arXiv:2601.00537v1 [cs.CV] 2 Jan 2026

# 提升 Segment Anything Model 对视觉非显著场景的泛化能力

Guangqian Guo，P engfei Chen、Yong Guo、Huafeng Chen、Boqiang Zhang、Shan Gao *IEEE* 会员

摘要——**以卓越的零样本分割能力著称的 Segment Anything Model（SAM）在学界引起了广泛关注。然而，在我们所称的视觉非显著场景中，前景与背景之间对比度较低，SAM 的性能会受到影响。在这类场景下，现有方法往往无法捕捉准确的轮廓，难以生成理想的分割结果。本文提出视觉非显著 SAM （ VNS-SAM），旨在增强 SAM 对视觉非显著场景的感知能力，同时保留其原有的零样本泛化能力。我们通过两项设计有效利用 SAM 的低级特征：掩码-边缘令牌交互解码器和非显著特征挖掘模块。这些设计使 SAM 解码器能够更深入地理解非显著特征，同时仅略微增加参数量和计算需求。VNS-SAM 的新增参数可在 4 小时内完成优化，体现了该方法的可行性与实用性。在数据方面，区别于以往针对单一任务的适配工作，我们构建了 VNS-SEG，这是一个涵盖多种 VNS 场景的统一数据集，包含 3.5 万余张图像。该数据集旨在帮助模型学习更稳健的 VNS 特征，并全面评测模型在 VNS 场景中的分割性能与泛化能力。在多种 VNS 分割任务上的大量实验表明，VNS-SAM 性能优异，尤其是在零样本设置下表现突出，展现出广泛应用于现实世界的潜力。代码和数据集已公开发布，网址为：https://guangqian-guo.github.io/VNS-SAM/。**

索引词——**视觉非显著特征、Segment Anything Model、基础模型微调**

## I. 引言

在多种场景中准确地分割目标 [1]–[8]是各种高级视觉应用的一项基础任务。近期，Segment Anything Models（SAMs） [9]–[11],作为基础分割模型，凭借出色的零样本分割能力，在社区中产生了重大影响。它可以使用点和边界框等视觉提示，对图像中的任意目标进行交互式分割。

Guangqian Guo、Huafeng Chen 和 Shan Gao 任职于西北工业大学无人系统研究院（中国西安 710072；电子邮箱：guogq21@mail.nwpu.edu.cn；chf@mail.nwpu.edu.cn；gaoshan@nwpu.edu.cn）。Pengfei Chen 任职于中国科学院大学电子、电气与通信工程学院（中国北京 101408；电子邮箱：chenpengfei20@mails.ucas.ac.cn）。Yong Guo 任职于马克斯·普朗克信息学研究所（MPI-INF）（电子邮箱：guoyongcs@gmail.com）。Boqiang Zhang 任职于中国科学技术大学（电子邮箱：cyril@mail.ustc.edu.cn）。本研究部分由国家自然科学基金（NSFC）项目 62372382 资助。© 2025年 IEEE。本文为作者已接收的手稿。允许个人使用本材料。其他所有用途均须获得 IEEE 的许可。

SAM 在庞大的 SA-1B 数据集上进行训练，其强大的泛化能力推动了各种下游任务取得突破并催生了新的范式，包括遥感 [12]–[14]，自动数据标注 [15]–[17],和医学图像分割 [18], [19]。

近期一些研究 [20]–[22] 指出，SAM 在面对复杂场景时性能会下降，例如伪装场景以及医学图像中的息肉。如图 1 所示，由于 VNS 场景中前景与背景具有高度的内在相似性，SAM 无法有效感知细微的判别区域，容易混淆前景与背景，从而生成错误的掩码。这一局限严重限制了 SAM 在真实世界中的适用性。一些研究 [23]–[27] 通过微调和适配器模块，针对特定下游任务对 SAM 进行专门优化。然而，这些方法仅聚焦于特定任务，因而忽视了不同复杂场景之间的共性知识，还可能损害 SAM 对其他场景固有的泛化能力。

本文尝试从统一的视角解决这一问题。我们发现，SAM 表现不佳的一些场景具有共同特征：前景与背景之间的对比度较低，且物体边界模糊（如图 1 所示）。我们将这种共性称为视觉非显著性（VNS），并将这类场景称为 VNS 场景。统一视角旨在共同提升 SAM 对统一 VNS 特征的学习能力，从而持续增强其在这些 VNS 场景中的表现。为了学习统一的 VNS 知识，受低级特征（如边缘和纹理）对 VNS 目标感知至关重要这一事实的启发 [28]–[35], ，我们尝试在 SAM 中有效利用这些特征，以增强其对 VNS 对象的感知能力；然而，这仍是一个尚待解决的问题。

为此，我们提出 ***VNS-SAM***，通过两个关键组件充分利用 SAM 的低级特征，从而增强其对 VNS 特征的感知能力。首先，我们通过增强对目标边缘的感知，促使 SAM 解码器高效地学习 VNS 特征。我们没有微调整个掩码解码器（图 2 (a)）或单个掩码令牌 [36] （图 2 (b)），而是设计了一种掩码-边缘令牌交互式解码器（图 2 (c)）。该设计的核心是引入 VNS 令牌（VNS 掩码令牌和 VNS 边缘令牌）之间的交互，并进行双层增强，以有效提升解码器学习 VNS 特征的能力，从而增强 SAM 的掩码预测。其次，我们旨在从高度优化的图像编码器中挖掘 VNS 特征，以丰富预测层的表征。要准确预测 VNS 对象，需要

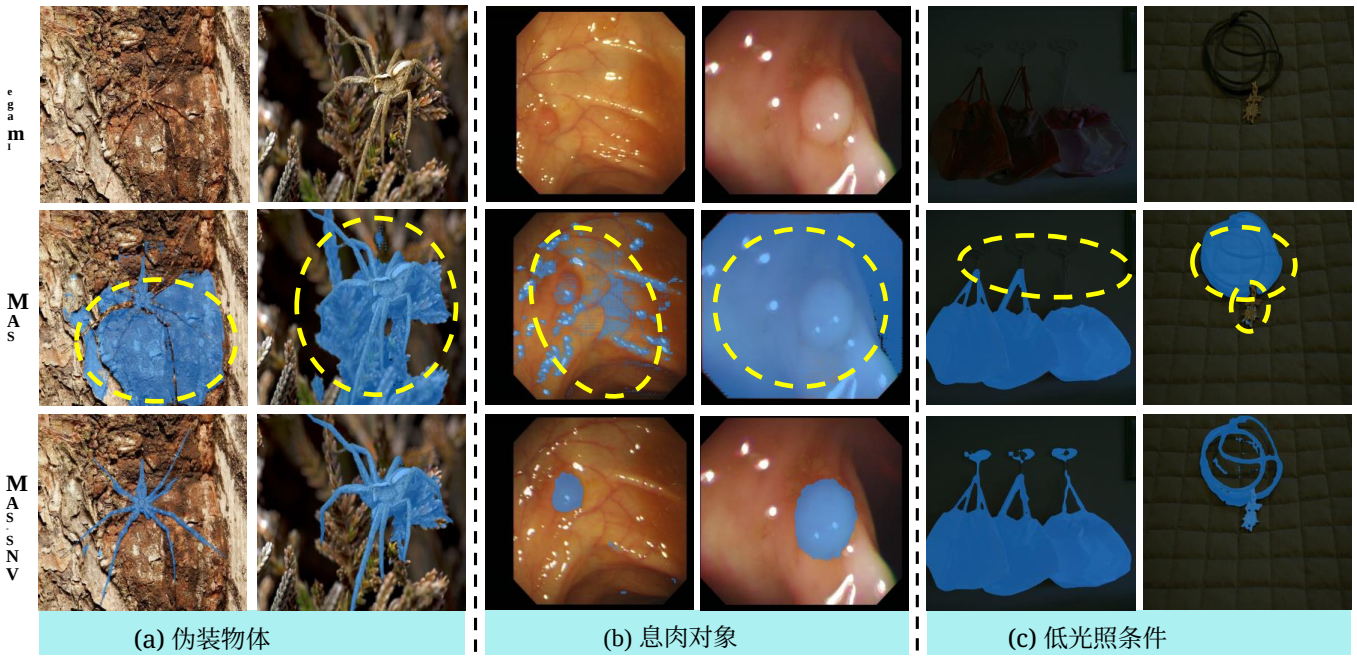



图1. SAM 和 VNS-SAM 在三种典型非显著场景下预测的掩码对比。SAM（第二行）在处理以下情况时常常表现不佳：(a) **伪装物体** 与周围环境完美融合；(b) **息肉对象** 中息肉组织和正常组织具有相同的纹理，给医学图像分析带来挑战；以及 (c) **低光照条件下的目标** 与背景之间缺乏显著的颜色对比度。SAM 无法准确识别物体边界和完整结构，导致分割细节缺失，并错误地预测背景。相比之下，VNS-SAM（第三行）能够生成更精准的分割结果。（彩色查看效果最佳）

充分探索细微的判别性特征。为此，我们设计了一种轻量级的非显著特征挖掘（NSFM）模块，从 SAM 编码器中提取信息最丰富的分量，从而实现更精准的预测。VNS-SAM 以 SAM 为基础，在训练阶段冻结其原始预训练参数，借助基础模型的泛化能力。值得注意的是，所提出的 VNS-SAM 仅增加 9.8 M 个参数，并且可在 4× RTX 4090 GPU 上于 4 小时内高效完成训练。

在数据方面，为了使模型学习 VNS 特征，我们没有针对单一数据集进行适配，而是为 VNS 场景建立了一个名为 ***VNS-SEG*** 的统一数据集。这不仅有助于模型学习更稳健的非显著特征，还能提升模型在多项任务中的表现。VNS-SEG 包含 3.5 万组成多样 VNS 场景的图像-掩码对，来源于现有知名数据集和我们合成的数据。VNS-SEG 的训练集由 23,232 张图像组成，评估集则包含涵盖 4 种 VNS 场景的 11 个子集。评估集分为已见集和未见集，以全面评测模型的零样本迁移能力。我们希望所构建的 VNS-SEG 数据集能够启发更多适用于 VNS 场景的分割模型，并为未来研究提供助益。

总体而言，本研究的主要贡献可概括为四个方面。

- 我们分析了 SAM 在一系列前景与背景对比度较低的场景中的局限性，并将这些场景统称为 VNS场景。因此，我们提出了 VNS-SAM，这是一种基于 SAM 构建的通用交互式分割模型，能够提升其应对各种 VNS 场景的鲁棒性。
- 我们在 VNS-SAM 中开发了掩码-边缘令牌交互式解码器和非显著特征挖掘模块，以促使模型挖掘细微的判别性特征。所提出的方法仅引入极少量参数，并可在不到 4 小时内高效完成训练。
- 我们构建了统一数据集 VNS-SEG，其中包含超过 3.5 万对图像-掩码对，用于最优地训练和评估模型。与单任务数据集相比，该统一数据集有助于模型学习更稳健的 VNS 特征。VNS-SEG 旨在为 VNS分割建立新的基准。
- 我们进行了广泛的实验，结果表明，VNS-SAM 在各种 VNS 场景中均能实现卓越的分割性能，同时保持强大的交互式分割泛化能力。

## II. 相关工作

**Segment Anything模型及其变体**。Segment Anything模型（SAM） [9], [10]凭借出色的零样本分割能力，在学界产生了重大影响。作为基础分割模型，SAM 使用规模庞大的 SA-1B 数据集 [9],进行训练，该数据集包含超过 1,100 万张图像和 10 亿个掩码。SAM 可以通过点和边界框等提示，对图像中的任意目标进行交互式分割。其强大的泛化能力已在各种下游任务中带来突破并开创新范式 [15]–[17], [37]–[42]。

尽管 SAM 功能强大，但在面对复杂的真实场景时，其性能会下降，例如具有复杂结构的目标 [36] 或伪装物体 [20], [21], [43]。

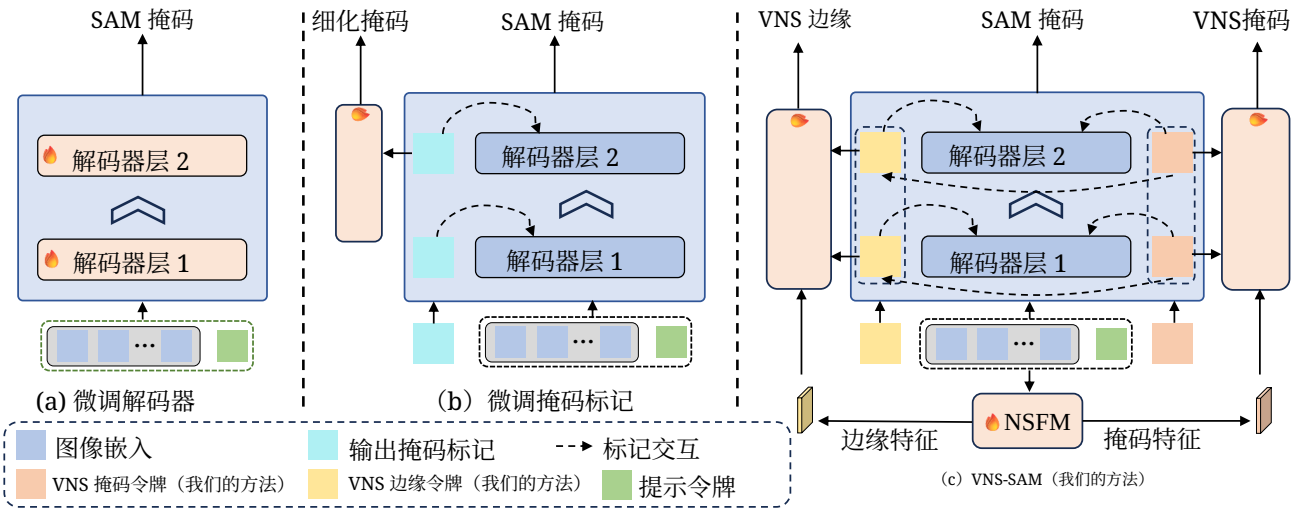



图 2. (a) 对 SAM 的整个解码器进行微调；(b) 对额外的输出掩码标记进行微调，以预测细化掩码；(c) 我们的 VNS-SAM 融合了边缘语义交互与双层解码器层增强。通过非显著特征挖掘（NSFM）模块提取编码器中的信息丰富的掩码和边缘特征，为预测层的表示提供增强，从而实现精准分割。（最佳彩色观看）

提升 SAM 在此类挑战性场景中的能力，是一个值得研究的课题。基于 SAM，已有一些改进变体被提出，大致可分为两条路线。一条路线 [12], [23], [26], [27] 是通过针对特定领域进行微调，将 SAM 用于特定下游任务。这些工作通常侧重于提升 SAM 在特定任务或数据集上的表现，但会牺牲模型固有的泛化能力。另一条路线 [36], [42], [44]–[47] 则是在保留其强大泛化能力的同时扩展 SAM 的能力。例如，MobileSAM [46] 和 EfficientSAM [47], 通过知识蒸馏等技术，使其能够用于实时分割。ASAM [45]通过对抗训练增强 SAM 的泛化能力。HQ-SAM [36]通过添加适配层并冻结 SAM 的原始参数，提升了 SAM 对复杂结构目标的分割质量。与这些现有方法不同，我们的方法旨在从统一的视角提升 SAM 在视觉非显著场景中的分割能力，同时保留其泛化能力。

**视觉非显著场景中的目标分割**。不同于一般场景，真实世界中存在一些具有挑战性的场景：目标的前景与背景具有相似的纹理和颜色，使目标难以被检测。我们将具有这一特征的场景称为视觉非显著（VNS）场景，例如，伪装场景 [48]–[57] 、医学图像中的息肉组织 [31], [58]–[62], 以及低光照环境 [63]–[65]。如何准确感知和理解这些 VNS 场景仍然是一个具有挑战性的问题。一些相关工作通常设计特定于任务的模型结构，例如特征编码器和掩码解码器，以解决单一任务。例如，在伪装目标检测任务中，SINet [48] 设计了一种仿生网络，用于逐步搜索和定位伪装目标。在医学领域的息肉分割任务中，PraNet [62] 集成了反向注意力模块，以突出息肉与其周围区域之间的边界。然而，这些方法——以及它们所使用的数据集都是特定于任务的（一个模型解决一个任务）。不同于现有工作，我们试图从另一种角度解决这一问题。我们构建了一个统一的非显著数据集，使模型能够有效学习更加稳健的 VNS 特征。此外，在 SAM 的基础上，我们开发了一种通用分割模型，在多个非显著场景中实现了优异的性能，同时保持了强大的泛化能力。

**边缘增强分割方法**。许多分割方法将有效的低级特征（例如边缘信息）引入网络，以增强模型对局部细节的感知能力，从而提升模型的分割能力 [28]–[31], [66]–[69]。这些方法的核心在于设计边缘感知模块，以捕获更丰富的上下文和细节信息，进而实现精准分割。例如，Zhou 等 [30] 设计了边界引导模块，用于学习边界增强的特征表示，以进行伪装目标检测。在本文中，我们提出利用 SAM 中的低级特征，并通过增强对目标边缘的感知，促使 SAM 解码器高效学习 VNS 特征，从而提升 VNS场景中的分割性能。

## III. VNS-SAM：视觉非显著分割一切模型

接下来，我们着重提升 SAM 在视觉非显著（VNS）场景中的分割质量，并开发一种更强大的泛化分割模型。首先，我们分析 SAM 在一系列前景与背景对比度较低的场景中的局限性，并在第 III-A 节中将这些场景统称为视觉非显著场景。我们强调，这些场景相当常见，而模型表现不佳极大地限制了 SAM 在现实中的应用。为了解决这些问题，我们提出了两种新技术，以提升 SAM 对 VNS 场景的泛化能力。首先，在第 III-B 节中，我们提出一种掩码-边缘令牌交互解码器，促使 SAM 解码器

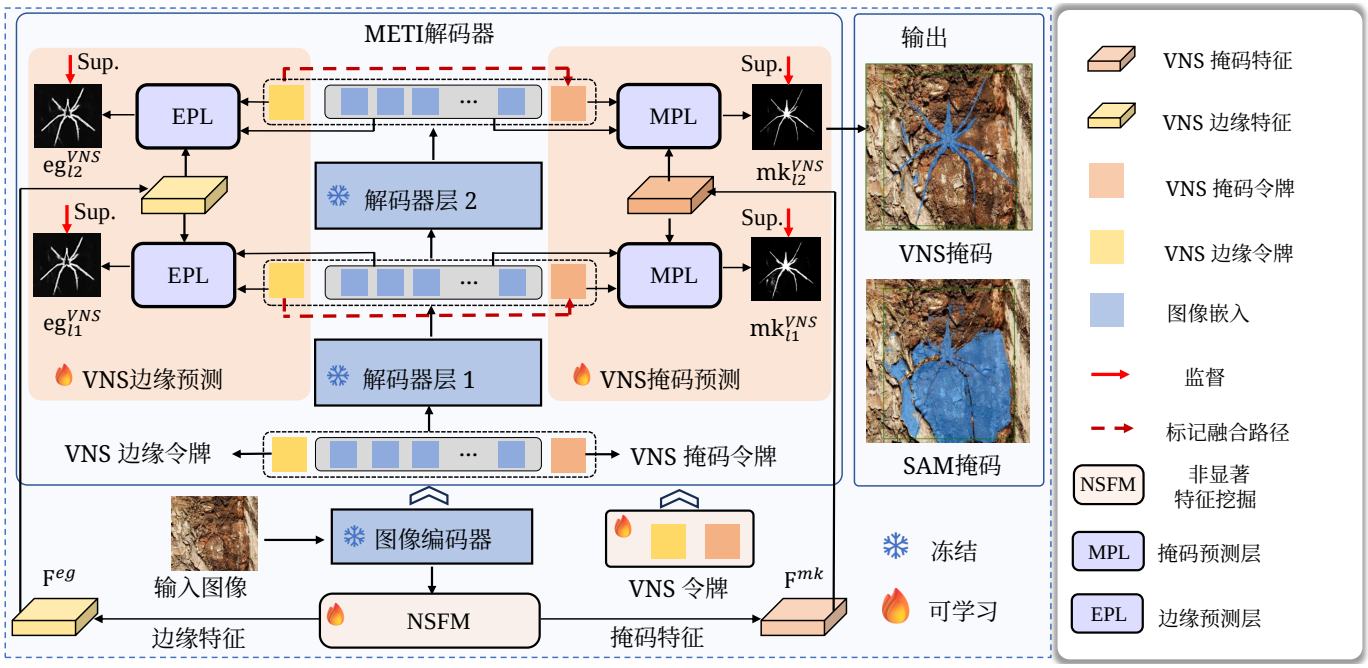



图3. 总体概览。在 SAM 的基础上，VNS-SAM 进行了两项关键改进。首先，通过边缘语义交互和双层解码器层增强，将 SAM 的原始解码器改进为掩码-边缘令牌交互（METI）解码器。其次，设计了一个轻量级非显著特征挖掘（NSFM）模块，从图像编码器中提取掩码和边缘特征，以丰富掩码和边缘预测层（即，MPL 和 EPL）的表征。在训练过程中，预训练 SAM 的参数保持冻结，仅训练 VNS-SAM 中新增的参数。在推理过程中，VNS-SAM 输出更精确的 VNS 掩码和原始 SAM 掩码。此处省略了提示编码器和提示令牌。

显式感知有用的非显著信息。具体而言，如图 3 所示，我们使用一对可学习的 VNS 令牌和双层增强来实现这一点。其次，在第 III-C 节中，我们开发了一个非显著特征挖掘模块，以丰富特征表示，提升 VNS 掩码预测的质量。这两种方法都很轻量，我们将展示它们如何显著提升 SAM 在 VNS 场景中的分割性能。

### *A.* 视觉非显著场景与 *SAM* 的局限性

与一般场景不同，真实世界中存在许多具有挑战性的场景，其中前景与背景的对比度较低，纹理和颜色相似，因此难以准确感知目标对象。例如，如图1所示，伪装物体与周围环境极为相似，使其难以被天敌捕食。在医学图像中，息肉与正常组织具有相同的纹理，且大多体积较小，给医学图像分析带来了挑战。此外，低照度条件下的物体与背景之间缺乏明显的颜色对比度。本文中，我们将此类场景的特征统称为视觉非显著（***VNS***）特征，并将具有 ***VNS*** 特征的场景称为 ***VNS*** 场景。由于缺乏提取 VNS 特征的能力，且没有相应的数据集用于训练，SAM 在 VNS 场景中的表现通常较差。

如图 1 所示，我们可以发现，SAM 难以感知 VNS 对象的前景，从而导致分割错误。这表明 SAM 在 VNS 场景中的鲁棒性较弱。为解决这一问题，不同于以往的方法，我们寻求持续增强 SAM 的分割能力，使其适用于各种 VNS场景，同时保留其原始数据的泛化能力。为此，我们在本节剩余部分设计了两种有效的技术。

### *B.* 掩码-边缘令牌交互（*METI*）解码器

在方法的第一部分，我们旨在促使 SAM 的解码器更深入地学习 VNS 特征。以往的一些方法 [28]–[31], [70] 证明，提取和学习低级特征（如边缘和局部细节）对于感知 VNS 目标至关重要。这促使我们充分利用 SAM 的低级特征，以增强对辨识度较低特征的感知，而这一方向鲜少得到研究。为此，我们将边缘语义融入 SAM 解码器。具体而言，我们引入一对视觉非显著令牌和双层级预测增强，构建了掩码-边缘令牌交互（METI）解码器。其详细结构如图 3 所示。

**视觉非显著令牌**。我们向 SAM 解码器中添加一对 VNS 令牌，其中包含一个 VNS 掩码令牌 $\mathbf{e}^{mk} \in \mathbb{R}^{1\times 256}$ 和一个 VNS 边缘令牌 $\mathbf{e}^{eg} \in \mathbb{R}^{1\times 256}$ 。通过复用 SAM 的原始层，VNS 令牌会与原始预训练输出令牌和提示令牌进行拼接。随后，这些令牌与图像嵌入（定义为 $\{\mathbf{e}^{sam}, \mathbf{e}^{mk}, \mathbf{e}^{eg}\}$）一起输入掩码解码器。在每个解码器层中，我们复用原始掩码解码器中的双向变换器块，分别实现令牌之间以及令牌与图像嵌入之间的特征交互。

$$\mathbf{F}, \{\mathbf{e}^{sam}, \mathbf{e}^{mk}, \mathbf{e}^{eg}\} \leftarrow \Phi_{\text{twt}}(\Phi_{\text{twt}}(\mathbf{F}, \{\mathbf{e}^{sam}, \mathbf{e}^{mk}, \mathbf{e}^{eg}\})), \tag{1}$$

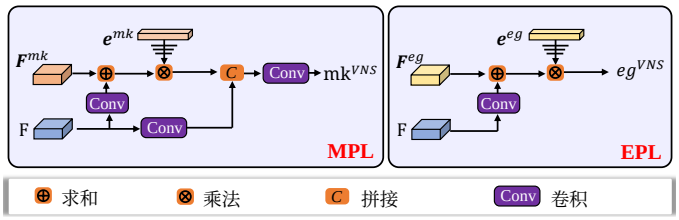



图 4. 掩码预测层（MPL）和边缘预测层（EPL）的细节。在 MPL 中，我们将经过高度优化的图像嵌入 F 重用为补充特征。（彩色查看效果最佳）

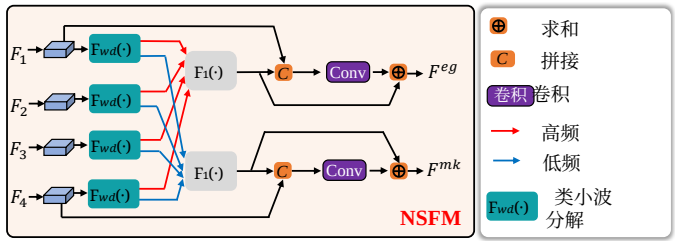



图 5. 非显著特征挖掘（NSFM）模块的细节。骨干网络提取的多层级特征被分解为不同的分量。随后，选取信息量最大的高频和低频分量，并聚合多层级特征，以提取边缘和掩码特征。（彩色查看效果最佳）

其中，左侧的 $\mathbf{F}$ 和 $\{\mathbf{e}^{sam},\mathbf{e}^{mk},\mathbf{e}^{eg}\}$ 分别表示经过两个解码器层更新后的嵌入特征和令牌。 $\Phi_{\text{twt}}(\cdot)$ 表示双向变换器层（包含一个自注意力单元以及一个图像到令牌和令牌到图像注意力模块）。在训练过程中，VNS 边缘令牌充当低级特征学习器，为 VNS 掩码令牌提供有效的低级信息。具体而言，在解码过程中，VNS 令牌之间的交互通过两种方式实现。一方面，VNS 边缘令牌通过原始解码器层中的交叉注意力机制与掩码令牌进行隐式交互，将有效的边缘和纹理信息传递给掩码令牌。此外，我们还通过一种直接的融合操作，显式增强两个令牌之间的交互。在每个解码器层之后，VNS 边缘令牌也会与 VNS 掩码令牌进行显式融合，如

$$\mathbf{e}^{mk} \leftarrow \mathcal{F}_{token}(\mathbf{e}^{mk}, \mathbf{e}^{eg}). \tag{2}$$

$\mathcal{F}_{token}(\cdot)$ 表示词元整合操作。这两个词元首先执行逐元素相加，随后通过线性层进行融合。这一简单操作进一步实现了将边缘表征从 VNS-边缘令牌显式聚合到 VNS 掩码令牌。

**双层级预测增强**。为了使两个解码器层都能持续学习细微的判别性特征，从而提升解码器理解非显著字符的能力，我们通过层级监督和预测增强，将原始解码器改进为双层级增强解码器。**层级监督：**在每个解码器层之后，将交互后的 VNS 掩码令牌和 VNS 边缘令牌输入增强掩码预测层（MPL）和边缘预测层（EPL），以获得掩码和边缘预测。在训练阶段，掩码和边缘预测由两个解码器层分别生成，并由真实标注进行监督。**预测增强：**如图 4 所示，在 MPL 和 EPL 中，首先分别将 VNS 掩码特征 $F^{mk}$ 和 VNS 边缘特征 $F^{eg}$ 与图像嵌入 $\mathbf{F}$融合。然后，每个 VNS 令牌（$\mathbf{e}^{mk}$ 和 $\mathbf{e}^{eg}$）都经过一个可学习的多层感知机层，再与相应的融合特征进行点积，得到用作 EPL 中边缘预测的单通道输出。此外，在 MPL 中，为了进一步利用 SAM 中经过高度优化的图像嵌入$\mathbf{F}$，我们将其输入两个可学习的卷积层，得到一个与 VNS 掩码令牌输出尺寸相同的单通道特征图，作为补充特征。最后，将这两个特征图拼接，并输入两个卷积层，其卷积核大小分别为 $3\times3$ 和 $1\times1$，以融合它们并输出最终的掩码预测。

### C. 非显著特征挖掘

在方法的第二部分，我们旨在从经过高度优化的图像编码器中挖掘有用的低级特征，以丰富预测层的表征。基于这一目标，我们提出了图 5 所示的可学习 NSFM 模块，它能有效地从多层级图像编码器中提取 VNS 掩码和边缘表征。生物学研究表明，特征的判别性特征主要存在于高频和低频分量中[71]。因此，我们首先对提取的特征进行分解，以获得不同的分量。随后，我们选取并聚合信息最丰富的分量，进一步进行掩码提取和边缘提取。需要注意的是，所提出的模块十分轻量，仅增加约 3M 个参数。

具体而言，对于由图像编码器提取的多层级特征，我们采用数学上严谨且广泛用于特征分析和分割的 Haar 离散小波分解 [72]，对多层级特征进行分解。具体来说，Haar 离散小波将每个特征分解为四个小波子带，$F_k^{HH}$，$F_k^{HL}$，$F_k^{LH}$，$F_k^{LL}$。

$$F_k^{HH}, F_k^{HL}, F_k^{LH}, F_k^{LL} = \mathcal{F}_{wd}(F_k), k = 1, 2, 3, 4, \tag{3}$$

其中，$\mathcal{F}_{wd}(\cdot)$表示 Haar 小波分解。我们选取信息量最大的高频 $F_k^{HH}$和低频 $F_k^{LL}$分量。随后，分别聚合多层高频和低频分量，以获得增强表示 $F_{agg}^{HH}$和 $F_{agg}^{LL}$。以高频分量为例，首先将多层高频特征进行拼接，并通过 $1\times1$ 卷积层进行通道降维。然后，应用有效的注意力层 [75], [76] 来探索层间特征相关性。经过上述操作后，我们获得了多层融合高频特征。上述操作可表示为：

$$F_{agg}^{HH} = \mathcal{F}_1(\{F_k\}_{k=1}^4) = \text{Attn}(\text{Conv}(\text{Cat}(\{F_k^{HH}\}_{k=1}^4))), \tag{4}$$

其中，Attn(·)表示连接通道注意力层和空间注意力层的注意力层。同样地，可以得到聚合后的低频特征 $F_{agg}^{LL}$。

表 I 我们的 VNS-SEG 训练集的数据组成。该训练集共包含 23,232 张图像，来源于知名的现有数据集和合成数据。收集的图像包含多种视觉非显著字符。

| 训练集 | 现有数据集 | | | | 合成数据集 | | | Sum |
|---|---|---|---|---|---|---|---|---|
| | CAMO [49] | COD10K [48] | Kvasir [61] | Clin.DB [59] | DIS-Dark [77] | Thin-Dark [78] | FSS-Dark [79] | |
| 数量 | 1000 | 3040 | 900 | 550 | 3000 | 4742 | 10000 | 23232 |

表 II VNS-SEG 评估集的数据组成，包含 4 种 VNS 场景下的 11 个子集。该评估集分为已见集和未见集，以全面评估模型在 VNS 场景中的分割性能和泛化能力。请注意，未见集中的所有数据均采集自真实场景，从而确保有效评估 VNS-SAM 在实际应用中的表现。

| *Eval-Seen-Set* | CAMO [49] | COD10K [48] | Kvasir [61] | ClinicDB [59] | DIS-Dark [77] | Thin-Dark [78] | Sum |
|---|---|---|---|---|---|---|---|
| 数量 | 250 | 2026 | 100 | 62 | 480 | 1000 | |
| 未见集评估 | NC4K [50] | ColonDB [60] | ETIS [58] | LIS [63] | CDS2K [51] | - | 12175 |
| 数量 | 4121 | 380 | 196 | 2230 | 1330 | - | |

如前文所述，高频分量包含丰富的纹理和边缘信息，因此我们利用它们提取视觉非显著边缘特征，而低频分量则用于提取视觉非显著掩码特征。对于高频部分，将浅层 $F_1$ 与 $F_{agg}^{HH}$ 拼接作为补充信息，然后使用 1×1 卷积层融合拼接后的特征，并降低通道维度。最后，通过跳跃连接将高频分量与补充表示进行合并，生成VNS 边缘特征 $F^{eg}$。同样地，我们可以得到VNS 掩码特征$F^{mk}$。如上文所述，VNS掩码和VNS 边缘特征分别用于在MPL和EPL中进行掩码预测和边缘预测。

### *D.* 训练与推理

**训练。** 在训练过程中，我们冻结预训练SAM的权重，仅更新新添加模块中的参数。我们使用多种采样提示，包括边界框、随机选取的点和粗略掩码。图像和提示被输入VNS-SAM，并在解码器中生成两个层级的掩码和边缘预测。对于掩码监督，我们采用结构损失

$\mathcal{L}^{stru}$ [80]该损失包含加权 IoU 损失和加权二元交叉熵损失，能够更加关注难分像素。对于边缘监督，我们使用 Dice 损失 $\mathcal{L}^{dice}$ [81]。总损失的公式如下：

$$\mathcal{L}_{total} = \sum_{k=1}^{2} \mathcal{L}_{l_k}^{stru}(\mathbf{mk}_{l_k}^{VNS}, \mathbf{mk}^{gt}) + \sum_{k=1}^{2} \mathcal{L}_{l_k}^{dice}(\mathbf{eg}_{l_k}^{VNS}, \mathbf{eg}^{gt}). \tag{5}$$

**推理。** 在推理阶段，我们舍弃边缘令牌的输出以及第一层的掩码输出。仅计算第二个解码器层的 VNS 掩码和原始 SAM 的输出。我们将预测掩码上采样至原始图像的分辨率，作为最终输出。

## IV. VNS-SEG：视觉非显著分割数据集

为了使分割模型能够有效学习 VNS 特征，我们精心构建了一个统一数据集：VNS-SEG，用于训练分割模型，并在多种 VNS 场景下评估其性能。该数据集包含超过 35K 对图像-掩码对。

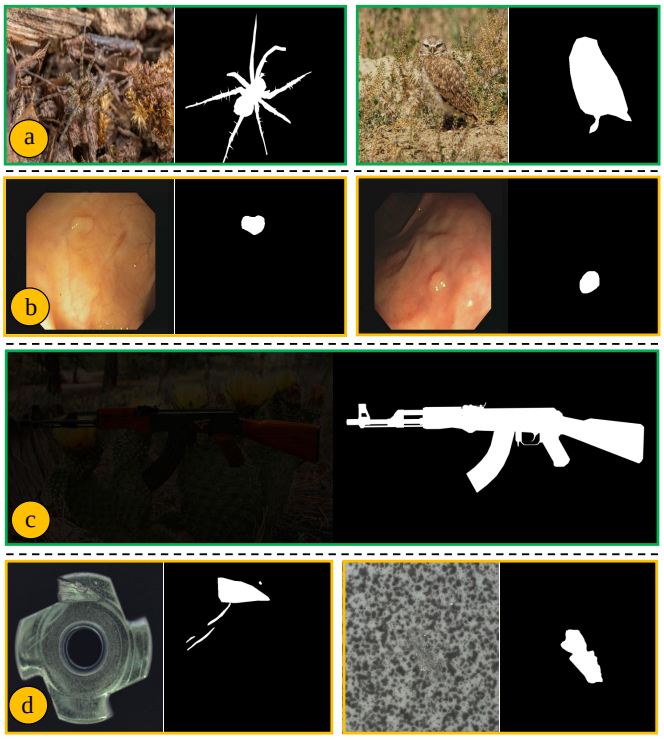



图 6. VNS-SEG 中包含多种 VNS 场景的图像及其对应掩码示例，即：(a) 伪装物体，(b) 息肉组织，(c) 低光照条件下的目标，以及 (d) 工业缺陷。可以看出，这些目标与背景高度相似，难以察觉，因此对当前的分割模型构成了挑战。

VNS-SEG 中的图像取自社区中一些知名的现有数据集以及合成数据，以丰富数据集的多样性。统一数据集使模型能够学习更鲁健的非显著特征，并提升其在多种 VNS 任务上的表现。图 6 展示了从 VNS-SEG 中随机选取的图像及对应的掩码标注。与普通目标相比，VNS 场景中的目标

表III VNS-SEG的***Eval-Seen-Set*** 性能比较。使用了三种类型的提示。我们在VNS-SEG数据集上对HQ-SAM进行了微调，将其称为HQ-SAM-F。我们的模型在不同的子数据集和提示下，始终优于基线SAM和其他竞争方法。粗体字表示最佳结果。

| 方法 | CAMO | | | | COD 10K | | | | Kvasir | | | | ClinicDB | | | | DIS-Dark | | | | Thin-Dark | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 交并比 | 边界交并比 | $E_\phi$ | $F_\beta^w$ | 交并比 | 边界交并比 | $E_\phi$ | $F_\beta^w$ | 交并比 | 边界交并比 | $E_\phi$ | $F_\beta^w$ | 交并比 | 边界交并比 | $E_\phi$ | $F_\beta^w$ | 交并比 | 边界交并比 | $E_\phi$ | $F_\beta^w$ | 交并比 | 边界交并比 | $E_\phi$ | $F_\beta^w$ |
| 基于点提示的评估 | | | | | | | | | | | | | | | | | | | | | | | | |
| SAM [9] | .659 | .450 | .757 | .633 | .687 | .593 | .830 | .665 | .719 | .480 | .800 | .652 | .602 | .447 | .753 | .528 | .594 | .476 | .723 | .558 | .775 | .617 | .827 | .752 |
| HQ-SAM [36] | .769 | .522 | .886 | .790 | .771 | .672 | .922 | .803 | .799 | .568 | .883 | .807 | .727 | .560 | .851 | .735 | .721 | .620 | .872 | .752 | .853 | .714 | .916 | .867 |
| HQ-SAM-F [36] | .780 | .542 | .880 | .769 | .774 | .675 | .911 | .771 | .782 | .554 | .856 | .695 | .690 | .500 | .828 | .634 | .720 | .624 | .861 | .723 | .854 | .718 | .905 | .840 |
| VNS-SAM（我们的 | | | | | | | | | | | | | | | | | | | | | | | | |
| 基于噪声框提示的评估 | | | | | | | | | | | | | | | | | | | | | | | | |
| SAM [9] | .740 | .510 | .852 | .765 | .725 | .629 | .879 | .751 | .756 | .524 | .853 | .784 | .828 | .648 | .944 | .859 | .529 | .434 | .695 | .545 | .656 | .514 | .716 | .646 |
| HQ-SAM [36] | .756 | .517 | .890 | .801 | .747 | .646 | .913 | .785 | .802 | .585 | .890 | .824 | .821 | .655 | .947 | .859 | .682 | .583 | .845 | .715 | .799 | .653 | .870 | .801 |
| HQ-SAM-F [36] | .761 | .531 | .890 | .803 | .773 | .679 | .925 | .804 | .824 | .612 | .909 | .843 | .836 | .678 | .957 | .868 | .708 | .620 | .865 | .738 | .823 | .682 | .888 | .824 |
| VNS-SAM（我们的 | | | | | | | | | | | | | | | | | | | | | | | | |
| 基于真实标注框提示的评估 | | | | | | | | | | | | | | | | | | | | | | | | |
| SAM [9] | .757 | .532 | .864 | .779 | .746 | .650 | .892 | .771 | .762 | .534 | .861 | .798 | .818 | .652 | .946 | .862 | .579 | .475 | .722 | .584 | .652 | .516 | .689 | .631 |
| HQ-SAM [36] | .776 | .535 | .898 | .812 | .757 | .658 | .918 | .794 | .810 | .594 | .888 | .829 | .825 | .671 | .946 | .859 | .703 | .605 | .854 | .729 | .810 | .672 | .870 | .807 |
| HQ-SAM-F [36] | .783 | .553 | .902 | .819 | .785 | .692 | .930 | .813 | .837 | .627 | .910 | .849 | .842 | .686 | .956 | .873 | .727 | .637 | .872 | .752 | .838 | .705 | .894 | .834 |
| VNS-SAM（我们的 | | | | | | | | | | | | | | | | | | | | | | | | |

更难被观察到，这大大增加了精确分割的难度，也对当前模型提出了挑战。

### *A.* 数据集构建

**训练集**。训练集包含 23,232 张带有精确掩码标注的图像。为构建训练集，我们从社区中精心选择了四个具有挑战性 VNS 特征的知名数据集，包括 COD10K [48] 和 CAMO [49]、Kvasir [61], 和 ClinicDB [59]。具体而言，COD10K 和 CAMO 来自伪装数据集，分别包含 3040 张和 1000 张图像。Kvasir 和 ClinicDB 来自息肉分割数据集，分别包含 900 张和 550 张图像。此外，为丰富数据的多样性，我们合成了一些低光照条件下的图像。具体做法是训练一个 CycleGAN [82] 模型，将正常图像转换为低光照图像。我们选择具有细粒度细节和复杂几何形状的 DIS [77], 、 ThinObject-5K [78], 和 FSS [79] 数据集作为源数据集，并将其转换为低光照数据集，即 *DIS-Dark*、*Thin-Dark* 和 *FSS-Dark*。这样，我们便可以直接利用原始数据集中的精确掩码标注，同时合成数据也能显著丰富数据集的多样性。具体组成细节如表 I 所示。

**评估集**。 评估集包含 11 个子集，并划分为已见集和未见集，以全面评估模型在 VNS 场景中的分割能力和泛化能力。详情见表 II。

已见集包含 6 个子集，即CAMO（250 张图像） [49], COD10K（2026 张图像） [48], Kvasir（100 张图像）[61], ClinicDB（62 张图像） [59], DIS-Dark（480 张图像） [77], 和 Thin-Dark（1000 张图像） [78]。具体而言，我们在这些数据集的测试集上评估各方法。请注意，如上所述，DIS-Dark 和 Thin-Dark 也是通过 CycleGAN[82] 合成的数据。这些数据与训练集中的数据属于相同的数据集。

未见集包含 5 个子集：NC4K [50],、ColonDB [60], 、ETIS [58], 、LIS [63], 和 CDS2K [51]。请注意，这些子集均采集自真实场景，因而能更好地评估模型在现实世界应用中的性能。具体而言，NC4K 是一个大规模伪装目标检测测试数据集，包含 4121 张图像。ColonDB 和 ETIS 是常用的息肉分割数据集，分别包含 380 张和 196 张图像。LIS 是一个低光照条件下的现实世界实例分割数据集。我们使用其 RGB-dark 测试集进行评估，其中包含 2230 张图像。此外，未见集还包括一个训练集中未涵盖的新型 VNS 场景：工业缺陷数据集 CDS2K。CDS2K 中的图像选自真实工业缺陷数据库，包含正样本和负样本划分。我们仅使用其中包含 1330 张图像的正样本进行评估。缺陷区域相对较小，且与背景具有相似的模式，因此极具挑战性。

### *B.* 评估指标

我们采用五项指标来评估模型的性能：

- **交并比** (IoU) 是一种广泛使用的分割准确率度量指标，用于衡量预测分割掩码与真实分割掩码之间的重叠程度。
- **边界交并比** (BIoU) [83]是传统 IoU 指标的扩展。它聚焦于分割目标的边界，从而能够更灵敏地评估边界准确率。
- **增强对齐度量**（$E_\phi$） [84] 是一种二值前景评估指标，适用于衡量二值图之间的局部和全局相似度。请注意，实验中报告的是均值 $E_\phi$ 。
- **加权 F 度量**($F_\beta^w$) [85] 基于 F 度量。它融入了空间信息，更重视对目标边界附近区域的准确分割，而较少关注背景。

表IV VNS-SEG在未见集评估上的性能比较。未见集中的数据均来自真实场景。我们的方法始终优于其他竞争方法，凸显了其在广泛实际应用中的潜力。

| 方法 | NC4K | | | | ColonDB | | | | ETIS | | | | CDS2K | | | | LIS | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 交并比 | 边界交并比 | $E_\phi$ | $F_\beta^w$ | 交并比 | 边界交并比 | $E_\phi$ | $F_\beta^w$ | 交并比 | 边界交并比 | $E_\phi$ | $F_\beta^w$ | 交并比 | 边界交并比 | $E_\phi$ | $F_\beta^w$ | 平均精度 | 平均精度$_{50}$ | 平均精度$_{75}$ |
| 基于点提示的评估 | | | | | | | | | | | | | | | | | | | |
| SAM [9] | .713 | .553 | .822 | .685 | .568 | .408 | .724 | .481 | .613 | .526 | .765 | .529 | .415 | .363 | .620 | .400 | - | - | - |
| HQ-SAM [36] | .794 | .625 | .915 | .817 | .722 | .549 | .864 | .744 | .731 | .633 | .876 | .751 | .558 | .488 | .795 | .617 | - | - | - |
| HQ-SAM-F [36] | .799 | .634 | .904 | .792 | .680 | .488 | .833 | .611 | .713 | .601 | .862 | .654 | .483 | .419 | .715 | .461 | - | - | - |
| VNS-SAM（我们的方法） | **.834** | **.683** | **.946** | **.868** | **.810** | **.624** | **.953** | **.871** | **.852** | **.745** | **.971** | **.881** | **.618** | **.544** | **.858** | **.677** | - | - | - |
| 基于噪声框提示的评估 | | | | | | | | | | | | | | | | | | | |
| SAM [9] | .760 | .592 | .882 | .784 | .822 | .616 | .954 | .874 | .851 | .732 | .961 | .886 | .593 | .517 | .815 | .644 | .298 | .570 | .281 |
| HQ-SAM [36] | .780 | .608 | .912 | .813 | .826 | .639 | .955 | .875 | .816 | .709 | .942 | .866 | .551 | .480 | .811 | .633 | .303 | .568 | .286 |
| HQ-SAM-F [36] | .801 | .640 | .921 | .828 | .841 | .650 | .963 | .888 | .866 | **.765** | .972 | .896 | .601 | .526 | .837 | .659 | .308 | .575 | .289 |
| VNS-SAM（我们的方法） | **.810** | **.657** | **.933** | **.847** | **.842** | **.657** | **.964** | **.895** | **.873** | .764 | **.979** | **.907** | **.617** | **.543** | **.848** | **.679** | **.318** | **.594** | **.304** |
| 基于真实标注框提示的评估 | | | | | | | | | | | | | | | | | | | |
| SAM [9] | .780 | .614 | .893 | .802 | .830 | .628 | .959 | .881 | .856 | .741 | .964 | .891 | .601 | .527 | .818 | .648 | .439 | .784 | .433 |
| HQ-SAM [36] | .792 | .623 | .917 | .822 | .829 | .647 | .956 | .877 | .828 | .724 | .946 | .874 | .560 | .491 | .815 | .640 | .437 | .773 | .428 |
| HQ-SAM-F [36] | .815 | .659 | .928 | .839 | .847 | .668 | .968 | .892 | .876 | .778 | .976 | .903 | .606 | .536 | .839 | .664 | .445 | .779 | .437 |
| VNS-SAM（我们的方法） | **.830** | **.685** | **.941** | **.860** | **.857** | **.684** | **.969** | **.904** | **.888** | **.790** | **.982** | **.918** | **.626** | **.559** | **.851** | **.687** | **.461** | **.787** | **.457** |

- **平均精度**（AP）将精确率-召回率曲线概括为一个数值，反映不同阈值下精确率与召回率之间的权衡。在我们的实验中，它用于评估实例分割的性能。

## V. 实验

本节将对所提出的 VNS-SAM 在 VNS-SEG 基准上的表现进行全面评估，包括已见集和未见集评估。我们还在通用 COCO [86] 基准上进行零样本实例分割。首先在第 V-A 节介绍实现细节，然后在第 V-B 节将 VNS-SAM 与基线及其他竞争方法进行比较。我们在第 V-C 节开展消融研究。之后，第 V-D 节将展示关于 VNS-SAM 和 VNS-SEG 的更多实验及进一步分析。最后，我们在第 V-E 节进行定量可视化。

### *A.* 实验详情

**实现细节**。 在训练阶段，VNS-SAM 使用所提出的 VNS-SEG 在 4× 4090 块 GPU 上训练 12 个训练轮次，仅耗时 4 小时。我们使用 Adam 优化器，初始学习率为 0.001（在训练 10 个轮次时乘以 10×），批量大小为 16。除非另有说明，实验默认使用基于 ViT-L 的模型。

在推理阶段，我们遵循 SAM 的相同流程，但将 VNS 掩码令牌生成的掩码预测作为 VNS 对象的结果。我们全面评估了模型在各种提示下的性能，包括边界框、噪声边界框和随机点。对于基于边界框提示的评估，我们使用真实掩码生成真实边界框，并将其作为边界框提示输入。对于基于噪声边界框提示的评估，我们按照[26]，通过向真实边界框添加噪声来生成噪声边界框，并将其作为提示输入。在我们的实验中，噪声比例默认设为 0.1。对于基于点提示的评估，我们随机采样从真实标注掩码中选取若干点，并将其用作输入提示。在我们的实验中，随机点的数量默认设为 10。

表 V 医学领域中其他特定领域 SAM 变体的比较。VNS-SAM 具有更强的泛化能力，配备用于任务特定适配的适配器后，性能还会进一步提升。

| 方法 | ClinicDB | | Kvasir | | ColonDB | | ETIS | | Avg. | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 交并比 | 边界交并比 | 交并比 | 边界交并比 | 交并比 | 边界交并比 | 交并比 | 边界交并比 | 交并比 | 边界交并比 |
| SAM [9] | .788 | .574 | .725 | .476 | .795 | .567 | .838 | .703 | .787 | .580 |
| SAM-Med2D [24] | .857 | .681 | **.866** | .628 | .825 | .614 | .781 | .646 | .832 | .642 |
| MedSAM [25] | .851 | .696 | .858 | **.692** | .830 | .638 | .840 | .738 | .845 | .691 |
| VNS-SAM（我们的+编码器适配器 | **.8** | | | | | | | | | |

表 VI VNS-SAM 各组件的消融研究。以原始数据 SAM 和微调后的 SAM（*FT*-解码器）作为基线。VNS-T 表示 VNS 令牌，DPE 表示双层级预测增强。每个引入的模块都对性能产生了积极影响。

| 模块 | METI | | NSFM | 交并比 | 边界交并比 | $E_\phi$ | $F_\beta^w$ | 可学习参数 |
|---|---|---|---|---|---|---|---|---|
| | VNS-T | DPE | | | | | | |
| SAM [9] | - | - | - | .720 | .561 | .825 | .736 | - |
| 微调-解码器 | - | - | - | .777 | .619 | .893 | .793 | 15.0 M |
| VNS-SAM | ✓ | | | .797 | .643 | .908 | .818 | 1.1 M |
| | ✓ | ✓ | | .809 | .663 | .917 | .833 | 6.1 M |
| | ✓ | | ✓ | .814 | .667 | .917 | .830 | 4.8 M |
| | ✓ | ✓ | ✓ | **.821** | | | | 9.8 M |

**低光照数据集合成**。我们首先在从真实场景采集的成对 LOL [65] 数据集上训练 CycleGAN[82]，共训练 100 个训练轮次，初始学习率为 0.0002，并在训练 50 个轮次后降低学习率。随后，使用预训练的 CycleGAN 将我们选定的数据集转换为低光照数据集。

表 VII VNS-TOKEN 的消融研究。VNS-MT 和 VNS-ET 分别表示 VNS掩码令牌和 VNS 边缘令牌。VNS-ET 对性能有积极贡献，尤其是对边界交并比的提升，表明其有效性。

| | 模块 | IoU | BIoU | $E_\phi$ | $F_\beta^w$ |
|---|---|---|---|---|---|
| 不使用 NSFM | 使用 VNS-MT | .795 | .632 | .900 | .809 |
| | 使用 VNS-MT+VNS-ET | **.805** | **.660** | **.911** | **.825** |
| 使用 NSFM | 使用 VNS 掩码标记 | .806 | .660 | .903 | .818 |
| | 使用 VNS 掩码标记+VNS 边缘标记 | **.821** | **.684** | **.929** | **.850** |

表 VIII 掩码和边缘预测层设计分析。

| 方法 | 可学习参数 | Eval-Seen-set 交并比 | 边界交并比 | $E_\phi$ | $F_\beta^w$ | 未见集评估 交并比 | 边界交并比 | $E_\phi$ | $F_\beta^w$ |
|---|---|---|---|---|---|---|---|---|---|
| EPL&EPL | 8.7 M | .820 | .670 | .928 | .844 | .790 | .663 | .932 | .831 |
| MPL&MPL | 10.9 M | **.821** | **.684** | .928 | .848 | **.802** | **.681** | **.936** | **.844** |
| MPL&EPL | 9.8 M | **.821** | | | | | | | |

表 IX 令牌融合方法分析。我们的方法优于其他替代方法。“TF”表示令牌融合，“CA”表示交叉注意力，“DG”表示动态门控。

| 方法 | 可学习参数 | Eval-Seen-set 交并比 | 边界交并比 | $E_\phi$ | $F_\beta^w$ | 未见集评估 交并比 | 边界交并比 | $E_\phi$ | $F_\beta^w$ |
|---|---|---|---|---|---|---|---|---|---|
| 不使用 TF | 9.7 M | .809 | .675 | .927 | .848 | .791 | .671 | .932 | .840 |
| CA | 9.9 M | .818 | .677 | .926 | .844 | .791 | .669 | .930 | .834 |
| DG | 10.2 M | .820 | .682 | **.929** | **.850** | **.802** | **.681** | **.936** | .841 |
| Ours | 9.8 M | **.821** | **.684** | **.929** | | | | **.936** | |

表 X NSFM 模块的详细消融实验。“WD”表示 Haar 小波分解，“HF”和“LF”分别表示高频和低频分量。

| 方法 | Eval-Seen-Set 交并比 | 边界交并比 | $E_\phi$ | $F_\beta^w$ | 未见集评估 交并比 | 边界交并比 | $E_\phi$ | $F_\beta^w$ |
|---|---|---|---|---|---|---|---|---|
| 不使用权重衰减 | .816 | .679 | .924 | .850 | .786 | .663 | .932 | .840 |
| 仅使用高频 | .814 | .674 | .920 | .841 | .790 | .670 | .933 | **.842** |
| 仅使用 LF 的 w | .814 | .675 | .922 | .843 | .795 | .674 | .933 | .841 |
| 我们的方法 NSFM | **.821** | **.684** | **.929** | **.859** | **.800** | **.680** | **.936** | **.842** |

## *B.* 性能比较

在本实验中，我们在 VNS-SEG 基准上评估 VNS-SAM 的性能，包括表 III 中的已见集评估和表 IV 中的未见集评估。我们采用三种不同的提示，全面评估模型的交互式分割性能。除了原始 SAM，我们还将我们的方法与 HQ-SAM 进行比较，后者是原始 SAM 的一种先进变体。此外，我们按照 [36]，中相同的设置，在 VNS-SEG 上微调 HQ-SAM，并将其称为 HQ-SAM-F，以更全面地验证 VNS-SAM 的有效性。

**VNS-SEG 已见集上的性能**。在表 III 中，我们评估了 VNS-SAM 在 VNS-SEG 已见集上的性能，该数据集包含六个子集：CAMO、COD10K、Kvasir、ClinicDB、DIS-Dark 和 Thin-Dark。为评估模型在各种场景下的鲁棒性，我们考虑了三种提示：常用的 GT框提示，以及点提示和噪声框提示等低质量提示。在真实世界的交互式分割应用中，提示可能无法像 GT 框那样始终准确地框住目标。结果清楚地表明，VNS-SAM 在所有六个子集和三种提示下的表现均显著优于基线 SAM。值得注意的是，VNS-SAM 在基于点提示的评估中表现出色，在 ClinicDB（0.602 对比0.833）和 DIS-Dark（0.594 对比 0.780）子集上的得分比基线高出 20 多分，凸显了其有效处理低质量输入的能力。尽管 HQ-SAM 的表现较 SAM 有所提升，但在 VNS 场景中的性能仍不理想。即使在 VNS-SEG 数据集上进行微调，HQ-SAM-F 的性能也有所提升，但仍然有限，表明它未能充分捕捉细微的非显著特征。相比之下，VNS-SAM 始终优于 SAM 和 HQ-SAM-F，展现出更强的能力，能够适应非显著特征和低质量提示带来的挑战。这进一步验证了我们提出的方法在真实世界交互式分割任务中的有效性。

**VNS-SEG 未见集评估上的性能**。在表 IV 中，我们评估了 VNS-SAM 在 VNS-SEG 未见集上的零样本性能。值得注意的是，未见集中的数据均来自真实世界，因此能更好地评估模型在实际应用中的性能。此外，还包括一种训练集中未出现过的新 VNS 场景，即工业缺陷场景。总的来说，由于数据来自真实世界且多种多样，未见集尤其具有挑战性。从结果来看，VNS-SAM 始终优于其他方法。这些结果凸显了 VNS-SAM 强大的零样本泛化能力，也彰显了其在广泛的现实世界应用中的潜力。

**与特定领域 SAM 变体的比较**。在表 V 中，我们将所提出的方法与 SAM 的特定领域变体进行了比较，包括 SAM-Med2D [24] 和 MedSAM[25]，这两者均代表了医学影像领域的先进方法。实验在四个息肉分割数据集 [58]–[61] 上进行。我们的 VNS-SAM 平均交并比更高（0.854），优于 SAM-Med2D （ 0.832 ）和 MedSAM （ 0.845 ），展现出对 VNS 场景的强大泛化能力。此外，我们在 VNS-SAM 的编码器中加入了可学习适配器 [26]（与 SAM-Med2D 类似），以实现更强大的任务特定适配。借助这一设计，我们的模型平均交并比达到 0.875，超过所有特定领域基线，进一步验证了该方法的灵活性和可扩展性。

## *C.* 消融研究

本节中，我们对 VNS-SAM 的各个分量进行了广泛的消融实验，并进一步展开讨论，以阐明其有效性。

**各组件的效果**。 我们进行了消融研究，以考察各个设计组件的效果，包括 VNS 令牌（VNS-T）、双层预测增强（DPE）和非显著特征挖掘模块（NSFM）。结果如表VI所示。我们报告了

表 XI 不同微调（*FT*）策略的比较。*FT*-解码器和 *FT*-词元分别表示对整个解码器进行微调和对输出词元进行微调。

| *FT*-策略 | 已见集 交并比 | 已见集 边界交并比 | 未见集 交并比 | 未见集 边界交并比 | COCO 交并比 | COCO 边界交并比 | 可学习参数 |
|---|---|---|---|---|---|---|---|
| SAM | .720 | .561 | .768 | .628 | .812 | .707 | - |
| *FT*-编码器&解码器 | **.827** | .682 | .771 | .650 | .443 | .334 | 1191 M |
| *FT*-解码器 | .777 | .619 | .699 | .559 | .626 | .507 | 15.0 M |
| *FT*-令牌 | .787 | .625 | .786 | .651 | .811 | .710 | 2.1 M |
| VNS-SAM（我们的方法） | .821 | | **.800** | | | | 9.8 M |

表 XII COCO 集上的性能比较。使用 VNS-score 将 COCO 数据集划分为显著（COCO-S）和非显著（COCO-NS）子集。VNS-SAM 的表现始终优于 SAM，尤其是在非显著场景中，展现了其在严峻条件下的鲁棒性。

| 方法 | COCO全量集 IoU | COCO全量集 BIoU | COCO-S IoU | COCO-S BIoU | COCO-NS IoU | COCO-NS BIoU |
|---|---|---|---|---|---|---|
| SAM [9] | .755 | .642 | .763 | .666 | .732 | .572 |
| VNS-SAM（我们的方法） | **.775** | **.661** | **.778** | **.681** | **.765** | **.604** |

在 Eval-Seen-Set 上的平均性能，以及各项技术引入的额外参数量。我们将原始数据 SAM 和经过微调的 SAM 作为基线，它们在六个数据集上的平均交并比分别为 0.720 和 0.777。**i) VNS-T**：为促使解码器学习 VNS 特征，我们首先添加一对 VNS 令牌，包括一个 VNS 掩码令牌和一个 VNS 边缘令牌。将 VNS 令牌融入模型并在任务数据上进行微调后，性能显著提升：与经过微调的 SAM 相比，性能提高约 2 个百分点，而可学习参数更少（1.1 M 与 15.0 M）。**ii) DPE**：在上一步的基础上，我们将 SAM 解码器改进为双层增强解码器，使两个解码层都能持续学习细微的判别性特征，从而提升解码器理解非显著特征的能力。DPE 持续提升交并比、边界交并比和 $F_\beta^w$，增幅均超过 1 个百分点。**iii)NSFM**：此外，我们引入 NSFM，从经过高度优化的图像编码器中挖掘有用的低级特征，以丰富预测层的表征。在 NSFM 的帮助下，性能持续提升（最后两行），最终交并比达到 0.821；与基线 SAM 相比提高超过 10 个百分点，与经过微调的 SAM 相比提高 5 个百分点。值得注意的是，NSFM 十分轻量，仅引入 3.7 M 个参数。总的来说，表VI中的结果表明，每个引入的模块都能提升 VNS-SAM 的性能。

**META 消融研究**。表VII展示了针对 VNS 令牌的分项消融研究，说明了 VNS 掩码令牌（VNS-MT）和 VNS 边缘令牌（VNS-ET）分别对模型性能的影响。显然，VNS 边缘令牌对提升模型性能至关重要，尤其是在边界交并比方面：不使用 NSFM 时提升 2.8 个点，使用 NSFM 时提升 2.4 个点。

**MPL 和 EPL 的设计分析**。为验证 MPL 和 EPL 结构设计的合理性，我们进行了详细的消融实验，结果见表 VIII。具体而言，(a) “EPL&EPL”：从 MPL 中移除 SAM 预训练特征，使其结构与 EPL 保持一致。(b) “MPL&MPL”：将 SAM 预训练特征融入 EPL，使其结构与 MPL 保持一致。“EPL&EPL”的整体性能不如原始设计，而“MPL&MPL”的性能与原始设计相当，但引入了更多参数。我们的方法结构更简单，同时性能更优。

**令牌融合策略分析**。在每个解码器层之后，VNS 边缘令牌会与 VNS 掩码令牌融合，以显式聚合边缘表征。我们探索了多种融合策略，包括 (a) 交叉注意力、(b) 动态门控，以及 (c) 我们采用线性融合的逐元素相加。结果如表 IX 所示。令牌融合操作可有效提升交并比和边界交并比，幅度约为 1 个点。这三种策略的性能相近，但我们的方法所需参数更少，形式也更为简单。

**NSFM 模块消融**。在 NSFM 中，我们采用 Haar 小波分解 [72] 将特征分解为四个频带，并选取低频和高频分量进行进一步处理，以增强分割鲁棒性。为验证我们方法的合理性，我们进行了详细的消融研究，如表 X 所示。结果表明，移除小波分解（WD）会导致性能下降，而仅使用高频（HF）或仅使用低频（LF）分量都会导致性能欠佳。我们的 NSFM 模块在所有指标上均取得最佳结果，证明了同时利用高频和低频特征能够有效提升分割性能。

**与其他微调策略的比较**。在表 XI 中，我们将我们的方法与其他微调策略进行比较，包括微调 SAM 的编码器和解码器（*FT*-编码器和解码器）、解码器（*FT*-解码器），以及输出掩码标记（*FT*-令牌）。我们在 VNS-SEG 和 COCO 数据集上评估性能。可以观察到：**i)** 微调编码器和解码器，或仅微调解码器，能够更好地捕捉非显著特征，从而提升模型在 VNS-SEG 已见集上的性能。然而，这会导致灾难性遗忘，使模型在 COCO 上的性能大幅下降。**ii)**与微调整个解码器相比，微调输出掩码标记能够有效提升性能，但在视觉非显著场景中的表现仍然有限。**iii)** 我们的方法有效利用 SAM 的低级特征来促进 VNS 特征的学习，大幅提升性能，同时保持强大的零样本分割能力。

### *D.* 进一步分析与讨论

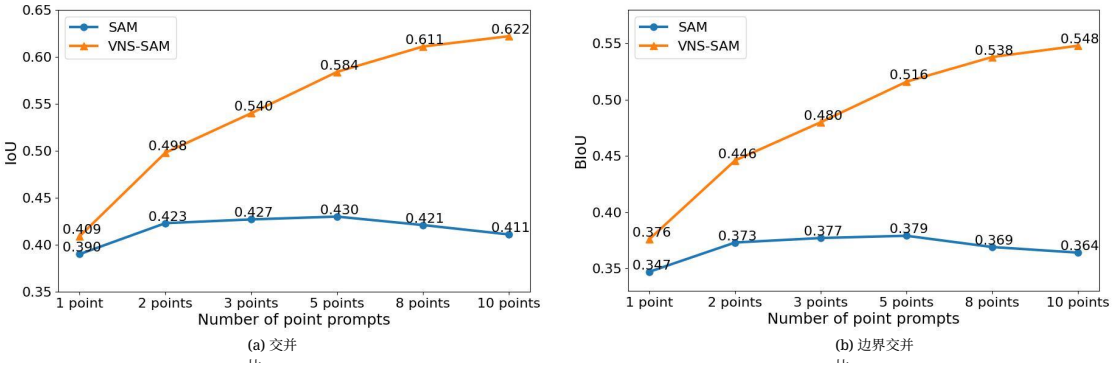
**在通用 COCO 数据集上的性能**。 为了进一步验证我们的方法在更多通用非显著场景中的效果，我们设计了视觉非显著性评分（VNS-score），用于量化图像的非显著性（更多详情见附录 A），并在 COCO 数据集中使用 VNS-score 提取非显著子数据集。我们计算每个

表XIII 不同 ViT 骨干网络下 SAM 与 VNS-SAM 的性能比较。VNS-SAM 在不同骨干网络和数据集上始终优于基线，同时仅增加少量额外参数。

| 骨干网络 | 方法 | *Eval-Seen-Set* | | | | 未见集评估 | | | | *COCO*全量集 | | | | 模型参数量（MB） | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | IoU | BIoU | $E_\phi$ | $F_\beta^w$ | IoU | BIoU | $E_\phi$ | $F_\beta^w$ | IoU | BIoU | $E_\phi$ | $F_\beta^w$ | 总计 | 可学习 |
| ViT-B | SAM [9] | .652 | .486 | .757 | .641 | .735 | .584 | .868 | .742 | .784 | .670 | .927 | .850 | 358 | 358 |
| | VNS-SAM | **.794** | **.643** | **.912** | **.817** | **.782** | **.652** | **.927** | **.822** | **.798** | **.689** | **.951** | **.866** | 367.4 | 9.4 |
| ViT-L | SAM [9] | .720 | .561 | .829 | .737 | .768 | .628 | .909 | .805 | .812 | .707 | .955 | **.881** | 1191 | 1191 |
| | VNS-SAM | **.821** | **.684** | **.929** | **.850** | **.800** | **.680** | **.944** | **.853** | **.816** | **.711** | **.956** | **.881** | 1200.8 | 9.8 |
| ViT-H | SAM [9] | .716 | .566 | .830 | .741 | .767 | .633 | .910 | .808 | .812 | .710 | **.956** | .878 | 2446 | 2446 |
| | VNS-SAM | **.833** | **.697** | **.936** | **.858** | **.800** | **.677** | **.934** | **.838** | **.814** | **.714** | **.956** | **.879** | 2456.2 | 10.2 |

表 XIV 我们统一的 VNS-SEG 数据集的效果。NC4K、ETIS 和 LIS 分别是用于伪装、息肉和低光照目标分割的未见数据集。使用 VNS-SEG 进行训练，在多项任务中始终取得出色结果，并优于各任务使用专用数据集训练所得的结果。

| 训练集 | VNS 特征 | NC4K | | | | ETIS | | | | LIS | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | IoU | BIoU | $E_\phi$ | $F_\beta^w$ | IoU | BIoU | $E_\phi$ | $F_\beta^w$ | AP | $AP_{50}$ | $AP_{75}$ |
| 基线 | - | .780 | .614 | .893 | .802 | .856 | .741 | .964 | .891 | .439 | .784 | .433 |
| COD10K+CAMO | 伪装 | .825 | .679 | .937 | .854 | .857 | .756 | .965 | .889 | .445 | .785 | .437 |
| Kvasir+ClinicDB | 息肉 | .730 | .530 | .890 | .768 | .850 | .748 | .958 | .885 | .399 | .708 | .392 |
| DIS-Dark+Thin-Dark+FSS-Dark | 低照度 | .787 | .634 | .920 | .828 | .867 | .769 | .972 | .906 | .459 | .786 | .455 |
| VNS-SEG | 统一 VNS 特征 | **.830** | **.685** | **.941** | **.860** | **.888** | **.790** | **.982** | **.918** | **.461** | **.787** | **.457** |




(a) 交并 (b) 边界交并

图 7. 在未见过的 CDS2K 子集上，使用不同数量输入点进行交互式分割的性能比较。VNS-SAM 在各种点数下始终优于 SAM，展现出更显著的提升。

对象-图像对的评分，并以 0.7 为阈值提取非显著子集（COCO-NS），其余样本则归为显著子集（COCO-S）。表 XII展示了这些精心筛选的子集上的基准测试结果，凸显了我们方法的有效性。在具有挑战性的 COCO-NS 子集上，VNS-SAM 的交并比达到 76.5%，边界交并比达到 60.4%，分别比 SAM 高出 3.3% 和 3.2%。值得注意的是，VNS-SAM 在 COCO-S 子集上也保持了有竞争力的性能（交并比为 77.8%，而 SAM 为 76.3%），这证实了其在非显著场景中的改进并未以牺牲整体分割性能为代价。

**不同骨干网络之间的比较**。在表XIII中，我们对 SAM 和 VNS-SAM 在多种 ViT [87] 骨干网络上的表现进行了全面比较，包括 ViT-Base（ViT-B）、ViT-Large（ViT-L）和 ViT-Huge（ViT-H）。我们全面评估了这些模型在 VNS-SEG 和 COCO 数据集的已见集和未见集上的表现，并报告了模型在已见集、未见集以及 COCO 数据集上的性能。此外，我们还列出了模型的总参数量和可学习参数量。这些结果表明，在不同规模的骨干网络和不同数据集上，VNS-SAM 的表现始终显著优于 SAM。就模型规模而言，基于 ViT-B、ViT-L 和 ViT-H 的 VNS-SAM 参数量分别仅增加 2.5%、0.8% 和 0.4%。

**VNS-SEG 的效果**。在表XIV中，我们比较了单任务数据训练与统一数据训练的结果，清楚地展示了 VNS-SEG 数据集的优势。对于伪装目标、息肉和低光照目标分割任务，我们分别使用常用的训练集 COD10K+CAMO、Kvasir+ClinicDB 和 DIS-Dark+Thin-Dark+FSS-Dark 来训练模型。我们使用三个未见数据集进行零样本评估，即 NC4K、ETIS 和 LIS。值得注意的是，使用 VNS-SEG 训练的模型在多个任务上始终取得出色的结果，并且优于在各任务专用数据集上训练的结果。这表明，统一的 VNS-SEG 数据集能够使模型学习到更稳健的非显著特征，优于以往用于训练的单任务数据集。

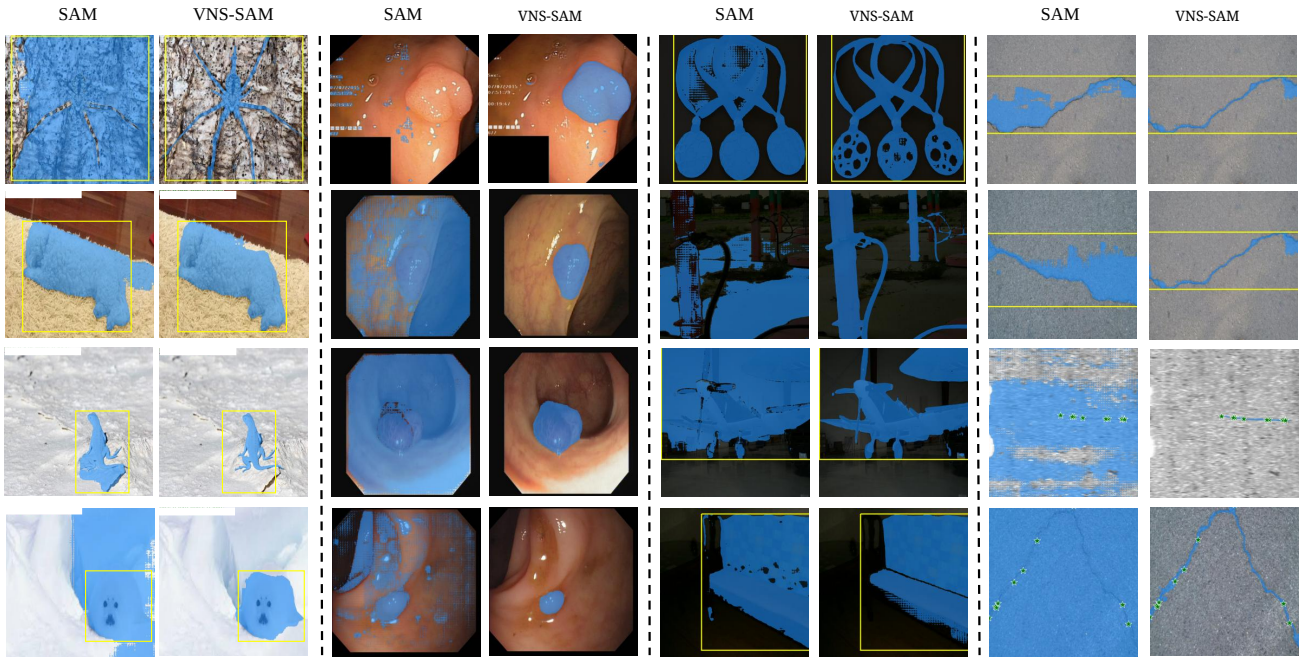



图 8. SAM 与 VNS-SAM 的分割结果视觉对比。面对具有挑战性的 VNS 场景时，SAM 无法准确区分前景和背景，导致分割错误。相比之下，我们的 VNS-SAM 在这些场景中更加稳健。

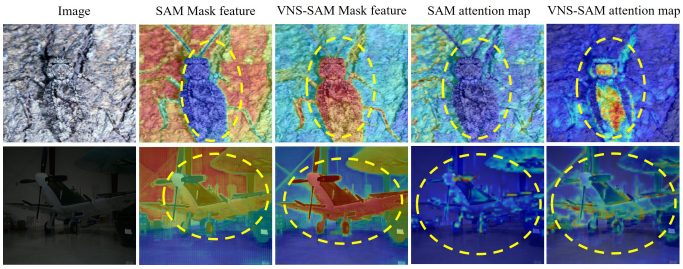



图 9. SAM 与 VNS-SAM 的掩码特征（预测层中与输出词元相乘的特征图）和注意力图（输出词元在最终解码器层中的交叉注意力图）的视觉比较。VNS-SAM 能够准确激活目标区域和边界，而 SAM 由于缺乏识别非显著特征的能力，无法区分前景与背景。

表 XV SAM、HQ-SAM 和 VNS-SAM 的计算需求。

| 方法 | 参数 | | 浮点运算次数 | 推理 |
|---|---|---|---|---|
| | 可学习 | 总计 | | FPS |
| SAM | 1191 M | 1191 M | ≈1550 G | 7.3 |
| HQ-SAM | 5.1 M | 1196.1 M | ≈1553 G | 7.1 |
| VNS-SAM | 9.8 M | 1200.8 M | ≈1559 G | 7.0 |

**基于点的交互式分割比较**。图 7 展示了 VNS-SAM 和 SAM 使用点提示时的交互式分割性能。该比较在未见子集 CDS2K 上，针对不同数量的输入点评估 VNS-SAM 和 SAM。对于不同数量的点提示（从 1 个点到 10 个点），VNS-SAM 的表现始终优于 SAM。请注意，随着提示所含的歧义减少（输入点增多），相对性能提升会更加显著。这表明 VNS-SAM 强大的分割能力。

**计算需求对比**。如表 XV 所示，与 HQ-SAM 相比，VNS-SAM 的参数总量仅略有增加（1200.8M 对 1196.1M），浮点运算次数也仅略有增加（1559G 对 1553G）。尽管存在这点额外开销，VNS-SAM 的分割性能却显著更高，展现出更优的性能与效率权衡。其推理速度也与之相当（7.0 帧每秒对 7.1 帧每秒），证实我们的方法既有效，也适用于实际部署。

### *E.* 可视化

本节展示一些可视化结果，并将我们的方法与 SAM 进行定性比较。

**分割结果可视化**。图 8 展示了 SAM 和我们的 VNS-SAM 在 VNS-SEG 评估集上的分割结果可视化。我们可以观察到，由于 VNS 特征具有挑战性，SAM 难以准确分割这些目标，导致细节严重缺失且背景预测错误，体现出其局限性。相比之下，我们的 VNS-SAM 能够精准分割 VNS 场景中不显眼的目标，展现出其对各种 VNS 特征的稳健感知能力。

**特征可视化**。在图 9 中，我们对比展示了 SAM 与 VNS-SAM 最后一层解码器的掩码特征图（第二列和第三列）以及交叉注意力图（第四列和第五列）。掩码特征来自解码器的最终掩码预测层，而交叉注意力图则来自最后一个令牌到图像层，分别对应 SAM 的输出掩码令牌和我们的 VNS 掩码令牌。可以观察到，VNS-SAM 能够准确激活目标区域和边界，而 SAM 则由于

缺乏识别非显著特征的能力，在前景与背景之间表现出混淆。这表明 VNS-SAM 增强了区分细微判别区域和细节的能力，这对于在非显著条件下实现有效分割至关重要。

## VI. 结论

本文探讨了 SAM 在前景与背景对比度较低的场景中性能下降的问题，我们将这类场景称为视觉非显著场景。为解决这一问题，我们提出了 VNS-SAM，在保留 SAM 原始零样本泛化能力的同时，增强其对 VNS 场景的感知能力。我们通过两种高效且有效的设计，充分利用 SAM 的低级特征来实现这一目标：掩码-边缘令牌交互式解码器和非显著特征挖掘模块。从数据角度出发，我们构建了统一的 VNS-SEG 数据集，涵盖多种 VNS 场景，与以往只针对单一场景的数据集形成对比。VNS-SEG 用于使模型学习稳健的非显著特征，并全面评估模型在 VNS 场景中的性能。大量实验表明，VNS-SAM 性能优越，展现出广泛应用于现实世界的潜力。此外，VNS-SEG 已见集和未见集上的表现为 VNS 分割树立了新标准。关于未来研究，我们希望构建的 VNS-SEG 数据集能够激励研究者开发更强大的、适用于 VNS 场景的分割模型。

## 附录

### A. 视觉非显著性评分

为了进一步分析，我们设计了一种**视觉非显著性评分**（VNS-score），用于量化图像的非显著性。该评分从两个方面进行计算：前景与背景之间的对比度，以及物体边界的清晰度。具体而言，前景-背景对比度 $C_{fb}$ 的计算综合考虑了两个关键因素：颜色对比度和纹理对比度。颜色对比度反映前景与背景之间的颜色差异，而纹理对比度反映两者纹理特征的差异。颜色对比度通过计算 LAB 色彩空间 [88]中前景区域和背景区域的颜色均值向量$\mu_{fg}^{LAB}$与 $\mu_{bg}^{LAB}$ 之间的差异来衡量。纹理对比度则基于灰度共生矩阵（GLCM） [89]进行计算。分别得到前景区域的对比度 $\mathrm{C}_{fg}^{GLCM}$和背景区域的对比度 $\mathrm{C}_{bg}^{GLCM}$。

$$C_{fb} = \frac{1}{2}(\|\mu_{fg}^{LAB} - \mu_{bg}^{LAB}\| + \|\mathrm{C}_{fg}^{GLCM} - \mathrm{C}_{bg}^{GLCM}\|). \quad (6)$$

边界清晰度 $B$ 用于衡量图像中物体边界的清晰程度。其定义如下：

$$B = \frac{\mathrm{Mean}(\|\nabla I_{\mathrm{edge}}\|)}{255}, \quad (7)$$

其中， $\|\nabla I_{\mathrm{edge}}\|$ 表示使用Sobel算子计算出的梯度幅值，具体是在物体边界区域内计算的。 $B$ 的值接近 0 表明物体边界非常模糊。

最后，VNS-score通过对 $C_{fb}$ 和 $B$进行加权求和得到，如下：

$$\text{VNS-score} = 1 - \frac{1}{2}(C_{fb} + B).1 \quad (8)$$

表 XVI 原始数据集、合成低光照数据集和真实低光照数据集的均值和标准差。

| 数据集 | 原始数据 | | 合成低光 | | 真实低光 | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD |
| 均值/标准差 | 131.82 | 58.72 | 11.13 | 7.81 | 8.31 | 10.3 |

### B. 合成数据集的真实性

我们计算了原始数据集（DIS、Thin 和 FSS）、由 CycleGAN 生成的相应合成数据集（DIS-Dark、Thin-Dark 和 FSS-Dark）以及真实低光数据集 LIS [63]中图像的均值和方差。结果如表 XVI 所示。与原始数据集相比，合成图像的均值和标准差（SD）与真实低光 LIS 数据集的结果更加接近（11.13 对比 8.31，以及 7.81 对比 10.3）。这表明，即使未达到绝对的照片真实感，CycleGAN 生成的数据仍能有效捕捉非显著场景的统计特性。

## 参考文献

[1] A. Kirillov、K. He、R. Girshick、C. Rother 和 P. Doll´ar，《全景分割》，载于*IEEE*计算机视觉与模式识别会议论文集，2019年，第9404–9413页。[2] D. Bolya、C. Zhou、F. Xiao 和 Y. J. Lee，《Yolact：实时实例分割》，载于*IEEE*国际计算机视觉会议论文集，2019年，第9157–9166页。[3] Y. Wang、Z. Xu、X. Wang、C. Shen、B. Cheng、H. Shen 和 H. Xia，《基于Transformer的端到端视频实例分割》，载于*IEEE*计算机视觉与模式识别会议论文集，2021年，第8741–8750页。[4] Y. Pang、X. Zhao、L. Zhang 和 H. Lu，《Caver：用于双模态显著目标检测的跨模态混合视图Transformer》，*IEEE*《图像处理汇刊》，第32卷，第892–904页，2023年。[5] G. Guo、P. Chen、X. Yu、Z. Han、Q. Ye 和 S. Gao，《拯救微小目标，拯救所有目标：用于微小目标检测的分层激活网络》，*IEEE*《视频技术电路与系统汇刊》，第34卷，第1期，第221–234页，2023年。[6] C. Wang、G. Guo、C. Liu、D. Shao 和 S. Gao，《有效旋转：学习用于航空目标检测的旋转鲁棒原型》，*IEEE*《地球科学与遥感汇刊》，第62卷，第1–14页，2024年。[7] S. Gao、G. Guo、H. Huang 和 C. P. Chen，《深入还是广泛？利用混合网络架构进行弱监督目标分类与定位》，*IEEE*神经网络与学习系统汇刊，2023年。[8] K. Wang、D. Lin、C. Li、Z. Tu 和 B. Luo，《无对齐RGB-T显著目标检测：语义引导的非对称相关网络与统一基准》，*IEEE*《多媒体汇刊》，第26卷，第10 692–10 707页，2024年。[9] A. Kirillov、E. Mintun、N. Ravi、H. Mao、C. Rolland、L. Gustafson、T. Xiao、S. Whitehead、A. C. Berg、W.-Y. Lo 等，《Segment Anything》，载于*2023*年*IEEE/CVF*国际计算机视觉会议（*ICCV*），2023年，第3992–4003页。[10] N. Ravi、V. Gabeur、Y.-T. Hu、R. Hu、C. Ryali、T. Ma、H. Khedr、R. R¨adle、C. Rolland、L. Gustafson、E. Mintun、J. Pan、K. V. Alwala、N. Carion、C.-Y. Wu、R. Girshick、P. Doll´ar 和 C. Feichtenhofer，《SAM 2：在图像和视频中分割万物》，*arXiv*预印本 *arXiv:2408.00714*，2024年。 [在线]。可访问：https://arxiv.org/abs/2408. 00714

[11] N. Carion、L. Gustafson、Y.-T. Hu、S. Debnath、R. Hu、D. Suris、C. Ryali、K. V. Alwala、H. Khedr、A. Huang等，《Sam 3：基于概念的万物分割》，arXiv预印本 arXiv:2511.16719，2025年。[12] K. Chen、C. Liu、H. Chen、H. Zhang、W. Li、Z. Zou 和 Z. Shi，《Rsprompter：基于视觉基础模型学习遥感实例分割提示》，IEEE《地球科学与遥感汇刊》，2024年。[13] L. P. Osco、Q. Wu、E. L. de Lemos、W. N. Gonc.alves、A. P. M. Ramos、J. Li 和 J. M. Junior，《面向遥感应用的万物分割模型（SAM）：从零样本到单样本》，《国际应用地球观测与地理信息期刊》，第124卷，103540，2023年，第.页。[14] L. Ding、K. Zhu、D. Peng、H. Tang、K. Yang 和 L. Bruzzone，《将Segment Anything模型适配于甚高分辨率遥感图像变化检测》，IEEE《地球科学与遥感汇刊》，2024年。[15] D. Wang、J. Zhang、B. Du、M. Xu、L. Liu、D. Tao 和 L. Zhang，《Samrs：利用Segment Anything模型扩展遥感分割数据集》，载于第三十七届神经信息处理系统大会数据集与基准赛道，2023年。[16] S. Ren、F. Luzi、S. Lahrichi、K. Kassaw、L. M. Collins、K. Bradbury 和 J. M. Malof，《从太空进行万物分割？》，载于IEEE冬季计算机视觉应用会议论文集，2024年，第8355–8365页。[17] G. Guo、D. Shao、C. Zhu、S. Meng、X. Wang 和 S. Gao，《P2p：将点监督转化为用于目标检测与分割的显式视觉提示》，载于第三十三届国际人工智能联合会议论文集，2024年。[18] J. Ma、Y. He、F. Li、L. Han、C. You 和 B. Wang，《在医学图像中分割万物》，《自然·通讯》，第15卷，第1期，第654页，2024年。[19] M. A. Mazurowski、H. Dong、H. Gu、J. Yang、N. Konz 和 Y. Zhang，《用于医学图像分析的Segment Anything模型：一项实验研究》，《医学图像分析》，第89卷，第102918页，2023年。[20] G.-P. Ji、D.-P. Fan、P. Xu、M.-M. Cheng、B. Zhou 和 L. Van Gool，《SAM在隐蔽场景中表现不佳——关于“万物分割”的实证研究》，arXiv预印本 arXiv:2304.06022，2023年。[21] L. Tang、H. Xiao 和 B. Li，《SAM能分割万物吗？当SAM遇上伪装目标检测》，arXiv预印本 arXiv:2304 04709，2023年。[22] T. Zhou、Y. Zhang、Y. Zhou、Y. Wu 和 C. Gong，《SAM能分割息肉吗？》，arXiv预印本 arXiv:2304.07583，2023年。[23] Y. Li、M. Hu 和 X. Yang，《Polyp-sam：迁移SAM用于息肉分割》，载于2024医学成像：计算机辅助诊断，第12927卷。SPIE，2024年，第759–765页。[24] J. Cheng、J. Ye、Z. Deng、J. Chen、T. Li、H. Wang、Y. Su、Z. Huang、J. Chen、L. Jiang等，《Sam-med2 16184, 2023 2308 d》，arXiv预印本 arXiv: . .[25] J. Ma、Y. He、F. Li、L. Han、C. You 和 B. Wang，《在医学图像中分割万物》，《自然·通讯》，第15卷，第1期，第654页，2024年。[26] T. Chen、L. Zhu、C. Ding、R. Cao、S. Zhang、Y. Wang、Z. Li、L. Sun、P. Mao 和 Y. Zang，《SAM无法分割万物？——SAM适配器：将SAM适配到表现不佳的场景：伪装、阴影等》，arXiv预印本 arXiv:2304.09148，2023年。[27] J. Wu、W. Ji、Y. Liu、H. Fu、M. Xu、Y. Xu 和 Y. Jin，《医学SAM适配器：将Segment Anything模型适配于医学图像分割》，arXiv预印本 arXiv:2304.12620，2023年。[28] H. Ding、X. Jiang、A. Q. Liu、N. M. Thalmann 和 G. Wang，《用于场景分割的边界感知特征传播》，载于IEEE国际计算机视觉会议论文集，2019年，第6819–6829页。[29] J.-X. Zhao、J.-J. Liu、D.-P. Fan、Y. Cao、J. Yang 和 M.-M. Cheng，《Egnet：用于显著目标检测的边缘引导网络》，载于IEEE国际计算机视觉会议论文集，2019年，第8779–8788页。[30] Y. Sun、S. Wang、C. Chen 和 T.-Z. Xiang，《边界引导的伪装目标检测》，arXiv预印本 arXiv:2207.00794，2022年。[31] T. Zhou、Y. Zhang、G. Chen、Y. Zhou、Y. Wu 和 D.-P. Fan，《用于息肉分割的边缘感知特征聚合网络》，arXiv预印本 arXiv:2309.10523，2023年。[32] Y. Ye、R. Yi、Z. Gao、Z. Cai 和 K. Xu，《深入探究清晰度：用于清晰边缘检测的引导式标签细化》，IEEE《图像处理汇刊》，2023年。[33] T. Liu、H. Zhu、Y. Wei、S. Wei、Y. Zhao 和 Y. Zhang，《通过边缘引导扩散实现精准人体解析》，IEEE《图像处理汇刊》，2024年。[34] K. Wang、D. Lin、C. Li、Z. Tu 和 B. Luo，《无对齐RGB-T显著目标检测：语义引导的非对称相关网络与统一基准》，IEEE《多媒体汇刊》，2024年。

[35] K. Wang, Z. Tu, C. Li, C. Zhang, 和 B. Luo，《面向多模态显著目标检测的自适应融合库学习》，IEEE《视频技术电路与系统汇刊》，2024年。[36] L. Ke, M. Ye, M. Danelljan, Y.-W. Tai, C.-K. Tang, F. Yu 等，《高质量分割一切》，神经信息处理系统进展，第36卷，页码29 914–29 934，2023年。[37] K. Wang, D. Lin, C. Li, Z. Tu, 和 B. Luo，《通过语义特征融合引导将Segment Anything模型应用于多模态显著目标检测》，arXiv预印本 arXiv:2408.15063，2024年。[38] T. Chen, Z. Mai, R. Li, 和 W.-l. Chao，《Segment Anything模型（SAM）增强的弱监督语义分割伪标签》，arXiv预印本 arXiv:2305.05803，2023年。[39] X. Yang 和 X. Gong，《基础模型辅助的弱监督语义分割》，载于《IEEE冬季计算机视觉应用会议论文集》，2024年，页码523–532。[40] Z. Wei, P. Chen, X. Yu, G. Li, J. Jiao, 和 Z. Han，《用于点提示实例分割的语义感知SAM》，载于《IEEE计算机视觉与模式识别会议论文集》，2024年，页码3585–3594。[41] H. Chen, P. Wei, G. Guo, 和 S. Gao，《SAM-COD：SAM引导的弱监督伪装目标检测统一框架》，载于《欧洲计算机视觉会议》，2024年，页码315–331。[42] G. Guo, Y. Guo, X. Yu, W. Li, Y. Wang, 和 S. Gao，《通过生成式潜空间增强实现任意质量图像分割》，载于《计算机视觉与模式识别会议论文集》，2025年，页码2366–2376。[43] W. Ji, J. Li, Q. Bi, T. Liu, W. Li, 和 L. Cheng，《Segment Anything并非总是完美：对SAM在不同现实世界应用中的研究》，《机器智能研究》。[44] W.-T. Chen, Y.-J. Vong, S.-Y. Kuo, S. Ma, 和 J. Wang，《RobustSAM：在退化图像上稳健地分割一切》，载于《IEEE计算机视觉与模式识别会议论文集》，2024年，页码4081–4091。[45] B. Li, H. Xiao, 和 L. Tang，《ASAM：通过对抗调优增强Segment Anything模型》，载于《IEEE计算机视觉与模式识别会议论文集》，2024年，页码3699–3710。[46] C. Zhang, D. Han, Y. Qiao, J. U. Kim, S.-H. Bae, S. Lee, 和 C. S. Hong，《更快的Segment Anything：面向移动应用的轻量级SAM》，arXiv预印本 arXiv:2306.14289，2023年。[47] Y. Xiong, B. Varadarajan, L. Wu, X. Xiang, F. Xiao, C. Zhu, X. Dai, D. Wang, F. Sun, F. Iandola 等，《EfficientSAM：利用掩码图像预训练实现高效的Segment Anything》，arXiv预印本 arXiv:2312.00863，2023年。[48] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, 和 L. Shao，《伪装目标检测》，载于《IEEE计算机视觉与模式识别会议论文集》，2020年，页码2777–2787。[49] T.-N. Le, T. V. Nguyen, Z. Nie, M.-T. Tran, 和 A. Sugimoto，《用于伪装目标分割的分支网络》，《计算机视觉与图像理解》，第184卷，页码45–56，2019年。[50] Y. Lv, J. Zhang, Y. Dai, A. Li, B. Liu, N. Barnes, 和 D.-P. Fan，《同时定位、分割并排序伪装物体》，载于《IEEE计算机视觉与模式识别会议论文集》，2021年，页码11 591–11 601。[51] D.-P. Fan, G.-P. Ji, P. Xu, M.-M. Cheng, C. Sakaridis, 和 L. Van Gool，《深度隐蔽场景理解进展》，《视觉智能》，第1卷，第1期，第16页，2023年。[52] Z. Huang, H. Dai, T.-Z. Xiang, S. Wang, H.-X. Chen, J. Qin, 和 H. Xiong，《基于Transformer的伪装目标检测特征收缩金字塔》，载于《IEEE计算机视觉与模式识别会议论文集》，2023年，页码5557–5566。[53] D.-P. Fan, G.-P. Ji, M.-M. Cheng, 和 L. Shao，《隐蔽目标检测》，IEEE《模式分析与机器智能汇刊》，第44卷，第10期，页码6024–6042，2021年。[54] C. He, K. Li, Y. Zhang, L. Tang, Y. Zhang, Z. Guo, 和 X. Li，《通过特征分解与边缘重建进行伪装目标检测》，载于《IEEE计算机视觉与模式识别会议论文集》，2023年，页码22 046–22 055。[55] Y. Zhong, B. Li, L. Tang, S. Kuang, S. Wu, 和 S. Ding，《在频域中检测伪装目标》，载于《IEEE计算机视觉与模式识别会议论文集》，2022年，页码4504–4513。[56] J. Zhao, X. Li, F. Yang, Q. Zhai, A. Luo, Z. Jiao, 和 H. Cheng，《Focus-diffuser：感知局部差异以检测伪装目标》，载于《欧洲计算机视觉会议》。施普林格，2024年，页码181–198。

[57] H. Chen, D. Shao, G. Guo, 和 S. Gao, 《只需一点提示：点监督伪装目标检测》，载于《欧洲计算机视觉会议》，2024年，页码332–348。[58] J. Silva, A. Histace, O. Romain, X. Dray, 和 B. Granado, 《面向早期结直肠癌诊断的无线胶囊内镜图像息肉嵌入式检测》，《国际计算机辅助放射学与外科学杂志》，第9卷，页码283–293，2014年。[59] J. Bernal, F. J. S´anchez, G. Fern´andez-Esparrach, D. Gil, C. Rodr´ıguez, 和 F. Vilari˜no, 《用于结肠镜检查中精准突出显示息肉的WM-DOVA映射：与医生生成的显著图的验证比较》，《计算机化医学影像与图形》，第43卷，页码99–111，2015年。[60] N. Tajbakhsh, S. R. Gurudu, 和 J. Liang, 《利用形状和上下文信息自动检测结肠镜视频中的息肉》，《IEEE医学影像汇刊》，第35卷，第2期，页码630–644，2015年。[61] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, 和 H. D. Johansen, 《Kvasir-SEG：一个息肉分割数据集》，载于《多媒体建模：第26届国际会议，MMM 2020，韩国大田，2020年1月5日至8日，论文集，第二部分26》。施普林格，2020年，页码451–462。[62] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, 和 L. Shao, 《PraNet：用于息肉分割的并行反向注意力网络》，载于《国际医学图像计算与计算机辅助手术会议》。施普林格，2020年，页码263–273。[63] L. Chen, Y. Fu, K. Wei, D. Zheng, 和 F. Heide, 《暗光环境下的实例分割》，《国际计算机视觉杂志》，页码1–21，2023年。[64] K. A. Hashmi, G. Kallempudi, D. Stricker, 和 M. Z. Afzal, 《FeatEnhancer：在低照度视觉条件下增强目标检测及其他任务的层次特征》，载于《IEEE国际计算机视觉会议论文集》，2023年，页码6725–6735。[65] C. Wei, W. Wang, W. Yang, 和 J. Liu, 《用于低光照增强的深度Retinex分解》，arXiv预印本arXiv:1808.04560，2018年。[66] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, 和 M. Jagersand, 《BASNet：边界感知显著目标检测》，载于《IEEE计算机视觉与模式识别会议论文集》，2019年，页码7479–7489。[67] R. Wang, S. Chen, C. Ji, J. Fan, 和 Y. Li, 《用于医学图像分割的边界感知上下文神经网络》，《医学图像分析》，第78卷，第102395页，2022年。[68] T. Liu, J. Zhang, X. Nie, Y. Wei, S. Wei, Y. Zhao, 和 J. Feng, 《用于高保真服装迁移的空间感知纹理变换器》，IEEE《图像处理汇刊》，第30卷，页码7499–7510，2021年。[69] T. Liu, Y. Wei, Y. Zhao, S. Liu, 和 S. Wei, 《Magic-Wall：通过增强墙面分割实现室内装饰可视化》，IEEE《图像处理汇刊》，第28卷，第9期，页码4219–4232，2019年。[70] C. Zhu, S. Gao, H. Chen, G. Guo, C. Wang, Y. Wang, C. S. Lei, 和 Q. Fan, 《Mamba为何有效？探索用于多模态图像融合的线性Transformer-Mamba网络》，arXiv预印本arXiv:2409.03223，2024年。[71] M. Stevens 和 S. Merilaita, 《动物伪装：当前问题与新视角》，《英国皇家学会哲学汇刊B辑：生物科学》，第364卷，第1516期，页码423–427，2009年。[72] R. S. Stankovic 和 B. J. Falkowski, 《Haar小波变换：现状与成就》，《计算机与电气工程》，第29卷，第1期，页码25–44，2003年。[73] Y. Yang, G. Yuan, 和 J. Li, 《 SFFNet：一种用于遥感分割的基于小波的空域与频域融合网络》，IEEE《地球科学与遥感汇刊》，第62卷，页码1–17，2024年。[74] G. Yue, Y. Li, S. Wu, B. Jiang, T. Zhou, W. Yan, H. Lin, 和 T. Wang, 《用于自动结直肠息肉分割的双域特征交互网络》，《IEEE仪器与测量汇刊》，第73卷，页码1–12，2024年。[75] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, 和 Q. Hu, 《ECA-Net：用于深度卷积神经网络的高效通道注意力》，载于《IEEE计算机视觉与模式识别会议论文集》，2020年，页码11 534–11 542。[76] X. Qin, Z. Wang, Y. Bai, X. Xie, 和 H. Jia, 《 FFA-Net：用于单幅图像去雾的特征融合注意力网络》，载于《AAAI人工智能会议论文集》，第34卷，第07期，2020年，页码11 908–11 915。[77] X. Qin, H. Dai, X. Hu, D.-P. Fan, L. Shao, 和 L. Van Gool, 《高精度二分图像分割》，载于《欧洲计算机视觉会议》。施普林格，2022年，页码38–56。[78] J. H. Liew, S. Cohen, B. Price, L. Mai, 和 J. Feng, 《深度交互式细长目标选择》，载于《IEEE冬季计算机视觉应用会议论文集》，2021年，页码305–314。

[79] X. Li、T. Wei、Y. P. Chen、Y.-W. Tai 和 C.-K. Tang, 《Fss-1000：用于少样本分割的1000 类数据集》，载于*IEEE*计算机视觉与模式识别会议论文集，2020年，第2869–2878页。[80] J. Wei、S. Wang 和 Q. Huang, 《$F^3$net：用于显著目标检测的融合、反馈与聚焦》，载于*AAAI*人工智能会议论文集，第34卷第07期，2020年，第12 321–12 328页。[81] E. Xie、W. Wang、W. Wang、M. Ding、C. Shen 和 P. Luo, 《野外透明物体分割》，载于计算机视觉——*ECCV 2020*：第*16*届欧洲会议，格拉斯哥，英国，*2020*年*8*月*23*日至*28*日，论文集，第*XIII*部分 *16*。施普林格，2020年，第696–711页。[82] J.-Y. Zhu、T. Park、P. Isola 和 A. A. Efros, 《使用循环一致性对抗网络进行非配对图像到图像转换》，载于*IEEE*国际计算机视觉会议论文集，2017年，第2223–2232页。[83] B. Cheng、R. Girshick、P. Doll´ar、A. C. Berg 和 A. Kirillov, 《边界交并比：改进以目标为中心的图像分割评估》，载于*IEEE*计算机视觉与模式识别会议论文集，2021年，第15 334–15 342页。[84] D.-P. Fan、C. Gong、Y. Cao、B. Ren、M.-M. Cheng 和 A. Borji, 《用于二值前景图评估的增强对齐度量》，*arXiv*预印本*arXiv:1805.10421*，2018 年。[85] R. Margolin 、 L. Zelnik-Manor 和 A. Tal, 《如何评估前景图？》，载于*IEEE*计算机视觉与模式识别会议论文集，2014年，第248–255页。[86] T.-Y. Lin、M. Maire、S. Belongie、J. Hays、P. Perona、D. Ramanan、P. Doll´ar 和 C. L. Zitnick, 《Microsoft coco：上下文中的常见物体》，载于计算机视觉——*ECCV 2014*：第*13*届欧洲会议，苏黎世，瑞士，*2014*年*9*月*6*日至*12*日，论文集，第*V*部分 *13*。施普林格，2014年，第740–755页。[87] A. Dosovitskiy、L. Beyer、A. Kolesnikov、D. Weissenborn、X. Zhai、T. Unterthiner、M. Dehghani、M. Minderer、G. Heigold、S. Gelly 等, 《一张图像胜过16×16个词：用于大规模图像识别的 Transformer》，载于国际学习表征会议，2020年。[88] S. Murali 和 V. Govindan, 《使用 LAB 色彩空间从单幅图像中检测并去除阴影》，控制论与信息技术，第13卷第1期，第95–103页，2013年。[89] R. M. Haralick、K. Shanmugam 和 I. H. Dinstein, 《用于图像分类的纹理特征》，*IEEE*《系统、人类与控制论汇刊》，第6期，第610–621页，1973年。