LawDIS: 基于语言-窗口的可控图像二值分割*

闫馨宇 [1,2,6] 孙美君 [1,2] 季葛鹏 [3]
Fahad Shahbaz Khan[6] Salman Khan[6] 范登平 [4,5†]
[1] 天津大学 [2] 天津市机器学习重点实验室 [3] 澳大利亚国立大学
[4] 南开国际先进研究院 (深圳福田) [5] 南开大学 [6] MBZUAI

摘要

本文提出了 *LawDIS*, 一个基于语言-窗口的可控图像二值分割 *(DIS)* 框架, 能够生成高质量的目标掩码。本文的框架将 *DIS* 任务重构为在潜在扩散模型中进行的以图像为条件的掩码生成任务, 从而实现了用户控制的无缝集成。*LawDIS* 具有从宏观到微观的控制模式。具体而言, 在宏观模式中, 引入语言控制的分割策略 *(LS)*, 根据用户提供的语言提示生成初始掩码。在微观模式中, 提出窗口控制的细化策略 *(WR)*, 允许在初始掩码中对用户定义的区域 *(*即尺寸可调的窗口*)* 进行灵活细化。通过一个模式切换器进行协调, 这些模式可以独立运行或联合运行, 使该框架非常适合高精度、个性化的应用。在 *DIS5K* 基准数据集上的大量实验证明, *LawDIS* 在所有指标上显著超过了 *11* 种最先进的方法。值得注意的是, 相较于次优模型 *MVANet*, 在 *DIS-TE* 上, *LawDIS* 在同时使用 *LS* 和 *WR* 策略时获得了 *4.6%* 的 F_β^ω 值的提升, 仅使用 *LS* 策略时获得了 *3.6%* 的提升。代码将在以下地址发布: https://github.com/XinyuYanTJU/LawDIS。

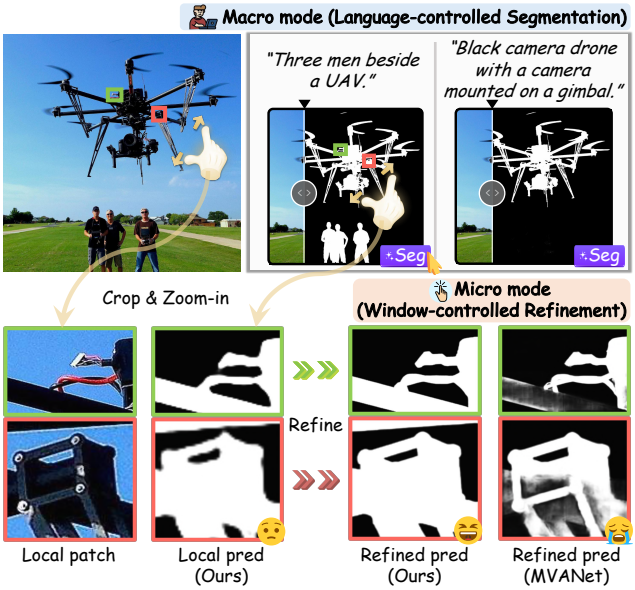

图 1. 为 DIS 任务提出的宏观与微观控制的示意图。宏观模式允许用户通过自定义的语言提示实现目标分割, 而微观模式支持在任意尺度的用户定义窗口上进行后期细化。经过细化后, 本文的方法能够得到更加精确的分割结果, 而次优模型 MVANet [45] 缺乏对裁剪局部图像块的适应能力, 导致更差的预测结果。

1. 引言

随着高性能相机设备的普及, 计算机视觉中的分割任务已从粗略定位 [36, 22] 发展为高精度刻画 [46]。因此, 图像二值分割 (DIS) 任务 [28], 专注于前景目标的高精度分割, 因其广泛的应用场景而受到了广泛关注, 例如: 三维重建 [21, 37], 图像编辑 [10, 20], 增强现实 [29, 43], 以及医学图像分割 [14, 13]。

相比于通用的分割任务 [18, 19, 35], DIS 任务具有更大的挑战, 主要体现在 2 个方面。首先, 广泛认可的 DIS 基准数据集 [28]覆盖了 200 多个目标类别和多样的视觉特征, 如显著目标 [9, 50]和伪装目标

*本文为ICCV'25 论文的中文翻译稿。由闫馨宇翻译, 并由范登平和季葛鹏校稿。

†通讯作者: 范登平 (dengpfan@gmail.com)

[6, 8, 12]。这种多样性要求模型具备较强的全局上下文理解能力, 以便在不同场景中正确前景目标。此外, 该任务强调在高分辨率图像中对目标的高精度细分, 甚至包括对目标内部细节的刻画。因此, 模型应具备细粒度特征提取能力, 以有效分割复杂结构和形状。

当前的 DIS 方法 [28, 49, 25, 16, 48, 45] 通常采用判别式学习范式 (即基于每像素的分类), 依赖模型自身的学习能力来提取目标语义。这种范式在实际应用中面临困难, 尤其是面临个性化需求时表现不佳。首先, 当图像中包含多个前景实体时, 在语义上会产生歧义, 难以明确应分割哪一个或哪一些目标。其次, 为了更好地捕捉高分辨率目标的几何细节, 大多数方法 [25, 16, 41] 引入了额外的高分辨率信息流, 通常通过将整张图像下采样为 1024^2 像素的输入。这种直接的方式在一定程度上弥补了被分割目标的细粒度信息, 但无限制地扩大输入尺寸在计算上是不可行的。近期的方法 [45, 48] 通过将整张图像划分为一系列小块来应对该问题, 从而以较少的像素损失实现几何细节的等效放大。然而, 由于这些方法是在预设尺寸的小块上进行训练的, 它们缺乏对不同大小图像块的适应能力。如图 1 所示, 先进模型 MVANet [45] 在输入与训练阶段尺寸不同的局部图像块时会失败。

为缓解上述问题, 本文提出了一个基于语言-窗口的可控图像二值分割 (DIS) 框架 LawDIS, 旨在满足用户的个性化需求。本文的框架将 DIS 任务重构为在潜在扩散模型中 [33] 进行的以图像为条件的掩码生成任务, 从而实现多种用户控制方式的无缝集成。此外, 本文为该生成式框架引入了从宏观到微观的控制模式。在宏观模式中, 本文提出语言控制的分割策略 (LS), 根据用户提供的语言提示生成初始掩码, 如图 1 上半部分所示。在微观模式中, 窗口控制的细化策略 (WR) 根据初始掩码对用户指定的不同尺度的区域进行细化, 如图 1 下半部分所示。通过模式切换器, 这两种模式可独立或联合运行, 从而在训练阶段实现自适应优化, 在推理阶段实现相互细化。在 DIS5K 基准数据集上的大量实验证明 [28], LawDIS 在所有评估指标上显著优于 11 种最新的先进方法 (SOTA)。具体而言, 与次优模型 MVANet 相比, 同时采用 LS 和 WR 策略时, LawDIS 在 F_β^ω 上提升了 4.6%, 仅使用 LS 策略时提升了 3.6%。

本文的主要贡献总结如下:

- 本文提出了 LawDIS, 一个基于语言-窗口的可控框架, 将 DIS 任务重构为以图像为条件的掩码生成问题, 从而实现用户控制的无缝集成, 用于生成高质量的目标掩码。
- 该框架引入了从宏观到微观的控制模式。在宏观模式中, 本文提出语言控制的分割策略 (LS), 根据用户提供的语言提示生成初始掩码。在微观模式中, 本文设计了窗口控制的细化策略 (WR), 对初始掩码中的用户指定区域 (即尺寸可调的窗口) 进行细化。
- 两种模式可通过模式切换器独立或联合运行, 在 DIS5K 基准上各项指标均取得了 SOTA 的性能。

2. 相关工作

图像二值分割。 Qin 等人 [28] 引入了 DIS 任务, 旨在在高分辨率图像中准确分割具有不同结构复杂度的目标, 无论其外观特征如何, 该任务已引起研究社区的广泛关注。随着全卷积网络 (FCN) [23] 的出现, 许多模型在判别式框架下对该任务进行建模, 将其视为一个逐像素分类问题。早期方法 [31, 30, 38, 28, 49, 25] 基于卷积结构, 并引入了如中间监督 [28]、频率先验 [49] 和统一-分解-统一 [25] 等策略来提升性能。近年来, 视觉 Transformer [5] 凭借其强大的长距离依赖建模能力受到青睐, 取得了显著的效果。然而, 由于缺乏卷积所具备的局部归纳偏置, 它们捕获局部结构的能力仍然相对较弱。为克服这一局限性, InSPyReNet [16]、BiRefNet [48] 和 MVANet [45] 在训练或推理阶段引入了不同分辨率的附加图像或图像块作为输入, 在一定程度上增强了对细节信息的捕捉能力。

上述依赖判别式学习范式的 DIS 方法, 主要致力于全局与局部信息的平衡与融合, 但忽略了两个关键挑战: 其一是不同应用场景下灵活的语义控制, 其二是对不可避免的模糊分割进行局部窗口的细化。与这些方法不同, 本文提出的 LawDIS 在生成式模型中重新定义了 DIS 任务, 利用其百科全书式的视觉-语言理解能力以及去噪机制的优势, 实现语言控制的分割与窗口控制的细化。

用于高分辨率分割的扩散模型。目前, 将潜在扩散模型 [33] 应用于高分辨率分割任务 [44] 是计算机视觉领域的一大趋势。鉴于获取高分辨率数据的挑战, 一种合理的做法 [27] 是利用稳定扩散模型强大的生成能力进行数据合成, 从而提升现有高分辨率分割方法的性能。此外, Wang 等人 [39] 提出了一种基于扩散的细化模型, 用于提升不同分割模型所生成目标掩码的质量。随后, 一种两阶段潜在扩散方法 [40] 被用于人像抠图任务。GenPercept [42] 则通过定制解码器将生成模型转化为确定性单步微调范式, 在一系列基础视觉密集感知任务上进行了大量实验, 包括 DIS 和人像抠图等高分辨率分割任务。与以往方法不同, 本文用切换器在单个稳定扩散模型上扩展出宏观与微观控制模式。这不仅可以在同一模型中同时实现高分辨率图像分割和结果细化, 还允许用户在语言和窗口两个层面上进行控制。

3. 方法

本文提出了一个用户可控的框架, 通过将 DIS 任务重新定义为一个以图像为条件的掩码生成范式。本文的框架支持两个层级的控制。具体来说, 宏观模式允许用户在自定义语言提示的引导下, 从高分辨率图像中指定并分割对象。与此同时, 微观模式作为一个通用的后期细化工具, 用于提升分割掩码的准确性, 特别是在结构复杂的区域。为了获得更好的整合性, 本文通过一个模式切换器, 在一个扩散模型中整合了宏观与微观模式, 从而在不同尺度下产生更好的几何表征。接下来将介绍用于构建生成范式的预备知识 (第 3.1节), 以及两个关键流程 (第 3.2节和第 3.3节), 从而重新定义 DIS 任务。

3.1. DIS 的生成式建模

本文将 DIS 任务在条件去噪扩散上重新定义, 重点在于对分割掩码 $s \in \mathbb{R}^{W\times H}$ 的条件概率分布 $D(s \mid x)$ 进行建模。其中条件 $x \in \mathbb{R}^{W\times H\times 3}$ 由对应的待分割 RGB 图像指定。利用预训练的 Stable Diffusion v2 模型 [33] 所建立的框架, 本文提出的 LawDIS 能够在一个低维的潜在空间中高效执行扩散过程。最初, 使用带有编码器 $\phi(\cdot)$ 和解码器 $\varphi(\cdot)$ 的变分自编码器 (VAE) 以在分割掩码与潜在空间之间转换数据, 即 $\mathbf{z}^{(s)} = \phi(s)$ 和 $s \approx \varphi(\mathbf{z}^{(s)})$。类似地, x 也被转换为潜在表示 $\mathbf{z}^{(x)} = \phi(x)$, 作为生成过程的条件。

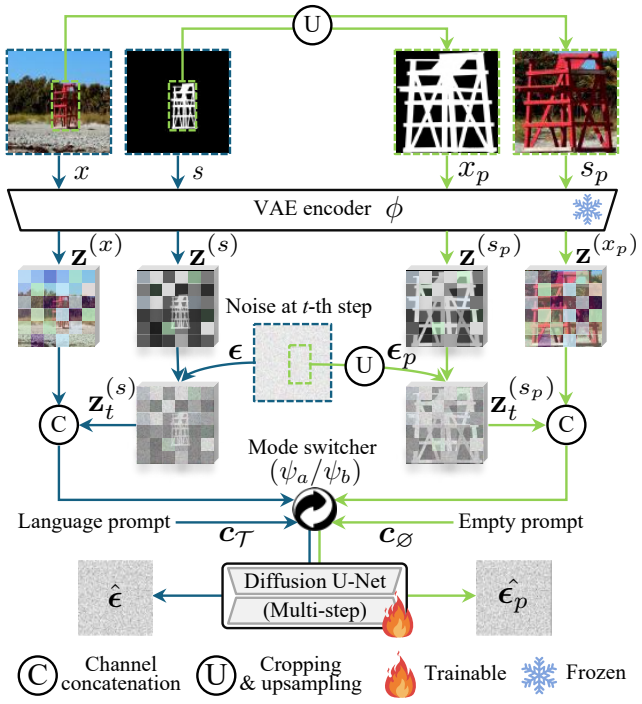

图 2. LawDIS 中 U-Net 训练流程的概览。在预训练的 Stable Diffusion 模型基础上, 引入了模式切换器, 将其扩展为宏观模式与微观模式以进行联合训练, 其中 ψ_a 用于激活宏观模式, ψ_b 用于激活微观模式。宏观模式以整张图像 x 及其分割图 s 为输入, 微观模式则以图像块 x_p 及其对应的分割图块 s_p 为输入。两组输入分别通过 VAE 编码器转化为潜在空间表示, 并分别与语言提示 $\boldsymbol{c}_\mathcal{T}$ 和空提示 $\boldsymbol{c}_\varnothing$ 配对后输入至 U-Net, 生成 $\hat{\boldsymbol{\epsilon}}$ 和 $\hat{\boldsymbol{\epsilon}}_p$, 并用于对分割潜在编码的标准扩散目标的优化。

其次, Stable Diffusion 利用 U-Net 在潜在空间中构建了一个正向加噪过程和一个反向去噪过程。在正向过程中, 从 $\mathbf{z}_0^{(s)} := \mathbf{z}^{(s)}$ 出发, 在每个时间步 $t \in \{1, \ldots, T\}$ 上逐步添加高斯噪声 $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$, 以构建一个离散的马尔可夫链 $\{\mathbf{z}_0^{(s)}, \mathbf{z}_1^{(s)}, \ldots, \mathbf{z}_T^{(s)}\}$。

在反向过程中, 具有已学习参数 θ 的 U-Net 模型 $f_\theta(\cdot)$ 会逐步预测在每一步 t 加到噪声样本 $\mathbf{z}_t$ 上的噪声 $\boldsymbol{\epsilon}$。参数 θ 的更新过程如下: 从训练集中采样一个数据对 (s, x), 并转换为 $(\mathbf{z}^{(s)}, \mathbf{z}^{(x)})$, 在随机的时间步 t 将噪声 $\boldsymbol{\epsilon}$ 加到 $\mathbf{z}^{(s)}$ 上, 预测的噪声 $\hat{\boldsymbol{\epsilon}}$ 通过 $f_\theta(\mathbf{z}_t^{(s)}, \mathbf{z}^{(x)}, t)$ 进行计算。其目标是最小化:

$$\mathbb{E}_{\epsilon\sim\mathcal{N}(0,1),t,\mathbf{z}_0^{(s)},\mathbf{z}^{(x)}}[\|\boldsymbol{\epsilon} - f_\theta(\mathbf{z}_t^{(s)}, \mathbf{z}^{(x)}, t)\|_2^2]. \quad (1)$$

在推理时, 从一个正态分布变量 $\mathbf{z}_T^{(s)} \in \mathcal{N}(\mathbf{0}, \mathbf{1})$ 开始逐步预测噪声 $\hat{\boldsymbol{\epsilon}}$, 然后在去噪调度器 (如 DDPM、DDIM 等) 的引导下逐渐去噪以获得 $\mathbf{z}_0^{(s)}$。随后, 通过解码器 φ 重建预测得到的干净的潜在表示 $\mathbf{z}_0^{(s)}$, 得到掩码预测 $\hat{s} = \varphi(\mathbf{z}_0^{(s)}) \sim D(s \mid x)$。

3.2. 两种模式的联合训练

模式切换器。为了使 LawDIS 能够在两种模式下执行不同的功能, 向稳定扩散模型中引入了一个模式切换器 ψ, 它被表示为一个经过位置编码的一维向量, 并与扩散模型的时间嵌入相加。ψ 被设置为 ψ_a 或 ψ_b, 分别激活宏观模式或微观模式。两种模式在训练过程中相互增强, 并在推理过程中无缝切换。

宏观模式。当激活 ψ_a 时, LawDIS 切换到宏观模式, 在该模式中, 采用语言控制的分割策略, 通过用户提供的语言提示来引导物体的分割。在训练过程中, 模型接收完整的图像 x、分割掩码 s 以及其对应的用户语言提示 $\mathcal{T}$。这些提示由视觉语言模型 (VLM) [1, 2] 生成。更多细节见补充材料。如第 3.1 节所述, x 和 s 会被编码至潜在空间中, 随后送入扩散过程中, 以学习预测添加的高斯噪声 $\boldsymbol{\epsilon}$。根据 [33], 本文使用 CLIP [32] 对语言提示 $\mathcal{T}$ 进行编码, 得到控制的嵌入向量 $\boldsymbol{c}_{\mathcal{T}}$, 并通过交叉注意力机制集成到扩散模型的 U-Net 中。宏观模式下的损失函数如下:

$$\mathcal{L}_{macro} = \|\boldsymbol{\epsilon} - f_\theta(\mathbf{z}_t^{(s)}, \mathbf{z}^{(x)}, \boldsymbol{c}_{\mathcal{T}}, t, \psi_a)\|_2^2. \quad (2)$$

微观模式。当选择 ψ_b 时, LawDIS 激活其微观模式, 采用窗口控制的细化策略, 精确刻画用户指定窗口内的细节。在训练过程中, 选取分割掩码 s 中前景目标的外接矩形作为局部窗口, 而非使用随机窗口。此策略确保前景目标完整地位于窗口内, 从而防止丢失目标或引入语义歧义。基于该窗口选择标准, 从图像 x 和分割掩码 s 中裁剪出对应区域, 得到局部图像块 x_p 和局部分割掩码 s_p。相应地, 噪声 $\boldsymbol{\epsilon}$ 也被裁剪为 $\boldsymbol{\epsilon}_p$。为避免局部图像块与语言提示之间产生不匹配, 采用空提示 $\varnothing$。上述输入被送入 U-Net, 以促进对局部噪声预测的学习。微观模式下的损失函数为:

$$\mathcal{L}_{micro} = \|\boldsymbol{\epsilon}_p - f_\theta(\mathbf{z}_t^{(s_p)}, \mathbf{z}^{(x_p)}, \boldsymbol{c}_{\varnothing}, t, \psi_b)\|_2^2. \quad (3)$$

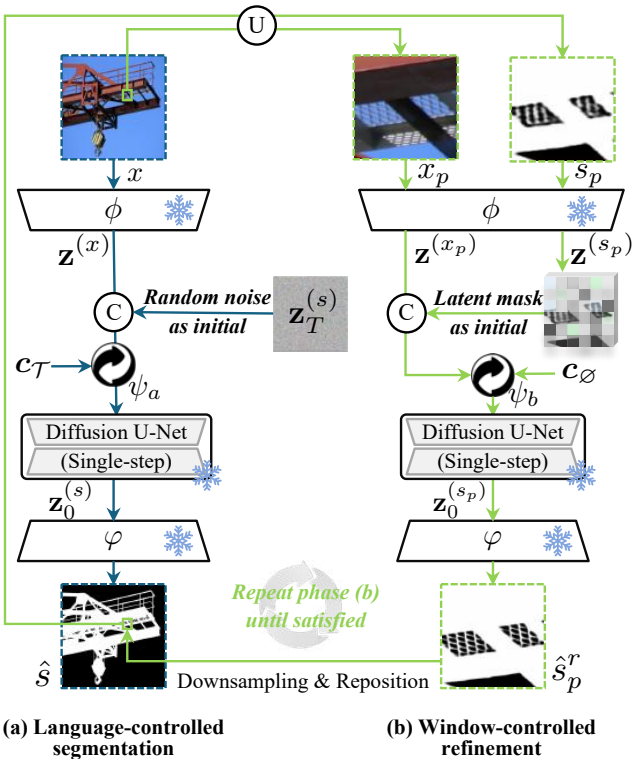

图 3. 推理流程概览, 包括两个步骤。第一步为语言控制的分割, LawDIS 切换至宏观模式, 根据语言提示生成初始分割结果。若用户对细节不满意, 则执行第二步——窗口控制的细化, 在可变分辨率的可控窗口中执行细节优化。细化后的局部图像块将替换初始结果中的对应区域。第二步可无限次地重复执行, 直至获得满意的分割结果。

联合训练方案。为了实现两种模式之间的协同增强, 联合地用两种模式训练 U-Net $f_\theta(\cdot)$, 如图 2 所示。原始的扩散损失函数 Equ.(1) 被重新表述为每种模式的损失之和: $\mathcal{L}_u = \mathcal{L}_{macro} + \mathcal{L}_{micro}$。

在训练完 U-Net 之后, 本文对最初为 RGB 图像重建设计的 VAE 解码器 φ 进行微调, 以使其适应 DIS 任务。首先, 为了保持先前训练结果的有效性, 在此过程中冻结编码器 ϕ 和 U-Net, 仅微调解码器 φ。其次, 通过在编码器和解码器之间添加短连接, 对解码器 φ 进行简单的结构调整。此外, 将解码器的输出层的通道数从 3 减少到 1, 其权重张量通过对通道进行平均来初始化。接下来, 随机将 (图像 x, 文本提示 $\mathcal{T}$, ψ_a) 或 (局部图像块 x_p, 空文本提示 $\varnothing$, ψ_b) 与服从正态分布的噪声 $\mathbf{z}_T^{(s)} \in \mathcal{N}(\mathbf{0}, \mathbf{1})$ 一起输入到模型中。这些输入依次通过 VAE 编码器、U-Net、去噪调度器

和 VAE 解码器, 从而获得预测的分割掩码 $\hat{s}$ 或局部分割掩码 $\hat{s}_p$。整个过程通过标注的掩码 s 或 s_p 进行监督, 其结构损失函数定义如下:

$$\mathcal{L}_d = \begin{cases} \mathcal{L}_{wbce}(\hat{s}, s) + \mathcal{L}_{wiou}(\hat{s}, s) & \text{if } \psi = \psi_a, \\ \mathcal{L}_{wbce}(\hat{s}_p, s_p) + \mathcal{L}_{wiou}(\hat{s}_p, s_p) & \text{if } \psi = \psi_b. \end{cases} \tag{4}$$

其中, $\mathcal{L}_{wbce}(\cdot)$ 和 $\mathcal{L}_{wiou}(\cdot)$ 的定义参见文献 [45, 49]。需要特别指出的是, 为了得到 $\hat{s}$ 或 $\hat{s}_p$, 模型需要执行 T 步去噪, 以从随机噪声中生成干净的潜在特征 $\mathbf{z}_0^{(s)}$ 或 $\mathbf{z}_0^{(s_p)}$。然而, 如果使用诸如 DDIM [34] 这种去噪调度器, 通常需要 50 步才能生成分割图像, 在实际设置中不切实际 [17]。因此, 本文引入轨迹一致性蒸馏 (Trajectory Consistency Distillation, TCD) [47] 作为一种即插即用的去噪调度器, 将采样过程简化为单个步骤。这不仅使得 VAE 解码器 φ 的微调变得可行, 避免了显存溢出问题, 同时也提升了推理效率。

3.3. 两阶段推理

图 3 展示了推理过程的整体概览, 该过程包括两个步骤。第一步是语言引导的分割, 此时 LawDIS 切换至宏观模式, 根据语言提示生成分割结果。第二步是可选的细化阶段, 仅在用户需要调整时才执行。在此过程中, 微观模式被选用, 用户能够在一个分辨率可变的可控窗口内对细节进行完善细节。该过程可被无限次重复, 直到获得令人满意的分割结果。下面提供了两个步骤的细节。

语言控制的分割。 切换器调制为 ψ_a, 激活宏观模式。给定输入图像 x, 使用 VAE 编码器 ϕ 将其转换至潜在空间 $\mathbf{z}^{(x)} = \phi(x)$。随后, 将分割掩码的潜在表示初始化为标准高斯噪声 $\mathbf{z}_T^{(s)} \in \mathcal{N}(\mathbf{0}, \mathbf{1})$。用于控制分割目标的语言提示 $\mathcal{T}$ 可以由视觉语言模型 (VLM) 生成, 也可以由用户自定义。接下来, 将它们共同送入去噪 U-Net 以预测噪声。利用 TCD 调度器 [47] 执行单步去噪, 获得干净的潜在特征 $\mathbf{z}_0^{(s)}$。最后, 通过微调后的 VAE 解码器对特征 $\mathbf{z}_0^{(s)}$ 进行解码, 即可得到语言控制的分割映射 $\hat{s} = \varphi(\mathbf{z}_0^{(s)})$。

窗口控制的细化。 切换器调制为 ψ_b, 激活微观模式。一个关键挑战在于: 由于缺乏全图上下文信息, 将从任意窗口裁剪的局部图像块送入网络进行细化时如何获得可靠的细化结果。为了解决这一问题, 本文提出使用来自全局分割结果的局部图像块代替噪声, 作为扩散的起始点, 从而间接地在两种模式之间传递上下文信息。

具体而言, 用户可以在初始分割图中点击任意区域, 选取不满意的部分作为待细化的窗口。对于每一个待细化窗口, 将裁剪出两个图像块: 其一是从输入图像 x 中裁剪出的局部图像块 x_p, 其二是从第一步中获得的语言控制分割图 $\hat{s}$ 中裁剪出的局部初始分割掩码 $\hat{s}_p$。这些图像块随后被上采样至模型的输入尺寸。接着, 利用编码器 ϕ 分别提取局部图像块的潜在特征 $\mathbf{z}^{(x_p)} = \phi(x_p)$ 和局部初始分割掩码的潜在特征 $\mathbf{z}^{(\hat{s}_p)} = \phi(\hat{s}_p)$。语言提示 $\mathcal{T}$ 此时被替换为空 $\varnothing$ 作为条件。局部初始分割掩码的潜在特征 $\mathbf{z}^{(\hat{s}_p)}$ 与局部图像块的潜在特征 $\mathbf{z}^{(x_p)}$ 合并后共同输入至去噪 U-Net, 以预测噪声。经过 TCD 去噪调度器的去噪以及 VAE 解码器的解码之后, 得到细节更清晰的优化后的局部掩码 $\hat{s}_p^r$。最后, 将 $\hat{s}_p^r$ 替换语言控制分割图 $\hat{s}$ 中对应位置的原始内容。该过程支持多个窗口同时进行细化。当多个窗口存在重叠区域时, 最终优化结果采用重叠区域占比更大的窗口中的值。

4. 实验

4.1. 实验设置

数据集。 本文在 DIS5K 基准数据集 [28] 上进行所有实验, 该数据集包含 5,470 对高分辨率图像与掩码, 涵盖 225 个语义类别。该基准被划分为 DIS-TR(3,000 张图像)、DIS-VD(470 张图像) 和 DIS-TE(2,000 张图像)。LawDIS 在 DIS-TR 上进行所有训练, 并在 DIS-VD 和 DIS-TE 上评估所有模型。DIS-TE 进一步划分为四个子集, 从 DIS-TE1 到 DIS-TE4, 每个子集包含 500 个样本, 用于表示逐渐增加的形状复杂度等级。

评估。 为了进行全面对比, 本文采用五种广泛使用的像素级指标来评估模型能力, 包括加权 F 值 (F_β^ω) [24]、最大 F 值 (F_β^{mx}) [26]、结构度量 ($\mathcal{S}_\alpha$) [4]、平均增强对准度量 (E_ϕ^{mn}) [7], 及平均绝对误差 (M) [26]。

实现细节。 本文使用 PyTorch 框架实现 LawDIS, 并在单张 NVIDIA A100-40GB GPU 上加速运行。为了将 Stable Diffusion v2 [33] 重用于条件掩码生成, 对

方法	DIS-TE1 (500 张图)					DIS-TE2 (500 张图)					DIS-TE3 (500 张图)				
	$F_\beta^\omega \uparrow$	$F_\beta^{mx} \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$E_\phi^{mn} \uparrow$	$F_\beta^\omega \uparrow$	$F_\beta^{mx} \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$E_\phi^{mn} \uparrow$	$F_\beta^\omega \uparrow$	$F_\beta^{mx} \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$E_\phi^{mn} \uparrow$
$BASNet_{19}$ [31]	0.577	0.663	0.105	0.741	0.756	0.653	0.738	0.096	0.781	0.808	0.714	0.790	0.080	0.816	0.848
U^2Net_{20} [30]	0.601	0.701	0.085	0.762	0.783	0.676	0.768	0.083	0.798	0.825	0.721	0.813	0.073	0.823	0.856
$HRNet_{20}$ [38]	0.579	0.668	0.088	0.742	0.797	0.664	0.747	0.087	0.784	0.840	0.700	0.784	0.080	0.805	0.869
$PGNet_{22}$ [41]	0.680	0.754	0.067	0.800	0.848	0.743	0.807	0.065	0.833	0.880	0.785	0.843	0.056	0.844	0.911
$IS\text{-}Net_{22}$ [28]	0.662	0.740	0.074	0.787	0.820	0.728	0.799	0.070	0.823	0.858	0.758	0.830	0.064	0.836	0.883
$InSPyReNet_{22}$ [16]	0.788	0.845	0.043	0.873	0.894	0.846	0.894	0.036	0.905	0.928	0.871	0.919	0.034	0.918	0.943
$FP\text{-}DIS_{23}$ [49]	0.713	0.784	0.060	0.821	0.860	0.767	0.827	0.059	0.845	0.893	0.811	0.868	0.049	0.871	0.922
$UDUN_{23}$ [25]	0.720	0.784	0.059	0.817	0.864	0.768	0.829	0.058	0.843	0.886	0.809	0.865	0.050	0.865	0.917
$BiRefNet_{24}$ [48]	0.819	0.860	0.037	0.885	0.911	0.857	0.894	0.036	0.900	0.930	0.893	0.925	0.028	0.919	0.955
$GenPercept_{24}$ [42]	0.794	0.844	0.038	0.871	0.909	0.828	0.875	0.040	0.887	0.925	0.840	0.890	0.039	0.893	0.939
$MVANet_{24}$ [45]	0.820	0.870	0.037	0.885	0.914	0.875	0.915	0.030	0.917	0.943	0.888	0.929	0.029	0.923	0.953
Ours-S	**0.886**	**0.917**	**0.025**	**0.917**	**0.946**	**0.906**	**0.934**	**0.024**	**0.932**	**0.958**	**0.908**	**0.937**	**0.025**	**0.931**	**0.960**
Ours-R	**0.890**	**0.919**	**0.024**	**0.917**	**0.948**	**0.914**	**0.936**	**0.022**	**0.932**	**0.961**	**0.919**	**0.942**	**0.022**	**0.932**	**0.964**

方法	DIS-TE4 (500 张图)					DIS-TE (1-4) (2,000 张图)					DIS-VD (470 张图)				
	$F_\beta^\omega \uparrow$	$F_\beta^{mx} \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$E_\phi^{mn} \uparrow$	$F_\beta^\omega \uparrow$	$F_\beta^{mx} \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$E_\phi^{mn} \uparrow$	$F_\beta^\omega \uparrow$	$F_\beta^{mx} \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$E_\phi^{mn} \uparrow$
$BASNet_{19}$ [31]	0.713	0.785	0.087	0.806	0.844	0.664	0.744	0.092	0.786	0.814	0.656	0.737	0.094	0.781	0.809
U^2Net_{20} [30]	0.707	0.800	0.085	0.814	0.837	0.676	0.771	0.082	0.799	0.825	0.656	0.753	0.089	0.785	0.809
$HRNet_{20}$ [38]	0.687	0.772	0.092	0.792	0.854	0.658	0.743	0.087	0.781	0.840	0.641	0.726	0.095	0.767	0.824
$PGNet_{22}$ [41]	0.774	0.831	0.065	0.841	0.899	0.746	0.809	0.063	0.830	0.885	0.733	0.798	0.067	0.824	0.879
$IS\text{-}Net_{22}$ [28]	0.753	0.827	0.072	0.830	0.870	0.726	0.799	0.070	0.819	0.858	0.717	0.791	0.074	0.813	0.856
$InSPyReNet_{22}$ [16]	0.848	0.905	0.042	0.905	0.928	0.838	0.891	0.039	0.900	0.923	0.834	0.889	0.042	0.900	0.922
$FP\text{-}DIS_{23}$ [49]	0.788	0.846	0.061	0.852	0.906	0.770	0.831	0.047	0.847	0.895	0.763	0.823	0.062	0.843	0.891
$UDUN_{23}$ [25]	0.792	0.846	0.059	0.849	0.901	0.772	0.831	0.057	0.844	0.892	0.763	0.823	0.059	0.838	0.892
$BiRefNet_{24}$ [48]	0.864	0.904	0.039	0.900	0.939	0.858	0.896	0.035	0.901	0.934	0.854	0.891	0.038	0.898	0.931
$GenPercept_{24}$ [42]	0.801	0.861	0.055	0.869	0.918	0.816	0.868	0.043	0.880	0.923	0.815	0.865	0.043	0.881	0.922
$MVANet_{24}$ [45]	0.866	0.913	0.038	0.910	0.940	0.862	0.907	0.034	0.909	0.938	0.861	0.904	0.035	0.909	0.937
Ours-S	**0.890**	**0.926**	**0.032**	**0.920**	**0.955**	**0.898**	**0.929**	**0.027**	**0.925**	**0.955**	**0.894**	**0.925**	**0.026**	**0.924**	**0.955**
Ours-R	**0.910**	**0.932**	**0.026**	**0.922**	**0.964**	**0.908**	**0.932**	**0.024**	**0.926**	**0.959**	**0.905**	**0.929**	**0.023**	**0.925**	**0.959**

表 1. DIS5K 数据集上与 11 个具有代表性的方法的定量对比。↓ 表示数值越小越好, ↑ 表示数值越大越好。最优结果和次优结果分别用 **红色** 和 **蓝色** 标出。

扩散 U-Net 的输入层进行复制, 以适配拼接后的图像特征与噪声掩码特征。通过复制权重张量来初始化复制层, 并将其数值减半, 以抑制激活值的放大 [15]。在扩散 U-Net 的训练阶段, 采用 1000 步的 DDPM 噪声调度 [11]; 而在 VAE 解码器的训练中, 由于计算资源的限制, 本文使用单步的 TCD 调度器 [47]。两个阶段均使用 32 的批大小, 优化器采用 Adam, 学习率为 3e-5; 其中扩散 U-Net 微调 30K 次迭代, VAE 解码器则微调 6K 次迭代。为了增强视觉多样性, 本文采用随机水平翻转增强以及退火的多分辨率噪声策略。在推理阶段, 为了提高效率, 本文在宏观与微观模式中均采用 TCD 噪声调度器。所有输入, 无论是全尺寸图像还是窗口区域, 在训练与推理中统一缩放至 1024^2 像素大小。

4.2. 与 SOTA 方法的比较

定量比较。 如表 1 所示, 本文在 DIS5K 基准上, 与 11 个众所周知的任务相关方法进行了对比, 包括 BASNet [31]、U^2Net [30]、HRNet [38]、PGNet [41]、IS-Net [28]、InSPyReNet [16]、FP-DIS [49]、UDUN [25]、BiRefNet [48]、GenPercept [42] 以及 MVANet [45]。在统一的输入分辨率 (1024^2 px) 设置下, LawDIS 在语言控制下的基础配置 (Ours-S) 在所有测试集上的所有评价指标中均超越了现有方法 [38, 41, 16, 49, 25, 42, 45, 48, 28]。这些结果充分体现了 LawDIS 在结合用户提示的情况下, 解决 DIS 中的挑战性问题 (尤其是针对多样类别目标) 方面的有效性。例如, 在 DIS-TE1 上, Ours-S 在 F_β^ω 指标上相比于次优模型 MVANet [45] 提升了 6.6%。在微观模式下,

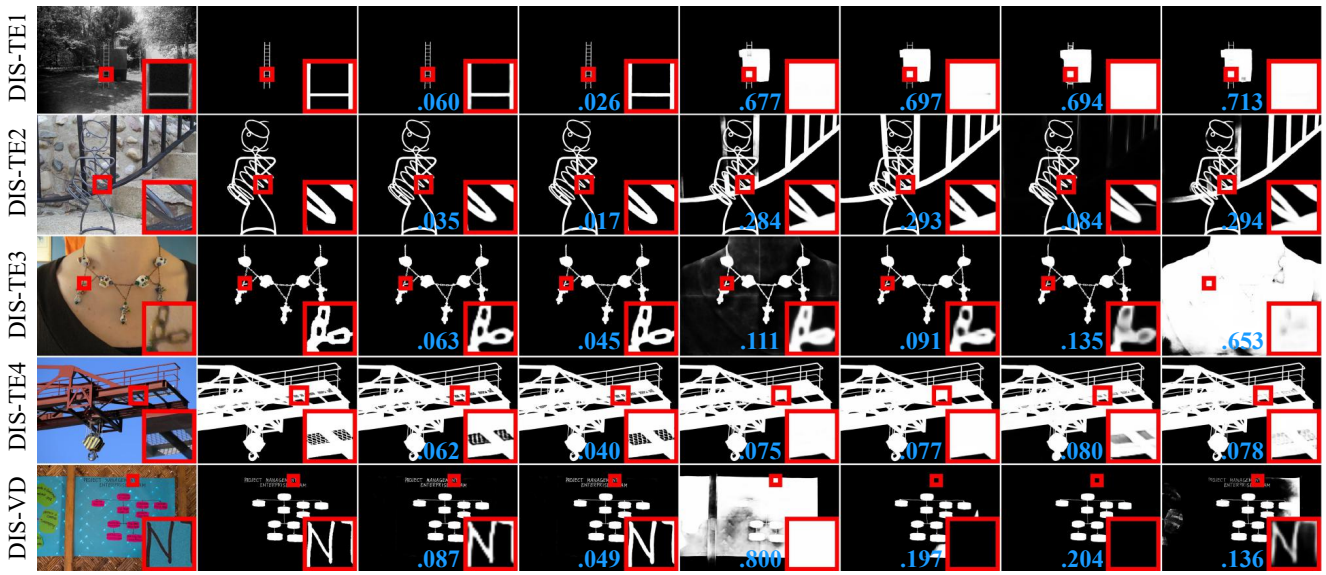

图 4. 本文的方法与四个主流模型的定性对比结果。为便于观察, 使用 $\mathcal{M}$ 指标对局部掩码进行评估。

消融设置	$F_\beta^{mx}\uparrow$	$\mathcal{M}\downarrow$	$\mathcal{S}_\alpha\uparrow$	$E_\phi^{mn}\uparrow$
baseline [33]	0.904	0.047	0.904	0.916
w/o micro-level training	0.912	0.037	0.909	0.943
w/o fine-tuning VAE decoder	0.919	0.040	0.915	0.933
Ours-S	**0.926**	**0.032**	**0.920**	**0.955**

表 2. 基线及通用设置的消融实验。

消融设置	$F_\beta^{mx}\uparrow$	$\mathcal{M}\downarrow$	$\mathcal{S}_\alpha\uparrow$	$E_\phi^{mn}\uparrow$
w/o user prompt (train & test)	0.912	0.036	0.908	0.944
w/o user prompt (test only)	0.915	0.036	0.909	0.947
***w/* user prompt (train & test)**	**0.926**	**0.032**	**0.920**	**0.955**

表 3. LawDIS 宏观控制的消融研究。

Ours-R 对宏观模式生成的初始分割结果 (Ours-S) 进行了进一步优化, 在 DIS-TE4 的 F_β^ω 上提升了 2.0%。这里, 为模拟用户交互的主观性, 沿着 GT 的对象轮廓选择了窗口候选。更多细节见 补充材料。通过将双重控制机制集成到 LawDIS 中, 本文在 DIS-TE1 上相较于 MVANet 实现了 7.0% 的 F_β^ω 增益。

定性对比。图 4 展示了本文的方法与当前最具竞争力的 DIS 模型之间的定性比较。在宏观层面, LawDIS 能够实现对目标区域更加完整的分割; 而在微观层面, 则在处理复杂结构和精细细节时展现出出更高的精度。例如, 第 5 行所示, 从 Ours-S 到 Ours-R, 字的边缘更加锐利, 局部区域的 $\mathcal{M}$ 分数从 0.087 下降至 0.049, 下降了 3.8%。

4.3. 消融实验

接下来, 本文在 DIS-TE4 子集上通过一系列消融实验评估各关键组件的影响。

基线及通用设置。为了揭示稳定扩散模型 [33]在 DIS 任务中的潜力, 本文通过对 LawDIS 进行三项修改来构建一个基线模型: 省略模式切换器, 训练扩散 U-Net 时不使用用户提示, 并且不微调 VAE 解码器。如表2 第一行所示, 基线变体未能取得优异表现。为了评估联合训练策略的影响, 移除模式切换器, 仅使用语言控制来训练 LawDIS。该变体 (表2第二行) 在所有指标上均较 Ours-S 有所下降, 表明双模式协同在为不同输入尺寸提供可扩展的几何表达方面起到了关键作用。此外, 本文还创建了另一变体 (表中第三行), 该变体未微调 VAE 解码器, 其输出通过通道平均生成单通道掩码预测。较差的性能说明, 微调 VAE 解码器对于实现高分辨率分割至关重要, 因为这使得解码过程能够通过细粒度细节补充去噪的掩码特征。

宏观控制的有效性。本文设计了两种设置的实验: 在表 3 的第一行, 用户提示在训练和测试阶段均设置为空; 在第二行, 用户提示仅在测试阶段被省略。与它们相比, 本文的默认设置, 在两个阶段都使用用户提示, 取得了更好的结果。如图 5 所示, 本文提出的模型展示了基于定制的语言提示下灵活分割各种目标物体的能力。相比之下, 其他方法 [45, 48] 缺乏处理语言提示的能力, 只能通过训练期间的记忆模式产生

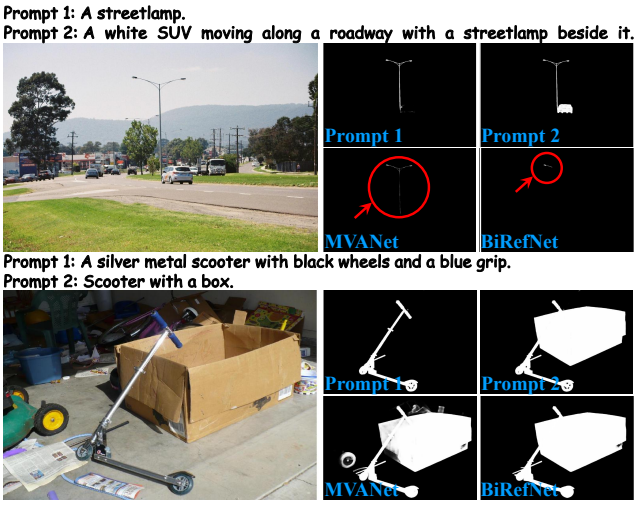

图 5. 不同宏观控制下的定性预测结果。

消融设置	$F_\beta^\omega \uparrow$	$\mathcal{M} \downarrow$	$BIoU^m \uparrow$	$HCE_\gamma \downarrow$
basic setting (*i.e.*, Ours-S)	0.890	0.032	0.795	2481
init. from Gaussian noise	-4.7%	+1.9%	-7.1%	-863
auto windows selection	+1.7%	-0.5%	+2.9%	-767
semi-auto windows selection	**+2.0%**	**-0.6%**	**+3.2%**	**-871**

表 4. LawDIS 微观控制的消融研究。

Methods	$F_\beta^\omega \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$E_\phi^{mn} \uparrow$
IS-Net† [28]	0.753 +6.4%	0.072 -1.6%	0.830 +2.7%	0.870 +4.2%
InSPyReNet† [16]	0.848 +4.2%	0.042 -1.1%	0.905 +0.8%	0.928 +2.6%
UDUN† [25]	0.792 +3.9%	0.059 -1.0%	0.849 +2.0%	0.901 +2.2%
BiRefNet† [48]	0.864 +2.5%	0.039 -0.6%	0.900 +0.7%	0.939 +1.3%
MVANet† [45]	0.866 +3.2%	0.038 -0.9%	0.910 +0.6%	0.940 +1.8%

表 5. 使用 WR 策略作为后优化工具以增强当前的 DIS 方法。性能提升以下划线标出。

固定的结果, 这进一步凸显了 LawDIS 的有效性。

微观控制的有效性。为了更好地揭示精细结构上的性能变化, 本文引入了两个额外指标, 人工矫正量 (HCE_γ) [28] 和边界交并比 ($BIoU^m$) [3]。在表 4 的第二行, 将补丁掩码的潜在变量替换高斯噪声作为输入, 导致 F_β^ω 从 0.890 下降到 0.843。这表明使用分割结果作为扩散过程的起点, 有助于模型获得更精细的掩码。此外, 表 4 第三行展示了一个完全自动的窗口选择过程, 窗口在 Ours-S 初始预测的目标边缘附近被选出, 无需用户干预。结果表明, 即使没有用户干预窗口选择, WR 策略仍有效提升了分割性能。

4.4. 讨论

提出的 WR 策略可以作为通用的后优化工具吗? 为了揭示其兼容性, 本文用 WR 策略去优化多种即插即用的 DIS 方法 [28, 25, 16, 48, 45] 预测出的初始掩码。如表 5 所示, LawDIS 在 DIS-TE4 子集上不同程度地

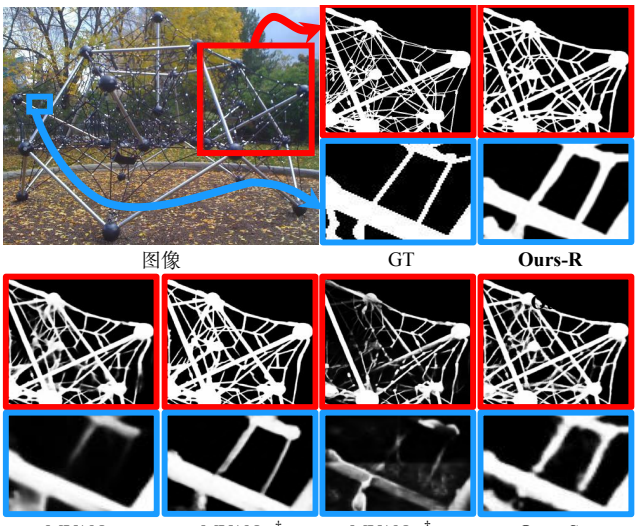

图 6. WR 策略的细化性能与 MVANet [45] 在局部区域上的分割性能的定性比较。† 表示通过应用 WR 策略得到的细化分割结果, 而 ‡ 表示由其他方法 (例如, MVANet) 获得的局部区域的分割结果。

Methods	$F_\beta^\omega \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$E_\phi^{mn} \uparrow$
IS-Net‡ [28]	0.753 -3.6%	0.072 +1.0%	0.830 -4.5%	0.870 -4.5%
InSPyReNet‡ [16]	0.848 -0.8%	0.042 +0.7%	0.905 -2.0%	0.928 -0.5%
UDUN‡ [25]	0.792 -3.8%	0.059 +1.6%	0.849 -3.6%	0.901 -2.8%
BiRefNet‡ [48]	0.864 -3.9%	0.039 +1.6%	0.900 -4.3%	0.939 -3.1%
MVANet‡ [45]	0.866 -0.2%	0.038 +0.6%	0.910 -1.4%	0.940 -0.7%
Ours‡ (Ours-R)	0.890 + **2.0%**	0.032 - **0.6%**	0.920 + **0.2%**	0.955 + **0.9%**

表 6. 局部区域上的分割性能的定量比较。

提升了每个模型的性能。例如, IS-Net [28] 和 MVANet [45] 的预测在 F_β^ω 上分别提高了 6.4% 和 3.2%。图 6 对比了来自整图分割方法 (MVANet) 的放大区域与通过 WR 的策略细化后的局部掩码 (MVANet†)。这表明 WR 策略作为后优化工具的可能性, 以提升现有 DIS 方法的准确性。由于稳定扩散模型是通过从噪声中去噪来生成掩码, 因此, 将去噪起点改为已有的分割结果局部区域的潜在特征, 可以使模型更专注于细节的优化, 而不是重新发现整体结构。

现有的 DIS 方法是否兼容局部图像块作为输入? 为研究这一问题, 采用与 LawDIS 默认设置相同的窗口选择方法和图像块替换技术, 并将局部图像块[1]输入到每个模型中以获取掩码。如表 6 所示, 除本文的方法外, 其余早期方法均表现出不同程度的性能下降。例如, UDUN [25] 和 BiRefNet [48] 的 F_β^ω 分别下降了 3.8% 和 3.9%。此外, 图 6 展示了 MVANet 在整张图像 (MVANet) 与局部图像块 (MVANet‡) 上的分割效

[1]图像块被调整为各模型所需的输入分辨率。

Methods	Fine-tuning VAE	Scheduler	Infer step	$F_\beta^\omega \uparrow$	$\mathcal{M} \downarrow$	FPS ↑
MVANet [45]	-	-	-	0.866	0.038	1.40
Ours-S	×	DDPM [11]	500	0.867	0.038	0.006
	×	DDIM [34]	10	0.835	0.044	0.55
	×	DDIM [34]	50	0.842	0.043	0.12
	×	TCD [47]	1	0.856	0.040	**3.09**
	✓	TCD [47]	1	**0.890**	**0.032**	3.07

表 7. LawDIS 的效率分析。

果对比, 后者表现出明显的性能下降。这表明, 早期方法依赖于固定分辨率输入, 无法适应可变尺寸的输入。

效率分析。本文使用一张 A100 GPU 进行效率分析。首先, 评估使用原始 DDPM 调度器进行 500 步去噪的结果, 结果如表 7 的第二行所示。然后, 探索两种加速版的 DDPM 去噪调度器的策略: 广泛使用的 DDIM 去噪调度器 [34] (见第三和第四行), 以及一步式的 TCD 去噪调度器 [47] (见第五行)。令人惊讶的是, 本文发现 TCD 去噪器在很大程度上保留了 DDPM 去噪器的分割性能, 同时显著提高了效率, 将 FPS 从 0.006 提高到 3.09。此外, 本文微调了 VAE(见第六行), 这在不影响效率的前提下提升了模型对高分辨率图像分割的适应能力。最后, 本文将提出的 LawDIS 模型与次优模型 MVANet [45] 进行了比较, 如表 7 的第一行和第六行所示, 本文提出的模型在分割性能和推理速度方面均表现出显著优势。

5. 总结

本文提出了一个名为 LawDIS 的模型, 它将 DIS 任务重新表述为一个生成式扩散框架, 在单个稳定扩散模型上扩展出宏观与微观控制模式, 从而实现了在两个层面上的可控 DIS。在宏观模式中, 本文提出了 LS 策略, 能够在语言提示的控制下生成分割结果。在微观模式中, 本文设计了 WR 策略, 支持在高分辨率图像上以可变尺度的可控窗口中进行无限次的细节优化。这两种模式通过切换器集成在一个统一的网络中, 从而在训练阶段实现协同增强, 并在推理阶段实现无缝切换。在 DIS5K 数据集上的大量实验表明, LawDIS 在所有评估指标上均显著优于现有的最先进的方法。

致谢。本项目受到了国家自然科学基金 (NO.62376189 & NO.62476143) 的支持。我们诚挚感谢马琦 (南开大学) 和刘静怡 (庆应义塾大学) 在建设性讨论中所做的贡献。

参考文献

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 4

[2] J. Chen, D. Zhu, X. Shen, X. Li, Z. Liu, P. Zhang, R. Krishnamoorthi, V. Chandra, Y. Xiong, and M. Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 4

[3] B. Cheng, R. Girshick, P. Dollár, A. C. Berg, and A. Kirillov. Boundary iou: Improving object-centric image segmentation evaluation. In *IEEE CVPR*, 2021. 8

[4] M.-M. Cheng and D.-P. Fan. Structure-measure: A new way to evaluate foreground maps. *IJCV*, 129(9):2622–2638, 2021. 5

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2

[6] D.-P. Fan, G.-P. Ji, M.-M. Cheng, and L. Shao. Concealed object detection. *IEEE TPAMI*, 44(10):6024–6042, 2021. 2

[7] D.-P. Fan, G.-P. Ji, X. Qin, and M.-M. Cheng. Cognitive vision inspired object segmentation metric and loss function. *SSI*, 6(6):5, 2021. 5

[8] D.-P. Fan, G.-P. Ji, P. Xu, M.-M. Cheng, C. Sakaridis, and L. Van Gool. Advances in deep concealed scene understanding. *VI*, 1(1):16, 2023. 2

[9] D.-P. Fan, J. Zhang, G. Xu, M.-M. Cheng, and L. Shao. Salient objects in clutter. *IEEE TPAMI*, 45(2):2344–2366, 2022. 1

[10] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. *IEEE TPAMI*, 34(10):1915–1926, 2011. 1

[11] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 6, 9

[12] G.-P. Ji, D.-P. Fan, Y.-C. Chou, D. Dai, A. Liniger, and L. Van Gool. Deep gradient learning for efficient camouflaged object detection. *MIR*, 20(1):92–108, 2023. 2

[13] G.-P. Ji, J. Liu, P. Xu, N. Barnes, F. S. Khan, S. Khan, and D.-P. Fan. Frontiers in intelligent colonoscopy. *MIR*, 2025. 1

[14] G.-P. Ji, G. Xiao, Y.-C. Chou, D.-P. Fan, K. Zhao, G. Chen, and L. Van Gool. Video polyp segmentation: A deep learning perspective. *MIR*, 19(6):531–549, 2022. 1

[15] B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, and K. Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *IEEE CVPR*, 2024. 6

[16] T. Kim, K. Kim, J. Lee, D. Cha, J. Lee, and D. Kim. Revisiting image pyramid structure for high resolution salient object detection. In *ACCV*, 2022. 2, 6, 7, 8

[17] M. Li, T. Yang, H. Kuang, J. Wu, Z. Wang, X. Xiao, and C. Chen. Controlnet++: Improving conditional controls with efficient consistency feedback. *ECCV*, 2024. 5

[18] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1

[19] C. Liu, X. Jiang, and H. Ding. Primitivenet: decomposing the global constraints for referring segmentation. *VI*, 2(1):16, 2024. 1

[20] C. Liu, X. Li, and H. Ding. Referring image editing: Object-level image editing via referring expressions. In *IEEE CVPR*, 2024. 1

[21] F. Liu, L. Tran, and X. Liu. Fully understanding generic objects: Modeling, segmentation, and reconstruction. In *IEEE CVPR*, 2021. 1

[22] Y. Liu, Y.-H. Wu, G. Sun, L. Zhang, A. Chhatkuli, and L. V. Gool. Vision transformers with hierarchical attention. *MIR*, 21(4):670–683, 2024. 1

[23] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *IEEE CVPR*, 2015. 2

[24] R. Margolin, L. Zelnik-Manor, and A. Tal. How to evaluate foreground maps. In *IEEE CVPR*, 2014. 5

[25] J. Pei, Z. Zhou, Y. Jin, H. Tang, and P.-A. Heng. Unite-divide-unite: Joint boosting trunk and structure for high-accuracy dichotomous image segmentation. In *ACM MM*, 2023. 2, 6, 8

[26] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *IEEE CVPR*, 2012. 5

[27] H. Qian, Y. Chen, S. Lou, F. Khan, X. Jin, and D.-P. Fan. Maskfactory: Towards high-quality synthetic data generation for dichotomous image segmentation. In *NeurIPS*, 2024. 3

[28] X. Qin, H. Dai, X. Hu, D.-P. Fan, L. Shao, and L. Van Gool. Highly accurate dichotomous image segmentation. In *ECCV*, 2022. 1, 2, 5, 6, 8

[29] X. Qin, D.-P. Fan, C. Huang, C. Diagne, Z. Zhang, A. C. Sant'Anna, A. Suarez, M. Jagersand, and L. Shao. Boundary-aware segmentation network for mobile and web applications. *arXiv preprint arXiv:2101.04704*, 2021. 1

[30] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *PR*, 106:107404, 2020. 2, 6

[31] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand. Basnet: Boundary-aware salient object detection. In *IEEE CVPR*, 2019. 2, 6

[32] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 4

[33] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE CVPR*, 2022. 2, 3, 4, 5, 7

[34] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 5, 9

[35] Y. Song, X. Wang, J. Yao, W. Liu, J. Zhang, and X. Xu. Vitgaze: gaze following with interaction features in vision transformers. *VI*, 2(1):1–15, 2024. 1

[36] A. M. Treisman and G. Gelade. A feature-integration theory of attention. *COGPSY*, 12(1):97–136, 1980. 1

[37] X. Tu, Z. He, Y. Huang, Z.-H. Zhang, M. Yang, and J. Zhao. An overview of large ai models and their applications. *VI*, 2(1):1–22, 2024. 1

[38] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE TPAMI*, 43(10):3349–3364, 2020. 2, 6

[39] M. Wang, H. Ding, J. H. Liew, J. Liu, Y. Zhao, and Y. Wei. Segrefiner: Towards model-agnostic segmentation refinement with discrete diffusion process. *NeurIPS*, 2023. 3

[40] Z. Wang, B. Li, J. Wang, Y.-L. Liu, J. Gu, Y.-Y. Chuang, and S. Satoh. Matting by generation. In *ACM SIGGRAPH*, 2024. 3

[41] C. Xie, C. Xia, M. Ma, Z. Zhao, X. Chen, and J. Li. Pyramid grafting network for one-stage high resolution saliency detection. In *IEEE CVPR*, 2022. 2, 6

[42] G. Xu, Y. Ge, M. Liu, C. Fan, K. Xie, Z. Zhao, H. Chen, and C. Shen. What matters when repurposing diffusion models for general dense perception tasks? In *ICLR*, 2025. 3, 6, 7

[43] H. Ying, Y. Yin, J. Zhang, F. Wang, T. Yu, R. Huang, and L. Fang. Omniseg3d: Omniversal 3d segmentation via hierarchical contrastive learning. In *IEEE CVPR*, 2024. 1

[44] Q. Yu, P.-T. Jiang, H. Zhang, J. Chen, B. Li, L. Zhang, and H. Lu. High-precision dichotomous image segmentation via probing diffusion capacity. In *ICLR*, 2025. 3

[45] Q. Yu, X. Zhao, Y. Pang, L. Zhang, and H. Lu. Multi-view aggregation network for dichotomous image segmentation. In *IEEE CVPR*, 2024. 1, 2, 5, 6, 7, 8, 9

[46] Y. Zeng, P. Zhang, J. Zhang, Z. Lin, and H. Lu. Towards high-resolution salient object detection. In *IEEE ICCV*, 2019. 1

[47] J. Zheng, M. Hu, Z. Fan, C. Wang, C. Ding, D. Tao, and T.-J. Cham. Trajectory consistency distillation. *arXiv preprint arXiv:2402.19159*, 2024. 5, 6, 9

[48] P. Zheng, D. Gao, D.-P. Fan, L. Liu, J. Laaksonen, W. Ouyang, and N. Sebe. Bilateral reference for high-resolution dichotomous image segmentation. *CAAI AIR*, 3:9150038, 2024. 2, 6, 7, 8

[49] Y. Zhou, B. Dong, Y. Wu, W. Zhu, G. Chen, and Y. Zhang. Dichotomous image segmentation with frequency priors. In *IJCAI*, 2023. 2, 5, 6

[50] M. Zhuge, D.-P. Fan, N. Liu, D. Zhang, D. Xu, and L. Shao. Salient object detection via integrity learning. *IEEE TPAMI*, 45(3):3738–3752, 2022. 1