0.6934	0.8849
Alpha-CLIP	gt	0.7467	0.9456
Ours	all one	0.7462	0.9003
Ours	gt	**0.7859**	**0.9497**

表 8: 在 CLIP 微调过程中，全 *1* 掩码与 真实掩码之间选择概率 P 的影响。

P	$cS_m\uparrow$	$cF_\beta^w\uparrow$	cMAE $\downarrow$	$cF_\beta\uparrow$	$cE_m\uparrow$	cIoU $\uparrow$
0.00	0.659	0.607	0.266	0.624	0.688	0.559
0.25	0.660	0.610	0.266	0.619	0.690	0.558
0.50	**0.668**	**0.615**	**0.265**	**0.631**	**0.697**	**0.568**
0.75	0.663	0.611	0.267	0.621	0.693	0.560
1.00	0.661	0.611	0.268	0.623	0.691	0.561

表 9: 通用提示模板与伪装特定提示模板在微调后 CLIP 上的比较。

模型	提示模板	**Top-1**↑
Fine-Tuned CLIP	A photo of a <cls>.	0.7762
Fine-Tuned CLIP	A photo of the <cls> camouflaged in the background.	**0.7859**

表 10: 单阶段与两阶段流水线的比较。

流水线	$cS_m\uparrow$	$cF_\beta^w\uparrow$	cMAE $\downarrow$	$cF_\beta\uparrow$	$cE_m\uparrow$	cIoU $\uparrow$
One Stage	0.657	0.605	0.274	0.615	0.685	0.554
Two Stage	**0.668**	**0.615**	**0.265**	**0.631**	**0.697**	**0.568**

表 11: 改进掩码解码器中条件多路注意力（CMA）和边缘增强（EDE）的消融实验。

模型	$cS_m\uparrow$	$cF_\beta^w\uparrow$	cMAE $\downarrow$	$cF_\beta\uparrow$	$cE_m\uparrow$	cIoU $\uparrow$
Baseline	0.644	0.599	0.281	0.610	0.651	0.549
+ EDE	0.650	0.605	0.278	0.615	0.666	0.554
+ CMA	0.652	0.607	0.273	0.621	0.683	0.551
Ours	**0.668**	**0.615**	**0.265**	**0.631**	**0.697**	**0.568**

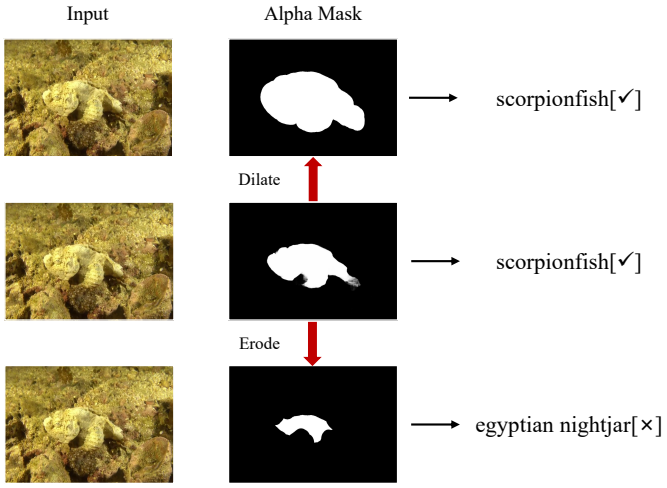


图 7: 使用错误分割预测时的分类结果。左：输入图像。中：膨胀、原始和腐蚀后的掩码。右：对应的分类结果。膨胀后的掩码和原始掩码均保留了目标对象（scorpionfish），因而得到正确分类。腐蚀后的掩码丢失了关键目标区域，从而导致误分类（egyptian nightjar）。

分类（*egyptian nightjar*）。这表明，尽管软裁剪对于适度的不准确性（如膨胀）具有一定容忍性，但对于严重腐蚀、即移除了关键目标内容的情况仍较为敏感。

为了定量分析这一现象，本文进一步研究了不同程度掩码腐蚀与膨胀下的分类准确率。见 图 8。准确率在使用原始预测掩码时达到峰值。在适度膨胀下性能基本稳定，但在严重腐蚀下则急剧下降。这些结果表明，只要关键目标区域得以保留，软先验掩码就能够保证可靠分类；而当

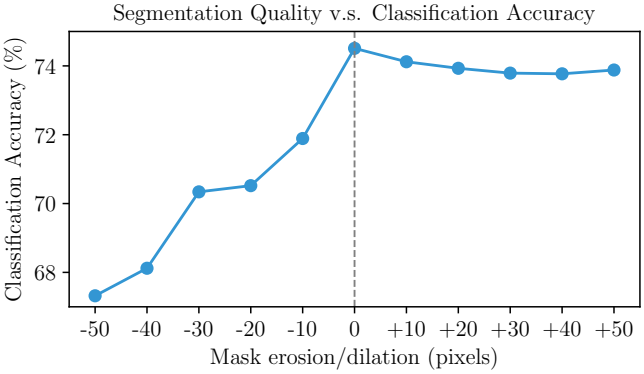


图 8: 不同程度掩码腐蚀（-）与膨胀（+）下的分类准确率。当掩码发生严重腐蚀时，分类准确率显著下降；而在适度膨胀下，准确率则相对稳定。

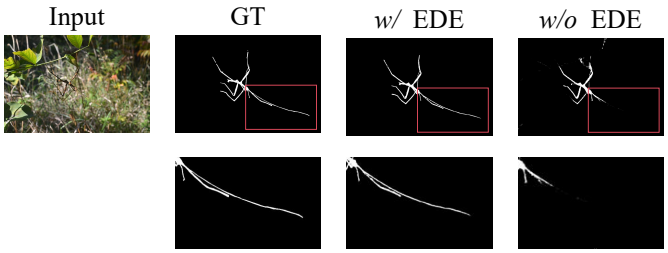


图 9: 有无 EDE 模块时的预测分割掩码。

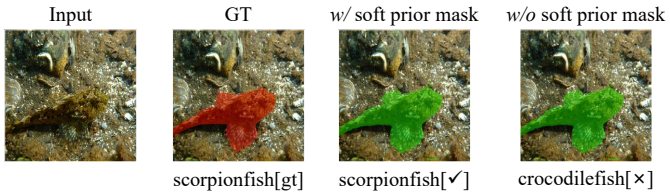


图 10: 第一阶段软先验掩码对分类的影响。[✓] 表示预测正确，[✗] 表示预测错误。

掩码发生严重结构性缺失时，分类精度将显著下降。

4.3.7 边界细化与软先验的贡献

本文进一步通过可解释性分析研究了所提出的边缘增强模块（EDE）和两阶段流水线的贡献。具体而言，本文并非仅验证它们是否能够提升性能，而是可视化并比较了有无 EDE 时的分割结果，以及单阶段和两阶段流水线的分类结果，从而更好地揭示这些组件对决策过程的影响。

如 图 9 所示，与未使用 EDE 的结果相比，加入 EDE 模块后预测得到的分割掩码具有更清晰、更准确的边界。

图 10 展示了有无软先验掩码时的分类结果。不使用掩码（*w/o*）时，直接对完整图像进行分类会误判为 crocodilefish。而使用掩码（*w/*）后，软裁剪策略能够引导模型聚焦于目标区域，从而正确识别出 scorpionfish。

4.3.8 不同设计对分割的影响

本文进一步给出可视化结果，以展示所提出模块在分割和分类中的优势。图 11 展示了基线、仅使用 CMA、仅使用 EDE 以及完整模型（CMA + EDE）的结果，并与真实标注进行了对比。基线模型无法准确定位和勾勒伪装目标，生成的掩码通常不完整或伴随明显噪声。引入 CMA 后，模型能够更好地识别目标区域，从而增强语义聚焦能力；引入 EDE 后，边界质量得到提升，轮廓更加清晰且连续。结合 CMA 和 EDE 的完整框架产生了与真实标注最一致的结果，从而验证了两者的互补作用。

4.3.9 分类策略的选择

图 12 比较了 3 种分类策略：*full-image*、*hard crop* 以及本文提出的 *soft crop*。完整图像策略容易受到背景干扰，因此常常导致误分类。硬裁剪会引入域间差异，因为 CLIP [41] 是在完整图像上进行预训练的。直接围绕掩码进行裁剪会丢弃有用的上下文信息，同时可能保留无关区域，从而破坏图像结构并降低分类精度。相比之下，本文提出的软裁剪策略在突出目标区域的同时保留了全局上下文，因此能够在不破坏 CLIP 预训练假设的前提下实现准确识别。

4.3.10 注意力与边缘特征可视化

本文在 图 13 中给出了可视化结果，以突出 CMA 和 EDE 的作用。

图 (b) 中的注意力图来源于 CMA 中语义条件提示与 SAM [23] 图像特征之间的交叉注意力。该注意力图能够清晰聚焦于伪装目标，同时抑制无关的背景纹理，从而在低对比度条件下增强语义聚焦能力。

对于 EDE，图 (c) 中的可视化结果对应于式 (8) 中的边缘相关特征图 X_{fusion}。经过通道聚合后，该特征图能够突出与目标边界一致的细而连续的轮廓，同时在背景区域保持较低响应。这表明在边缘预测之前，其已经实现了更优的边界质量建模。

4.4 跨领域泛化能力

为了测试所提方法的泛化性，本文将实验进一步扩展到两个下游任务：(i) 基于 Kvasir [40] 数据集的医学息肉分割，以及 (ii) 基于 ACOD-12K [43] 数据集的农业遮挡作物检测。这一评估为检验本文方法在主基准任务之外的适应能力和实际应用价值提供了更严格的实验依据。

4.4.1 数据集概述

Kvasir [40] 数据集是胃肠内窥镜息肉分割中广泛使用的基准数据集。它包含带有像素级精确息肉标注的高分辨率结肠镜图像。由于息肉在大小、形状和纹理上存在显著变化，因此该数据集具有较高挑战性。

ACOD-12K [43] 数据集是一个面向真实温室密集场景中遮挡作物检测的大规模农业数据集。该数据集包含的图像中，作物常常被叶片、茎秆和支撑结构部分或完全遮挡，

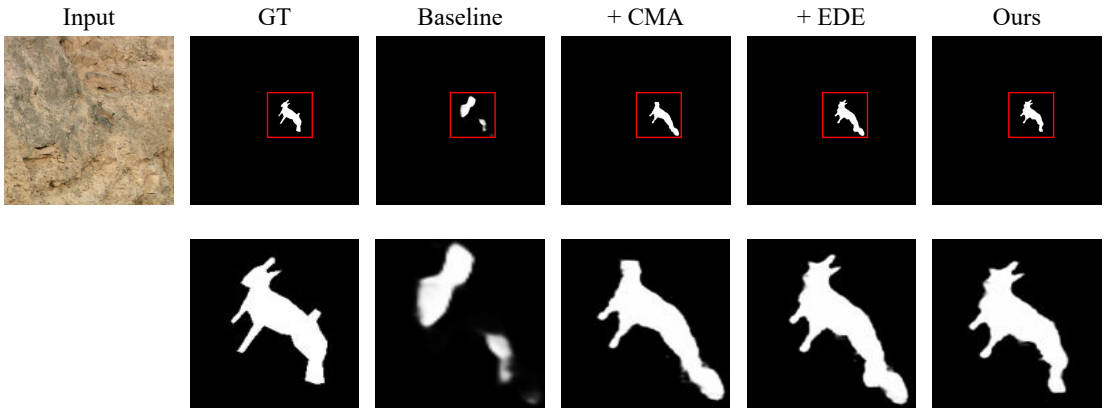


图 11: CMA 和 EDE 的贡献：展示了基线（baseline）、仅使用 CMA、仅使用 EDE 以及完整模型与真实标注之间的对比结果。

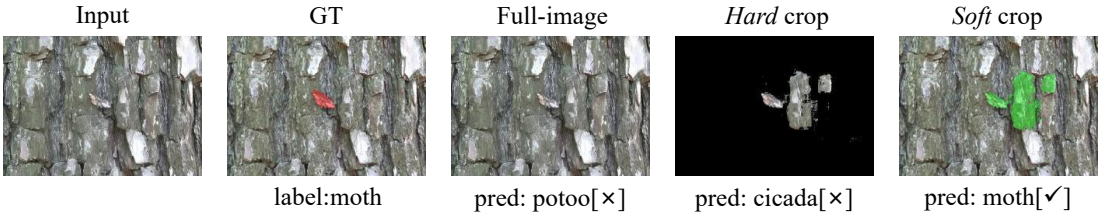


图 12: 分类策略比较：*full-image*、*hard crop* 以及本文提出的 *soft crop*。[✓] 表示预测正确，[✗] 表示预测错误。

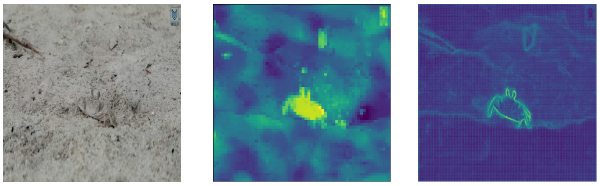


图 13: CMA 和 EDE 的可视化： (a) 包含低对比度伪装目标的输入图像，(b) CMA 中语义条件提示与 SAM 图像特征之间的交叉注意力图，(c) 来自 EDE 的通道聚合边缘特征图。

同时伴随复杂背景、光照变化以及微弱视觉线索，使得检测任务极具挑战性。

4.4.2 实验设置

为了评估本文方法在下游任务中的表现，本文分别在以下两个场景中对模型进行了微调：

1. 医学领域：按照 [15] 的设置，本文在 Kvasir [40] 数据集上对模型进行了微调。
2. 农业领域：按照 [43] 的设置，本文在 ACOD-12K [43] 数据集上对模型进行了微调。

示例分割结果见 图 14，定量比较结果见 表 12 and 13。与强基线方法和最先进方法相比，本文方法在两个数据集

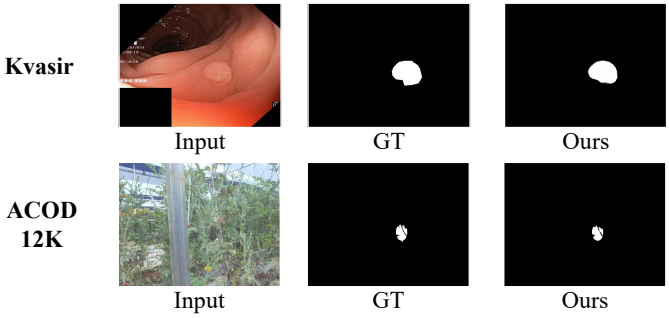


图 14: Kvasir（医学）和 ACOD-12K（农业）数据集上的定性比较，展示了原始图像、真实掩码以及本文方法预测得到的分割结果。

表 12: 在 **Kvasir** 医学息肉分割数据集上的定量比较。加粗值表示最佳结果，带下划线的值表示次优结果。

模型	**wFm**↑	**Sm**↑	**Emax**↑	**MAE**↓
U-Net	0.7940	0.8580	0.8930	0.0550
U-Net++	<u>0.8080</u>	<u>0.8620</u>	<u>0.9100</u>	<u>0.0480</u>
SFA	0.6700	0.7820	0.8490	0.0750
Ours	**0.8551**	**0.9057**	**0.9223**	**0.0370**

上均取得了最佳结果。这些结果验证了本文方法在多样化真实应用领域中的鲁棒性与实际适用性。

表 13: 在 **ACOD-12K** 农业遮挡作物检测数据集上的定量比较。加粗值表示最佳结果，带下划线的值表示次优结果。

模型	**Sm↑**	**wFm↑**	**Emax↑**
SINet	0.7450	0.4740	0.8260
PFNet	0.8050	0.6850	<u>0.9420</u>
ZoomNet	<u>0.8320</u>	<u>0.7470</u>	0.9400
FSPNet	0.7190	0.5260	0.8190
Ours	**0.8455**	**0.8002**	**0.9625**

5 结论

本文提出了 COCUS，一种面向 OVCOS 的两阶段框架，其显式地将分割与分类过程解耦。在第一阶段中，利用微调后的 CLIP 模型提取视觉嵌入和文本嵌入，并以此引导改进后的 SAM 及其重新设计的掩码解码器，从而增强目标定位能力和边界刻画精度。在第二阶段中，将预测得到的分割掩码与输入图像进行融合，以引导模型将注意力聚焦于目标区域，从而在不依赖裁剪输入的情况下实现区域感知分类。大量 OVCOS 和 COS 基准实验表明，本文方法优于现有开放词汇分割方法。改进后的 SAM 在 COS 基准上也取得了更优结果。这些实验验证了两阶段框架和边缘感知增强在复杂伪装场景中的有效性。

参考文献

[1] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Visual prompting: Modifying pixel space to adapt pre-trained models. *arXiv preprint arXiv:2203.17274*, 3(11-12):3, 2022.

[2] Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li. Salient object detection: A survey. *Computational Visual Media*, 5:117–150, 2019.

[3] Maxime Bucher, Tuan-Hung VU, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. In *Advances in Neural Information Processing Systems*, volume 32, pages 466–477, 2019.

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 801–818, September 2018.

[5] Tianrun Chen, Lanyun Zhu, Chaotao Deng, Runlong Cao, Yan Wang, Shangzhan Zhang, Zejian Li, Lingyun Sun, Ying Zang, and Papa Mao. Sam-adapter: Adapting segment anything in underperformed scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 3367–3375, October 2023.

[6] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *Advances in Neural Information Processing Systems*, volume 34, pages 17864–17875, 2021.

[7] Xuelian Cheng, Huan Xiong, Deng-Ping Fan, Yiran Zhong, Mehrtash Harandi, Tom Drummond, and Zongyuan Ge. Implicit motion handling for video camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13864–13873, June 2022.

[8] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, June 2023.

[9] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4123, June 2024.

[10] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Annals of operations research*, 134(1):19–67, 2005.

[11] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11583–11592, June 2022.

[12] Zheng Ding, Jieke Wang, and Zhuowen Tu. Open-vocabulary universal image segmentation with maskclip. In *International Conference on Machine Learning*, pages 8090–8102, 2023.

[13] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6024–6042, 2022.

[14] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2777–2787, June 2020.

[15] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI*, pages 263–273, 2020.

[16] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, pages 1–14, 2022.

[17] 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.

[18] Wei Ji, Jingjing Li, Cheng Bian, Zongwei Zhou, Jiaying Zhao, Alan L Yuille, and Li Cheng. Multispectral video semantic segmentation: A benchmark dataset and baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1094–1104, 2023.

[19] Wei Ji, Ge Yan, Jingjing Li, Yongri Piao, Shunyu Yao, Miao Zhang, Li Cheng, and Huchuan Lu. Dmra: Depth-induced multi-scale recurrent attention network for rgb-d saliency detection. *IEEE Transactions on Image Processing*, 31:2321–2336, 2022.

[20] Qi Jia, Shuilian Yao, Yu Liu, Xin Fan, Risheng Liu, and Zhongxuan Luo. Segment, magnify and reiterate: Detecting camouflaged objects the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4713–4722, June 2022.

[21] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, June 2023.

[22] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15190–15200, October 2023.

[23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, October 2023.

[24] 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.

[25] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, pages 1–13, 2022.

[26] Jingjing Li, Wei Ji, Miao Zhang, Yongri Piao, Huchuan Lu, and Li Cheng. Delving into calibrated depth for accurate rgb-d salient object detection. *International Journal of Computer Vision*, 131(4):855–876, 2023.

[27] Shuai Li, Minghan Li, Pengfei Wang, and Lei Zhang. Opensd: Unified open-vocabulary segmentation and detection. *arXiv preprint arXiv:2312.06703*, 2023.

[28] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, June 2023.

[29] Jiang-Jiang Liu, Qibin Hou, Zhi-Ang Liu, and Ming-Ming Cheng. Poolnet+: Exploring the potential of pooling for salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):887–904, 2022.

[30] 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.

[31] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. SegCLIP: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 23033–23044, 23–29 Jul 2023.

[32] 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 *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 11591–11601, June 2021.

[33] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

[34] Gellert Mattyus, Wenjie Luo, and Raquel Urtasun. Deeproadmapper: Extracting road topology from aerial images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3438–3446, Oct 2017.

[35] Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8772–8781, June 2021.

[36] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571, 2016.

[37] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2160–2170, June 2022.

[38] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9413–9422, June 2020.

[39] Youwei Pang, Xiaoqi Zhao, Jiaming Zuo, Lihe Zhang, and Huchuan Lu. Open-vocabulary camouflaged object segmentation. In *European Conference on Computer Vision*, pages 476–495, 2024.

[40] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, et al. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pages 164–169, 2017.

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763, 18–24 Jul 2021.

[42] Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Alpha-clip: A clip model focusing on wherever you want. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13019–13029, June 2024.

[43] Liqiong Wang, Jinyu Yang, Yanfu Zhang, Fangyi Wang, and Feng Zheng. Depth-aware concealed crop detection in dense agricultural scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17201–17211, June 2024.

[44] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero- and few-label semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8256–8265, June 2019.

[45] 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 *Advances in Neural Information Processing Systems*, volume 34, pages 12077–12090, 2021.

[46] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, June 2023.

[47] Jilan Xu, Junlin Hou, Yuejie Zhang, Rui Feng, Yi Wang, Yu Qiao, and Weidi Xie. Learning open-vocabulary semantic segmentation models from natural language supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2935–2944, June 2023.

[48] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2945–2954, June 2023.

[49] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision*, pages 736–753, 2022.

[50] Tianyu Yan, Zifu Wan, Xinhao Deng, Pingping Zhang, Yang Liu, and Huchuan Lu. Mas-sam: segment any marine animal with aggregated features. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24, pages 6886–6894, 2024.

[51] Lingfeng Yang, Yueze Wang, Xiang Li, Xinlong Wang, and Jian Yang. Fine-grained visual prompting. In *Advances in Neural Information Processing Systems*, volume 36, pages 24993–25006, 2023.

[52] Bowen Yin, Xuying Zhang, Deng-Ping Fan, Shaohui Jiao, Ming-Ming Cheng, Luc Van Gool, and Qibin Hou. Camoformer: Masked separable attention for camouflaged object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10362–10374, 2024.

[53] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. In *Advances in Neural Information Processing Systems*, volume 36, pages 32215–32234, 2023.

[54] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *European conference on computer vision*, pages 106–122, 2022.

[55] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, June 2021.

[56] Pingping Zhang, Tianyu Yan, Yang Liu, and Huchuan Lu. Fantastic animals and where to find them: Segment any marine animal with dual sam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2578–2587, June 2024.

[57] Hang Zhao, Xavier Puig, Bolei Zhou, Sanja Fidler, and Antonio Torralba. Open vocabulary scene parsing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2002–2010, Oct 2017.

[58] Xiaoqi Zhao, Youwei Pang, Jiaxing Yang, Lihe Zhang, and Huchuan Lu. Multi-source fusion and automatic predictor selection for zero-shot video object segmentation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2645–2653, 2021.

[59] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, June 2022.

[60] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[61] Chaoyang Zhu and Long Chen. A survey on open-vocabulary detection and segmentation: Past, present, and future. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):8954–8975, 2024.