MaskFactory：面向二分图像分割的高质量合成数据生成

钱浩天 [1*] 陈胤达 [1*] 楼圣韬 [1] Fahad Shahbaz Khan[3,4]
金小刚 [1†] 范登平 [2,3]
[1]State Key Lab of CAD&CG, 浙江大学 [2]VCIP&CS, 南开大学
[3]MBZUAI [4]Linköping University

摘要

二分图像分割（Dichotomous Image Segmentation, DIS）任务需要高度精确的标注，而传统的数据集构建方法耗费人力、成本高昂，并且需要大量的领域专业知识。尽管使用合成数据来解决 DIS 面临的这些挑战是一种有前景的方案，但当前的生成模型与技术仍难以应对场景偏移、噪声引起的误差以及训练样本多样性有限等问题。为此，我们提出了一种称为 MaskFactory 的新方法，它为生成多样且精确的数据集提供了一种可扩展的解决方案，显著减少了数据准备的时间与成本。首先，我们提出一种通用的掩码编辑方法，将刚性编辑与非刚性编辑技术相结合以生成高质量的合成掩码。具体而言，刚性编辑利用扩散模型中的几何先验，在零样本条件下实现精确的视角变换；非刚性编辑则采用对抗训练与自注意力机制，实现复杂且拓扑一致的修改。随后，我们使用多条件可控生成方法生成高分辨率图像与精确分割掩码的配对数据。最后，我们在广泛使用的 DIS5K 数据集基准上的实验表明，与现有方法相比，本方法在质量与效率方面均表现更优。代码可在 `https://qian-hao-tian.github.io/MaskFactory/` 获取。

1 引言

二分图像分割（DIS）旨在从自然图像中精确地分割出物体 [1, 2]，是医学图像 [3, 4, 5]、自动驾驶 [6] 和连接组学研究 [7, 8] 等计算机视觉应用中的重要任务。然而，为 DIS 任务收集数据集的传统方法耗费人力、成本高昂，并且需要大量的领域专业知识。近年来，合成数据已成为一种有前景的解决方案，可用于大规模生成多样且精确的数据集，为模型训练提供了可扩展且经济高效的手段。鉴于高质量合成数据对模型训练的重要性，发展此类面向 DIS 任务的数据集生成方法成为关键。尽管生成模型已被用于辅助合成数据集的制作，但它们在可控性、精确性和多样性方面仍存在显著局限。

近期的合成方法 [9, 10, 11] 利用扩散模型 [12]，基于文本线索与注意力图来合成图像-掩码对。尽管取得了进展，它们仍面临三个主要挑战：(1) *可控性*：文本引导的生成方案可能偏离真实世界场景，尤其是在细粒度标签的情况下，导致生成图像失控。(2) *精确性*：注意力图的使用会引入噪声，降低掩码保真度并给需要精确像素级分割的应用（如 DIS [2]）带来

*共同第一作者。范登平担任本工作的项目负责人
†通讯作者。联系方式：jin@cad.zju.edu.cn

本文为NeurIPS 2024 的中文翻译版，由秦曦翻译，范登平、杜泽文校稿。

图 1: 展示了第一阶段编辑得到的掩码以及第二阶段生成的对应图像。在示例中，我们将公园长椅和桌子的视角从正视图变换为俯视图，并编辑了它们的形状，将公园长椅从弧形边缘改为方形边缘，将桌子从方形改为圆形。

挑战。(3) 多样性：先前的工作 [13, 14, 15] 相对单一，生成多样化训练样本的能力有限，因而不适合提供具有充分多样性的训练数据。

由于提示词的复杂性，生成多样化的人物图像常常构成挑战，这限制了其在需要极高可变性的任务中的适用。值得注意的是，对细粒度数据集（如 DIS 数据集）的扩充目前缺乏有效的解决方案。事实上，我们观察到现有的 DIS 数据集主要由单一来源的正射影像构成，相似物体之间的形状多样性有限。生成能够准确反映真实世界分布的更多样化 DIS 图像，仍然是一个尚未解决的挑战。

为应对这些挑战，我们提出了 MaskFactory，这是一种新颖的两阶段方法，能够高效地为 DIS 任务生成高质量的合成数据集。基于物体的几何特性，我们的方法同时考虑目标物体的刚性变换与非刚性变换。这意味着在生成合成数据集时，我们不仅考虑视角的变化（刚性变换），也考虑形变（非刚性变换），如图 1 所示。在第一阶段，刚性变换由从大规模扩散模型中学习到的几何先验驱动，实现精确的视角变化，并模拟观察角度与尺度的多样化变化。非刚性变换利用大语言模型提供的提示词，通过基于注意力的编辑精确地改变目标物体的形状，并通过拓扑保持对抗训练确保编辑前后的拓扑一致性。因此，即使对于具有复杂几何结构的掩码输入，我们的方法也能够为 DIS 任务生成高质量且多样的合成掩码。在第二阶段，我们利用多种控制输入来引导生成过程，包括掩码、Canny 边缘以及表示类别信息的提示词。这些输入被送入扩散模型，以生成与准备好的分割掩码相匹配的高分辨率图像。该方法确保了图像与掩码之间的高度一致性，同时提升了数据集的真实感与多样性。

我们的方法显著提升了生成数据集的真实感与多样性。我们在专门为评估 DIS 性能而设计的 DIS5K 数据集 [2] 上验证了 MaskFactory 的有效性。实验结果表明，我们的方法在结构保持与误差指标方面优于现有的数据集合成方法，取得了 8.8% 的平均性能增益。这些发现凸显了 MaskFactory 在提供 DIS 任务所需的数据多样性与标注精度方面的潜力，同时显著降低了与数据集准备相关的时间与成本。

总之，我们工作的贡献如下：

- 我们提出了 MaskFactory，这是一种为 DIS 任务生成高质量的数据集的新颖方法，能够提高质量、精度和效率。
- 我们提出了一种两阶段方法，通过掩码编辑合成高质量且多样的物体掩码，并使用多条件可控生成方法生成对应的高分辨率图像。
- 在 DIS5K 数据集上的实验结果表明，与现有的数据集合成方法相比，MaskFactory 具有更优的性能。

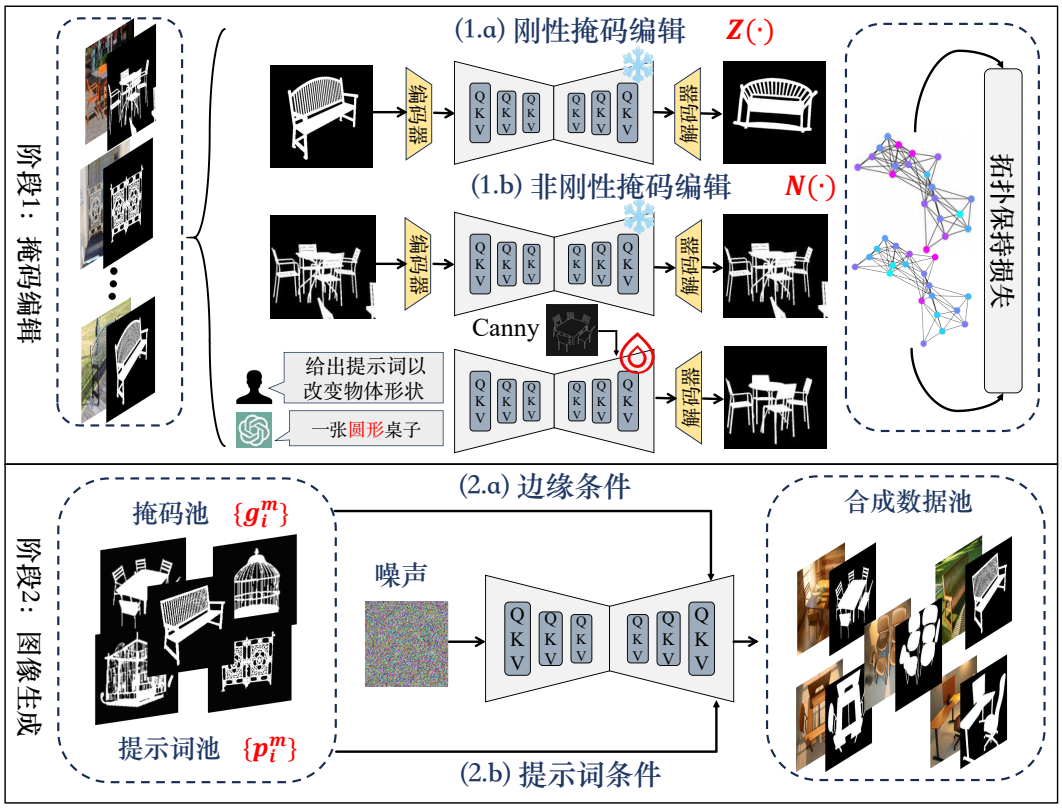


图 2: MaskFactory 的工作流程。在第一阶段，我们通过对现有的真值掩码施加刚性与非刚性编辑来生成新的掩码。在第二阶段，我们使用生成的掩码及其对应提取的 Canny 边缘作为条件，并结合表示类别的提示词来生成 RGB 图像。该过程为我们的生成模型构成了配对数据。

2 相关工作

合成数据 合成数据在多种机器学习与计算机视觉任务中受到了广泛关注，例如自然语言处理 [16]、目标检测 [17] 和图像分割 [10]。例如，Lin 等人 [17] 证明合成数据能够提升目标检测的性能，尤其是在难以获取真实世界数据的场景下。生成对抗网络（GAN）与变分自编码器（VAE）是两种流行的基于深度学习的合成数据生成方法 [18]。近期研究聚焦于提升合成数据的质量与多样性 [19] 并探索其在小样本学习中的应用 [20]。此外，合成数据已被证明能够增强机器学习模型的可解释性与可说明性 [21]。得益于扩散模型 [12, 22] 的最新进展，DatasetDM [9] 和 Dataset Diffusion [10] 等方法能够为多种计算机视觉任务生成高质量的合成数据。然而，这些方法由于引入了伪标签，可能会带来额外的误差。这些误差源于伪标签可能带有噪声且不准确，从而导致得到次优的训练数据。尽管有一些研究 [23] 采用类似 ControlNet 的控制方案 [24] 来缓解这些问题，但生成的合成数据仍可能存在噪声与误差，使其不适用于 DIS 任务。因此，本文聚焦于为 DIS 任务生成高质量的合成数据，其要求生成的图像-掩码对需要高度精确与准确。

二分图像分割 随着高分辨率成像技术的出现，DIS 取得了显著进展。代表性方法包括 IS-Net [2] 中的中间监督策略、频率先验方法 [1] 以及 [25] 提出的 unite-divide-unite 策略。然而，这些方法虽然提升了分割精度，却往往难以捕获极其精细的细节。因此，近期的渐进式细化策略（例如 BASNet [26]）强调了梯度图、多尺度输入等辅助信息的重要性。这些方法提出利用梯度特征与真值监督来增强复杂区域中弱特征的学习。随后，BiRefNet [1] 保持高分辨率输入，并采用双边参考框架以更好地捕获复杂细节。最新的工作，如 Yu 等人 [27] 的多视图聚合网络以及 Liu 等人 [28] 的交互式分割方法，通过整合多样化的提示以及增强特征表示以实现高质量分割，进一步推进了该领域的发展。

3 方法

3.1 MaskFactory 概述

当前的图像分割方法由于对有限的人工标注数据依赖而被明显制约，这既限制了性能也阻碍了泛化能力。为缓解这一问题，研究者们通常采用伪标签生成来扩充训练数据集。然而在 DIS 的场景下，这些方法常常引入伪影，降低分割质量。此外，现有的图像编辑技术往往无法保持二值掩码的拓扑结构，导致目标区域内出现不连续与重叠。

为应对这些挑战，我们提出了 MaskFactory 框架，旨在从原始数据集 $\mathcal{D} = \{(I_i^r, I_i^m)\}_{i=1}^N$ 生成大量高质量的合成图像-掩码对 $\mathcal{G} = \{(g_i^r, g_i^m)\}_{i=1}^M$。该方法旨在提升 DIS 模型的性能。如图 2 所示，我们的框架包括两个主要阶段：掩码编辑与图像生成。

在掩码编辑阶段，来自原始数据集的源掩码通过刚性与非刚性编辑方法进行变换，得到一组高精度的合成掩码 $\mathcal{G}^m = \{g_i^m\}_{i=1}^M$。刚性掩码编辑从多种视角生成合成掩码，而非刚性掩码编辑采用拓扑保持对抗训练机制，根据语义提示词编辑掩码，同时保持源掩码的结构完整性。

在图像生成阶段，模型利用多条件可控生成方法，以合成掩码作为条件约束，生成与合成掩码精确对应的逼真 RGB 图像 $\mathcal{G}^r = \{g_i^r\}_{i=1}^M$。

3.2 掩码编辑阶段

3.2.1 刚性掩码编辑

刚性掩码编辑旨在通过刚性变换保留源掩码的细节信息。我们利用 Zero123[29] 方法，该方法采用视角条件扩散模型 ψ_θ 来操控掩码的视角。给定目标视角的相对相机旋转与平移 $\mathbf{T}_i$，ψ_θ 基于源掩码 I_i^m 合成新掩码 g_i^m ，即 $g_i^m = \psi_\theta(I_i^{m'}, \mathbf{T}_i)$，其中 $I_i^{m'}$ 是源掩码 I_i^m 的反相图像，以确保主要成分为非零。

3.2.2 非刚性掩码编辑

非刚性掩码编辑受 MasaCtrl [30] 启发，是 MaskFactory 的关键组成部分。与 MasaCtrl 直接利用文本提示词 P_i 操控源掩码 I_i^m 以生成合成掩码 g_i^m 不同，我们引入了拓扑保持对抗训练机制，以缓解二值掩码编辑中的伪影与结构退化。文本提示词 P_i 来自一个提示词池 $\{p_i^m\}$ ，该提示词池是使用 GPT-4 基于原始图像生成的。这些提示词提供了指导掩码编辑过程的详细描述。该模块由生成器 G_θ 与判别器 D_ϕ 组成。生成器在文本提示词 P_i 与源掩码 I_i^m 的引导下，将噪声 $\mathbf{z}$ 转换为合成掩码 g_i^m。互注意力机制将 g_i^m 的查询特征 $\mathbf{Q}_t$ 与 I_i^m 的键和值特征 $\mathbf{K}_s, \mathbf{V}_s$ 对齐，以确保编辑过程中的一致性。为避免前景-背景混淆，从交叉注意力图中提取掩码 $\mathbf{M}$ 以引导模型的关注区域。

拓扑保持对抗训练 为保持源掩码的结构信息，我们首先使用边缘检测算子 $\mathcal{E}$ 提取边缘图 $E_s = \mathcal{E}(I_i^m)$，获得关键点 $\mathcal{V} = \{v_j\}_{j=1}^{N_v}$。然后基于这些关键点构建图 $\mathcal{T} = (\mathcal{V}, \mathcal{E}_s)$。判别器 D_ϕ 分别对合成掩码 g_i^m 与源掩码的结构图 $\mathcal{T}_g$ 和 $\mathcal{T}_s$ 进行对抗训练，以确保拓扑一致性。

判别器 D_ϕ 的训练目标是最大化：

$$\max_{\phi} \mathbb{E}_{\mathcal{T}_s \sim p_{\text{data}}(\mathcal{T}_s)}[\log D_\phi(\mathcal{T}_s)] + \mathbb{E}_{\mathcal{T}_g \sim p_{\text{gen}}(\mathcal{T}_g)}[\log(1 - D_\phi(\mathcal{T}_g))], \tag{1}$$

其中 $p_{\text{data}}(\mathcal{T}_s)$ 和 $p_{\text{gen}}(\mathcal{T}_g)$ 分别表示源掩码与合成掩码的结构图分布。相反，生成器 G_θ 的训练目标是最小化判别器的判别能力：

$$\min_{\theta} \mathbb{E}_{\mathcal{T}_g \sim p_{\text{gen}}(\mathcal{T}_g)}[\log(1 - D_\phi(\mathcal{T}_g))]. \tag{2}$$

通过拓扑保持对抗训练，非刚性编辑模块在编辑过程中有效地保留了源掩码的结构信息，生成高质量、无伪影的合成掩码。

非刚性掩码编辑的总损失函数 $\mathcal{L}_{\text{total}}$ 包括对抗损失 $\mathcal{L}_{\text{GAN}}$、内容损失 $\mathcal{L}_{\text{content}}$ 以及结构保持损失 $\mathcal{L}_{\text{structure}}$：

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{GAN}} + \lambda_1 \mathcal{L}_{\text{content}} + \lambda_2 \mathcal{L}_{\text{structure}}, \tag{3}$$

其中 λ_1 和 λ_2 是平衡因子。

对抗损失 $\mathcal{L}_{\text{GAN}}$ 定义为：

$$\mathcal{L}_{\text{GAN}} = \mathbb{E}_{\mathcal{T}_s \sim p_{\text{data}}(\mathcal{T}_s)}[\log D_\phi(\mathcal{T}_s)] + \mathbb{E}_{\mathcal{T}_g \sim p_{\text{gen}}(\mathcal{T}_g)}[\log(1 - D_\phi(\mathcal{T}_g))]. \tag{4}$$

内容损失 $\mathcal{L}_{\text{content}}$ 度量合成掩码与文本提示词之间的语义一致性：

$$\mathcal{L}_{\text{content}} = \|g_i^m - I_i^m\|_1, \tag{5}$$

其中 g_i^m 是合成掩码，I_i^m 是源掩码。

拓扑保持损失 $\mathcal{L}_{\text{structure}}$ 评估合成掩码与源掩码的结构图之间的差异：

$$\mathcal{L}_{\text{structure}} = \|\mathcal{T}_g - \mathcal{T}_s\|_1. \tag{6}$$

这些组成部分确保编辑过程保持源掩码的结构与语义一致性。

3.3 图像生成阶段

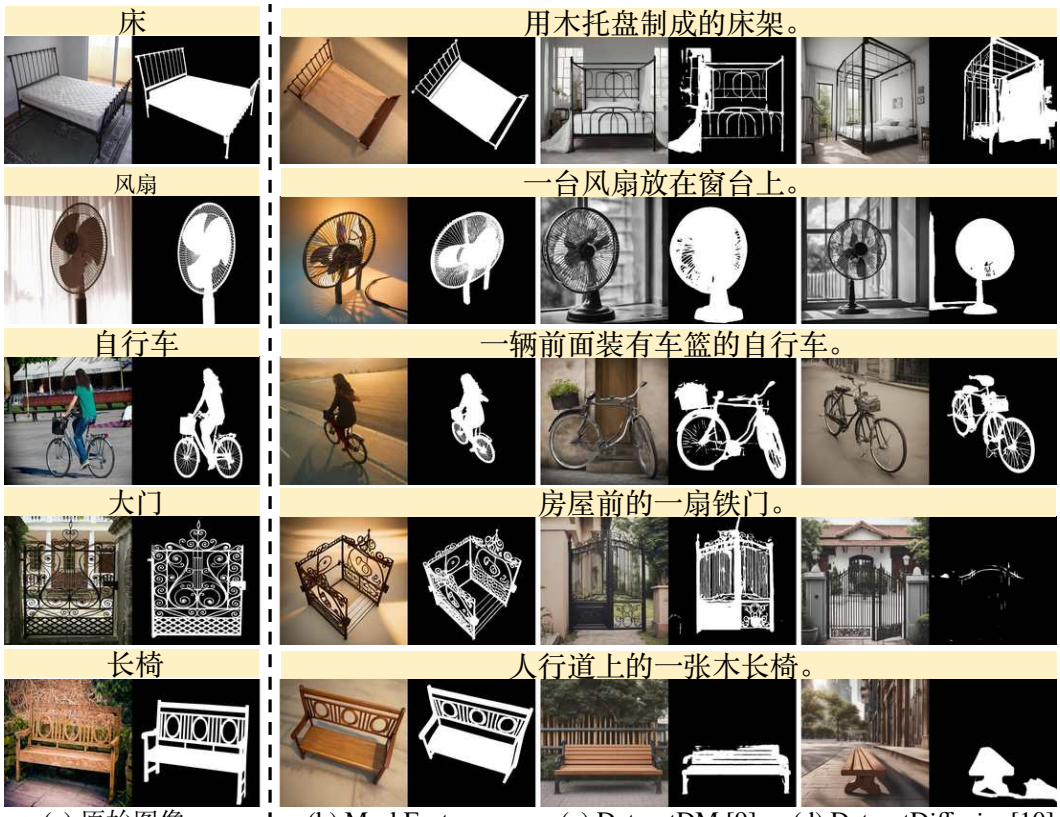
在掩码编辑阶段之后，我们得到一个合成掩码池 $\{g_i^m\}_{i=0}^M$，其中包含通过上述变换生成的精细合成掩码。在随后的图像生成阶段，我们使用多条件可控生成方法，为合成掩码池中的掩码精确地生成对应的 RGB 图像 $\{g_i^r\}_{i=0}^M$。RGB 图像的主要分割区域与对应的合成掩码对齐。受 ControlNet [24] 启发，我们引入了多条件可控生成方法以实现精确的 RGB 图像生成。该方法同时注入分割条件 c_i^s 和 Canny 条件 c_i^y 以引导随机高斯噪声的去噪过程，确保生成的 RGB 图像 $\{g_i^r\}_{i=0}^M$ 与合成掩码 $\{g_i^m\}_{i=0}^M$ 精确对应。

由于合成掩码本身即作为分割条件 c_i^s，当给定合成掩码 g_i^m 时，我们只需使用 Canny 算子提取 Canny 条件 c_i^y：

$$c_i^s = g_i^m, \quad c_i^y = Canny(g_i^m) \tag{7}$$

在获得 Canny 条件与分割条件后，我们将其输入由一组神经网络层构成的模块 B_θ。最后，这些条件被注入预训练扩散模型 M_θ，控制噪声 z 的去噪过程以生成对应的 RGB 图像。更多细节请读者参阅 [24]。

$$g_i^r = M_\theta(z, B_\theta(c_i^s), B_\theta(c_i^y)) \tag{8}$$



(a) 原始图像 (b) MaskFactory (c) DatasetDM [9] (d) DatasetDiffusion[10]

图 3: 与基线方法的比较。

4 实验与结果

4.1 数据集与指标

我们在 DIS5K 数据集上开展实验，该数据集包含 5,479 张高分辨率图像，涵盖各种背景中的伪装物体、显著物体或精细物体。DIS5K 数据集划分为三个子集：用于训练的 DIS-TR（3,000 张图像）、用于验证的 DIS-VD（470 张图像）以及用于测试的 DIS-TE（2,000 张图像）。对于数据增强，我们利用训练子集（DIS-TR）中的掩码部分。

为评估我们的模型，我们采用一组多样化的指标以确保全面的性能评估。这些指标包括：max F_1[31]，它在精确率与召回率之间取得平衡，给出反映整体准确性的调和平均值；F_β^ω[32]，一种补偿类别不平衡的加权 F 度量，取值范围为 0 到 1，值越高表示性能越优；M（平均绝对误差）[33]，计算预测掩码与真值掩码之间的平均绝对差异，值越低表示准确性越好；S_α[34]，一种结构相似性度量，评估图像中重要结构的保持情况，值越接近 1 表示性能越好；以及 E_M^ϕ[35]，一种同时考虑像素级与图像级信息以进行更全面评估的增强度量，值越高表示性能越好。总的来说，这些指标为评估我们分割模型的有效性与可靠性提供了稳健的框架。

4.2 实现细节

我们的图像编辑框架使用 PyTorch 实现，并在 8 张 NVIDIA GeForce RTX 3090 GPU 上训练。我们同时采用非刚性与刚性编辑方法来操控 DIS-TR 数据集中的图像，分别利用空间变换器与仿射变换来建模非刚性与刚性形变。我们模型使用的超参数如下：批大小为 16，图像尺寸为 512×512，编辑迭代次数为 5，学习率为 0.001，权重衰减为 0.0001，扩散步数为 1000，扩散步长为 0.1。式（3）中的超参数设置为 $\lambda_1 = 0.8$ 和 $\lambda_2 = 0.5$。

表 1: 不同生成方法的实验结果比较，展示了使用不同数量的生成图像对分割任务性能的影响。

	数据集	IS-Net	DatasetDM [9]				Dataset Diffusion [10]				MaskFactory			
			2500	5000	7500	10000	2500	5000	7500	10000	2500	5000	7500	10000
DIS-VD	$maxF_1 \uparrow$	0.791	0.792	0.791	0.788	0.785	0.793	0.780	0.771	0.767	0.831	0.833	0.834	**0.835**
	$F_\beta^\omega \uparrow$	0.717	0.720	0.719	0.712	0.710	0.726	0.726	0.716	0.710	0.725	0.754	0.757	**0.759**
	$M \downarrow$	0.074	0.076	0.076	0.075	0.077	0.074	0.077	0.076	0.077	0.073	0.073	**0.071**	0.072
	$S_\alpha \uparrow$	0.813	0.814	0.810	0.807	0.805	0.826	0.821	0.804	0.790	0.832	0.855	0.860	**0.866**
	$E_M^\phi \uparrow$	0.856	0.869	0.864	0.864	0.860	0.868	0.859	0.852	0.838	0.880	0.911	0.914	**0.923**
DIS-TE1	$maxF_1 \uparrow$	0.740	0.744	0.744	0.740	0.736	0.741	0.727	0.721	0.719	0.776	0.777	0.779	**0.784**
	$F_\beta^\omega \uparrow$	0.662	0.670	0.668	0.663	0.659	0.675	0.669	0.667	0.654	0.677	0.700	0.703	**0.705**
	$M \downarrow$	0.074	0.075	0.077	0.075	0.076	0.076	0.074	0.077	0.078	0.073	**0.071**	**0.071**	0.073
	$S_\alpha \uparrow$	0.787	0.791	0.786	0.781	0.778	0.790	0.776	0.760	0.754	0.803	0.822	0.826	**0.829**
	$E_M^\phi \uparrow$	0.820	0.825	0.819	0.813	0.810	0.827	0.816	0.802	0.798	0.853	0.866	0.868	**0.875**
DIS-TE2	$maxF_1 \uparrow$	0.799	0.810	0.807	0.801	0.793	0.801	0.799	0.782	0.776	0.808	0.814	0.819	**0.822**
	$F_\beta^\omega \uparrow$	0.728	0.741	0.736	0.736	0.734	0.739	0.737	0.721	0.712	0.739	0.764	0.765	**0.772**
	$M \downarrow$	0.070	0.072	0.070	0.071	0.071	0.072	0.071	0.073	0.071	0.068	0.067	0.068	**0.067**
	$S_\alpha \uparrow$	0.826	0.833	0.828	0.823	0.821	0.833	0.832	0.811	0.795	0.849	0.855	0.861	**0.865**
	$E_M^\phi \uparrow$	0.858	0.870	0.868	0.866	0.865	0.861	0.860	0.850	0.834	0.868	0.895	0.901	**0.903**
DIS-TE3	$maxF_1 \uparrow$	0.830	0.846	0.845	0.839	0.833	0.838	0.834	0.831	0.812	0.841	0.850	0.867	**0.870**
	$F_\beta^\omega \uparrow$	0.758	0.770	0.766	0.763	0.757	0.769	0.752	0.733	0.728	0.761	0.782	0.783	**0.785**
	$M \downarrow$	0.064	0.066	0.066	0.065	0.066	0.066	0.065	0.066	0.065	**0.062**	**0.062**	**0.062**	0.063
	$S_\alpha \uparrow$	0.836	0.848	0.844	0.839	0.832	0.844	0.837	0.823	0.819	0.842	0.859	0.866	**0.878**
	$E_M^\phi \uparrow$	0.883	0.894	0.887	0.883	0.878	0.900	0.896	0.878	0.858	0.911	0.915	0.918	**0.926**
DIS-TE4	$maxF_1 \uparrow$	0.827	0.833	0.826	0.820	0.816	0.830	0.818	0.804	0.801	0.856	0.876	0.876	**0.879**
	$F_\beta^\omega \uparrow$	0.753	0.759	0.754	0.750	0.746	0.760	0.759	0.758	0.747	0.789	0.824	0.824	**0.830**
	$M \downarrow$	0.072	0.074	0.073	0.073	0.075	0.074	0.072	0.076	0.074	0.071	**0.069**	0.071	0.072
	$S_\alpha \uparrow$	0.830	0.846	0.839	0.838	0.833	0.841	0.830	0.822	0.815	0.842	0.852	0.860	**0.862**
	$E_M^\phi \uparrow$	0.870	0.885	0.881	0.873	0.868	0.873	0.866	0.845	0.836	0.891	0.917	0.917	**0.923**

我们使用 DIS5K 数据集的 DIS-TR 子集训练分割模型，使用 2 张 NVIDIA GeForce RTX 3090 GPU。网络的输入尺寸为 512×512，学习率为 0.0001，批大小为 48。模型使用 Adam 优化器共优化 800 个 epoch。由我们模型生成的编辑图像与原始训练集合并，构成新的训练集，使模型能够同时从原始图像与编辑图像中学习。

我们在 DIS5K 数据集上评估我们的模型，使用每个数据集的测试集。为进行全面比较，我们还使用默认参数复现了两种最先进的基于扩散的图像生成方法 DatasetDM 与 Dataset Diffusion。将生成的图像与我们方法生成的图像进行比较，结果可视化如图 3 所示。

4.3 结果

按生成图像数量划分的结果 我们在 DIS5K 数据集上，将所提方法 MaskFactory 与两种最先进的基线 DatasetDM [9] 和 Dataset Diffusion [10] 进行性能评估。DIS5K 数据集包含五个子数据集：DIS-VD、DIS-TE1、DIS-TE2、DIS-TE3 与 DIS-TE4。我们使用稳健的分割方法 IS-Net [2] 作为 DIS 任务的基线。对于两种基线 DatasetDM 与 Dataset Diffusion，我们使用其开源实现中提供的默认参数。

首先，我们在 DIS-VD 验证集上确定表现最佳的模型，然后在其他子数据集上评估其性能。模型在 DIS-TR 数据集上训练，该数据集分别用规模为 2500、5000、7500 与 10000 张图像的生成数据集进行扩充。

实验结果如表 1 所示。我们提出的方法 MaskFactory 在所有子数据集与评估指标上均一致地优于基线。随着生成图像数量的增加，MaskFactory 的性能不断提升，在 10000 张生成图像

表 2: 使用不同分割方案的实验结果比较，展示了在固定数量的生成图像下的性能增益。

	方法	IS-Net [2]		FP-DIS [36]		UDUN [25]		BiRefNet [1]		SAM-HQ [37]	
		不使用本方法	使用本方法	不使用本方法	使用本方法	不使用本方法	使用本方法	不使用本方法	使用本方法	不使用本方法	使用本方法
DIS-TE1	$maxF_1 \uparrow$	0.740	0.784	0.784	0.805	0.784	0.799	0.866	0.882	0.897	0.905
	$M \downarrow$	0.074	0.073	0.060	0.063	0.059	0.057	0.036	0.033	0.019	0.018
	$S_\alpha \uparrow$	0.787	0.829	0.821	0.859	0.817	0.830	0.889	0.900	0.907	0.911
	$E_M^\phi \uparrow$	0.820	0.875	0.855	0.885	0.846	0.849	0.915	0.916	0.943	0.949
DIS-TE2	$maxF_1 \uparrow$	0.799	0.822	0.827	0.849	0.829	0.849	0.906	0.910	0.889	0.894
	$M \downarrow$	0.070	0.067	0.059	0.061	0.058	0.055	0.031	0.029	0.029	0.030
	$S_\alpha \uparrow$	0.823	0.865	0.845	0.862	0.843	0.866	0.913	0.921	0.883	0.889
	$E_M^\phi \uparrow$	0.858	0.903	0.889	0.893	0.886	0.894	0.947	0.957	0.928	0.937
DIS-TE3	$maxF_1 \uparrow$	0.830	0.870	0.868	0.911	0.865	0.888	0.920	0.937	0.851	0.853
	$M \downarrow$	0.064	0.063	0.049	0.051	0.050	0.047	0.029	0.027	0.045	0.043
	$S_\alpha \uparrow$	0.836	0.878	0.871	0.892	0.865	0.885	0.918	0.918	0.851	0.854
	$E_M^\phi \uparrow$	0.883	0.926	0.903	0.924	0.913	0.931	0.951	0.957	0.897	0.906
DIS-TE4	$maxF_1 \uparrow$	0.827	0.879	0.846	0.882	0.846	0.851	0.906	0.916	0.763	0.764
	$M \downarrow$	0.072	0.072	0.061	0.063	0.059	0.053	0.038	0.035	0.088	0.084
	$S_\alpha \uparrow$	0.830	0.862	0.852	0.856	0.849	0.873	0.901	0.911	0.799	0.806
	$E_M^\phi \uparrow$	0.870	0.923	0.891	0.935	0.891	0.895	0.933	0.941	0.843	0.850
DIS-VD	$maxF_1 \uparrow$	0.791	0.835	0.823	0.851	0.823	0.847	0.897	0.905	0.842	0.847
	$M \downarrow$	0.074	0.072	0.062	0.065	0.059	0.058	0.036	0.033	0.045	0.044
	$S_\alpha \uparrow$	0.813	0.866	0.843	0.873	0.838	0.857	0.905	0.909	0.848	0.850
	$E_M^\phi \uparrow$	0.856	0.923	0.873	0.880	0.876	0.901	0.931	0.941	0.896	0.903

时取得最佳结果。值得注意的是，MaskFactory 在所有子数据集上均取得了最高的 $maxF_1$ 分数，相比 IS-Net 基线的提升幅度在 0.044 到 0.052 之间。

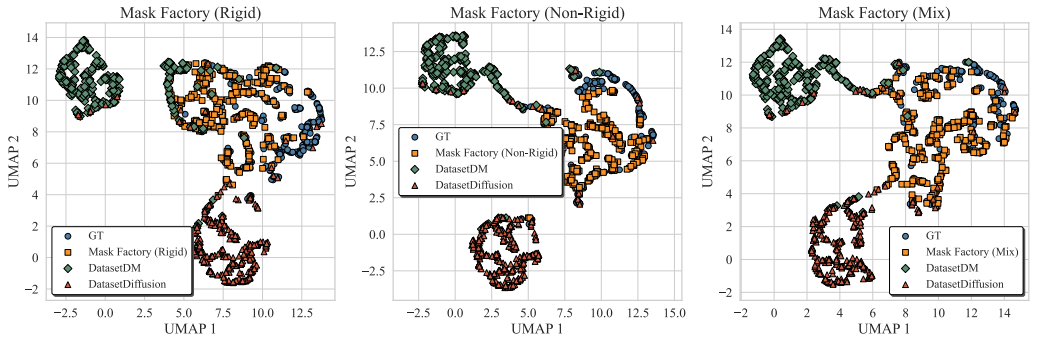
相比之下，尽管 DatasetDM 与 Dataset Diffusion 也表现出一定的性能增益，但当生成数据超过 5000 张图像时，它们出现了性能退化。在这些情况下，分割网络的性能停滞甚至下降。对于 DIS 分割任务，伪标签的使用可能引入额外误差，导致性能下降。

此外，MaskFactory 在加权 F 度量 (F_β^ω)、S 度量 (S_α) 和增强对齐度量 (E_M^ϕ) 方面均表现出更优的性能。与基线相比，MaskFactory 的平均绝对误差 (M) 也一致地更低，表明分割结果更为准确。

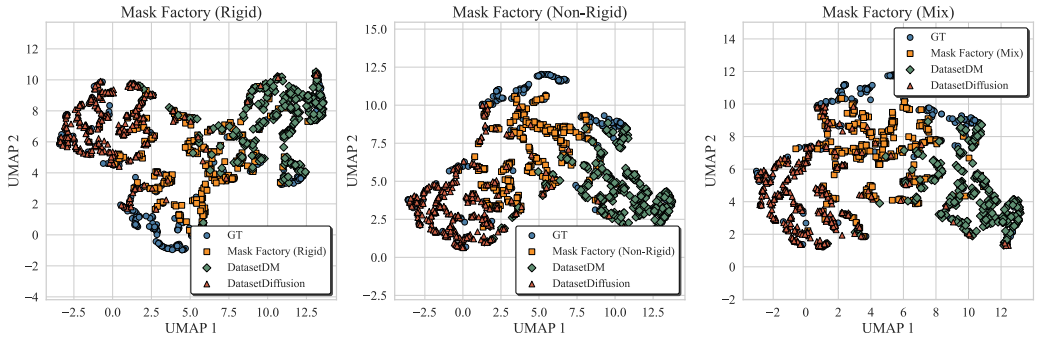
按分割网络划分的结果 在 IS-Net 的性能取得稳定提升之后，我们通过在 DIS5K 数据集上，将相同配置应用于若干最先进的分割网络，考察了本方法的泛化能力。每个网络均使用扩充了 10,000 对生成图像的 DIS-TR 数据集进行训练。本研究考虑的网络包括 FP-DIS [36]、UDUN [25]、BiRefNet [1] 与 SAM-HQ[37]，均采用其默认参数实现。

实验结果如表 2 所示，表明我们提出的方法在 DIS5K 的五个子数据集上一致地提升了所有被评估网络的性能。值得注意的是，在 $maxF_1$ 与 E_M^ϕ 指标上观察到了显著的提升。例如，在 DIS-TE1 子数据集上，我们的方法将 IS-Net 的 $maxF_1$ 分数从 0.740 提升到 0.784，FP-DIS 从 0.784 提升到 0.805，UDUN 从 0.784 提升到 0.799，BiRefNet 从 0.866 提升到 0.882，SAM-HQ 从 0.897 提升到 0.905。

此外，应用我们的方法后，所有网络的平均绝对误差 (M) 均有所下降，表明分割结果更为准确。S 度量 (S_α) 在所有子数据集与网络上也一致提升，凸显了本方法在捕获结构相似性方面的有效性。



(a) 生成的掩码图像的 UMAP 分布差异



(b) 生成的对应 RGB 图像的 UMAP 分布差异

图 4: UMAP 分布差异

在用于验证的 DIS-VD 子数据集上，我们的方法将 IS-Net 的 $maxF_1$ 分数从 0.791 提升到 0.835，FP-DIS 从 0.823 提升到 0.851，UDUN 从 0.823 提升到 0.847，BiRefNet 从 0.897 提升到 0.905，SAM-HQ 从 0.842 提升到 0.847。这些发现凸显了本方法的泛化能力，因为它在不需要任何针对特定网络修改的情况下提升了多种分割网络的性能。

可视化结果 如图 5（附录）所示，我们的方法取得了精确的结果，同时包含刚性与非刚性变换。非刚性变换（如 (b)、(c) 列所示）实现了形状编辑，例如移除桌角或将两个背包合并为一个。相比之下，刚性变换（如 (d)、(e)、(f) 列所示）主要涉及视角变化，展示了原始掩码在三维空间中的旋转。值得注意的是，我们的方法有效地保留了原始图像的拓扑结构信息，包括椅背上的孔洞。这使得能够以低成本生成高精度、多样化的数据对。为研究生成图像与真实图像之间的差异，我们分析了图像与掩码的分布。具体而言，我们使用 CLIP 模型 [38] 从 300 张真实图像以及由 DatasetDM、DatasetDiffusion 与 MaskFactory 生成的图像与掩码中提取特征。然后我们对这 300 个特征应用 UMAP [39] 进行降维。掩码与图像的特征分布分别可视化于图 4(a) 与图 4(b)。

表 3: 与原始数据集的相似度。

生成类型	CLIP	UMAP
DatasetDM [9]	0.6683	0.7017
Dataset Diffusion [10]	0.6217	0.6349
MaskFactory（刚性）	0.8791	0.8961
MaskFactory（非刚性）	**0.9147**	**0.9346**
MaskFactory（全部）	0.8967	0.9103

在掩码编辑领域，与其他生成方法相比，MaskFactory 展现出更优的掩码保真度，这主要归功于拓扑一致性约束的引入。这使得特征分布与真实图像的特征分布高度吻合。相反，对于 RGB 图像，来自扩散 VAE 的先验在生成图像与真实图像分布之间引入了更大的差异。然而，与其他两种方法相比，MaskFactory 生成的分布与真实图像分布具有更大的重叠。

此外，我们使用余弦相似度对差异进行了量化，如表 3 所示。结果表明我们的方法取得了与真实图像最接近的分布，进一步验证了 MaskFactory 在生成逼真掩码与图像方面的有效性。

5 讨论

5.1 消融实验

掩码生成类型消融 在我们的研究中，我们实现了掩码刚性编辑、非刚性编辑与混合编辑——每种都利用了我们为特定应用场景量身定制的新颖掩码控制技术。刚性编辑面向需要精确几何调整的场景，主要聚焦于视角与尺度变换。非刚性编辑面向需要高适应性的应用，处理拓扑一致的形变以及复杂、动态的图像编辑。混合编辑结合了两种方法的优势，提供了全面的解决方案。我们进一步评估了每种编辑策略带来的性能增益。我们的实验结果如表 4 所示。

表 4: 关于生成类型的消融。

类型	$maxF_1 \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_M^\phi \uparrow$
刚性	0.768	0.074	0.807	0.867
非刚性	0.771	0.074	0.796	0.858
混合	0.784	0.073	0.829	0.875

从表中可以观察到，混合编辑在大多数指标上取得了最高的性能。具体而言，它取得了最高的 $maxF_1$ 分数与 S_α，表明其具有更优的结构保真度与分割质量。M 与 E_M^ϕ 的轻微提升进一步凸显了混合编辑在创建多样且逼真的图像-掩码组合方面的多功能性与有效性。

损失函数消融 我们在模型中引入了内容损失与结构损失。判别损失通过判别器实现，评估生成图像与真实图像之间的差异，旨在提升输出的真实感与质量。边缘约束损失聚焦于在图像编辑过程中保持边缘连贯性，这对于保留细节结构信息至关重要。我们进行了消融实验以评估每种损失函数的影响。实验结果如表 5 所示。

表 5: 关于损失函数的消融。

损失函数			$maxF_1 \uparrow$	$M \downarrow$
$\mathcal{L}_{\text{GAN}}$	$\mathcal{L}_{\text{content}}$	$\mathcal{L}_{\text{structure}}$		
✓			0.778	0.073
	✓		0.745	0.075
		✓	0.751	0.074
✓	✓	✓	0.784	0.073

从表中可以观察到，三种损失函数的组合 ($\mathcal{L}_{\text{GAN}}$, $\mathcal{L}_{\text{content}}$ 和 $\mathcal{L}_{\text{structure}}$) 取得了最高的 $maxF_1$ 分数，表明其在结构保真度与真实感方面表现最佳。

5.2 局限性

尽管我们的方法取得了良好的结果，但它仍然面临一些显著问题。虽然我们尝试了不同的条件，结果如表 6 所示。ControlNet 有时会生成前景-背景区分生硬的不自然图像，需要额外的图像协调处理。复杂场景可能产生不真实的元素，例如位置不当的物体。此外，我们的方法依赖于预先标注的图像-掩码对，这限制了其自主生成数据的能力，并且需要高质量的初始标注。

表 6: 关于条件的消融。

掩码	提示词	Canny	$maxF_1 \uparrow$	$M \downarrow$
✓			0.778	0.075
✓	✓		0.782	0.073
✓		✓	0.764	0.080
✓	✓	✓	0.784	0.073

5.3 结论

本文提出了 MaskFactory，这是一种新颖的两阶段方法，用于为 DIS 任务生成高质量的合成数据集。通过结合刚性与非刚性掩码编辑技术，并使用多条件控制进行图像生成，MaskFactory 生成了多样且精确的合成图像-掩码对，显著减少了数据集准备的时间与成本。在 DIS5K 基准上的实验表明，与现有方法相比，MaskFactory 在质量与效率方面均具有更优的性能。

致谢

我们感谢 Yuyan Huang、刘弘与季文俊在本项目过程中富有成效的讨论。钱浩天与金小刚受浙江省重点研发计划（No.2024C01069）资助。范登平受国家自然科学基金（No.62476143）资助。

参考文献

[1] Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. Bilateral reference for high-resolution dichotomous image segmentation. *CAAI AIR*, 3:9150038, 2024.

[2] Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, and Luc Van Gool. Highly accurate dichotomous image segmentation. In *ECCV*, 2022.

[3] Lian Liu, Han Zhou, Jiongquan Chen, Sijing Liu, Wenlong Shi, Dong Ni, Deng-Ping Fan, and Xin Yang. Instructive feature enhancement for dichotomous medical image segmentation. In *MICCAI*, 2023.

[4] Yinda Chen, Che Liu, Xiaoyu Liu, Rossella Arcucci, and Zhiwei Xiong. Bimcv-r: A landmark dataset for 3d ct text-image retrieval. In *MICCAI*, 2024.

[5] Yinda Chen, Wei Huang, Xiaoyu Liu, Shiyu Deng, Qi Chen, and Zhiwei Xiong. Learning multiscale consistency for self-supervised electron microscopy instance segmentation. In *ICASSP*, 2024.

[6] Chenfeng Xu, Bichen Wu, Zining Wang, Wei Zhan, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. Squeezesegv3: Spatially-adaptive convolution for efficient point-cloud segmentation. In *ECCV*, 2020.

[7] Xiaoyu Liu, Miaomiao Cai, Yinda Chen, Yueyi Zhang, Te Shi, Ruobing Zhang, Xuejin Chen, and Zhiwei Xiong. Cross-dimension affinity distillation for 3d em neuron segmentation. In *CVPR*, 2024.

[8] Yinda Chen, Wei Huang, Shenglong Zhou, Qi Chen, and Zhiwei Xiong. Self-supervised neuron segmentation with multi-agent reinforcement learning. In *IJCAI*, 2023.

[9] Weijia Wu, Yuzhong Zhao, Hao Chen, Yuchao Gu, Rui Zhao, Yefei He, Hong Zhou, Mike Zheng Shou, and Chunhua Shen. Datasetdm: Synthesizing data with perception annotations using diffusion models. *NeurIPS*, 2023.

[10] Quang Nguyen, Truong Vu, Anh Tran, and Khoi Nguyen. Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. *NeurIPS*, 2023.

[11] Yinda Chen, Che Liu, Wei Huang, Sibo Cheng, Rossella Arcucci, and Zhiwei Xiong. Generative text-guided 3d vision-language pretraining for unified medical image segmentation. *arXiv preprint arXiv:2306.04811*, 2023.

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[13] Aysim Toker, Marvin Eisenberger, Daniel Cremers, and Laura Leal-Taixé. Satsynth: Augmenting image-mask pairs through diffusion models for aerial semantic segmentation. In *CVPR*, 2024.

[14] Lihe Yang, Xiaogang Xu, Bingyi Kang, Yinghuan Shi, and Hengshuang Zhao. Freemask: Synthetic images with dense annotations make stronger segmentation models. *NeurIPS*, 2023.

[15] Yinda Chen, Haoyuan Shi, Xiaoyu Liu, Te Shi, Ruobing Zhang, Dong Liu, Zhiwei Xiong, and Feng Wu. Tokenunify: Scalable autoregressive visual pre-training with mixture token prediction. *arXiv preprint arXiv:2405.16847*, 2024.

[16] Dingquan Wang and Jason Eisner. Synthetic data made to order: The case of parsing. In *EMNLP*, 2018.

[17] Shaobo Lin, Kun Wang, Xingyu Zeng, and Rui Zhao. Explore the power of synthetic data on few-shot object detection. In *CVPR*, 2023.

[18] Alvaro Figueira and Bruno Vaz. Survey on synthetic data generation, evaluation methods and gans. *Math.*, 10(15):2733, 2022.

[19] Aman Kishore, Tae Eun Choe, Junghyun Kwon, Minwoo Park, Pengfei Hao, and Akshita Mittel. Synthetic data generation using imitation training. In *CVPR*, 2021.

[20] Tung Nguyen, Sudhanshu Agrawal, and Aditya Grover. Expt: Synthetic pretraining for few-shot experimental design. *NeurIPS*, 2023.

[21] Yang Liu, Sujay Khandagale, Colin White, and Willie Neiswanger. Synthetic benchmarks for scientific research in explainable machine learning. In *NeurIPS Dataset Track*, 2021.

[22] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[23] Qi Chen, Xiaoxi Chen, Haorui Song, Zhiwei Xiong, Alan Yuille, Chen Wei, and Zongwei Zhou. Towards generalizable tumor synthesis. *CVPR*, 2024.

[24] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023.

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

[26] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *CVPR*, 2019.

[27] Qian Yu, Xiaoqi Zhao, Youwei Pang, Lihe Zhang, and Huchuan Lu. Multi-view aggregation network for dichotomous image segmentation. In *CVPR*, 2024.

[28] Qin Liu, Jaemin Cho, Mohit Bansal, and Marc Niethammer. Rethinking interactive image segmentation with low latency high quality and diverse prompts. In *CVPR*, 2024.

[29] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *CVPR*, 2023.

[30] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *CVPR*, 2023.

[31] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *CVPR*, 2009.

[32] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *CVPR*, 2014.

[33] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, 2012.

[34] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *ICCV*, 2017.

[35] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *IJCAI*, 2018.

[36] Yan Zhou, Bo Dong, Yuanfeng Wu, Wentao Zhu, Geng Chen, and Yanning Zhang. Dichotomous image segmentation with frequency priors. In *IJCAI*, 2023.

[37] Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. *NeurIPS*, 36, 2023.

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[39] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[40] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

[41] Fei Shen and Jinhui Tang. Imagpose: A unified conditional framework for pose-guided person generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[42] Fei Shen, Hu Ye, Jun Zhang, Cong Wang, Xiao Han, and Yang Wei. Advancing pose-guided image synthesis with progressive conditional diffusion models. In *The Twelfth International Conference on Learning Representations*.

[43] Ling Li, Junliang Xing, Xinchun Yu, and Xiao-Ping Zhang. Deviation wing loss for high-performance 2d pose estimation. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024.

[44] Ling Li, WenRui Yang, Xinchun Yu, Junliang Xing, and Xiao-Ping Zhang. Translating motion to notation: Hand labanotation for intuitive and comprehensive hand movement documentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4092–4100, 2024.

[45] Haopeng Sun. Ultra-high resolution segmentation via boundary-enhanced patch-merging transformer, 2024.

[46] Haopeng Sun, Lumin Xu, Sheng Jin, Ping Luo, Chen Qian, and Wentao Liu. Program: Prototype graph model based pseudo-label learning for test-time adaptation. In *The Twelfth International Conference on Learning Representations*, 2024.

[47] Jianjian Yin, Shuai Yan, Tao Chen, Yi Chen, and Yazhou Yao. Class probability space regularization for semi-supervised semantic segmentation. *Computer Vision and Image Understanding*, page 104146, 2024.

[48] Xiaowen Ma, Rongrong Lian, Zhenkai Wu, Hongbo Guo, Mengting Ma, Sensen Wu, Zhenhong Du, Siyang Song, and Wei Zhang. Logcan++: Adaptive local-global class-aware network for semantic segmentation of remote sensing imagery, 2024.

[49] Xiaowen Ma, Mengting Ma, Chenlu Hu, Zhiyuan Song, Ziyan Zhao, Tian Feng, and Wei Zhang. Log-can: Local-global class-aware network for semantic segmentation of remote sensing images. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[50] Hailin Tao, Jinjiang Li, Zhen Hua, and Fan Zhang. Dudb: Deep unfolding based dual-branch feature fusion network for pan-sharpening remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[51] Dayu Hu, Suyuan Liu, Jun Wang, Junpu Zhang, Siwei Wang, Xingchen Hu, Xinzhong Zhu, Chang Tang, and Xinwang Liu. Reliable attribute-missing multi-view clustering with instance-level and feature-level cooperative imputation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1456–1466, 2024.

[52] Dayu Hu, Zhibin Dong, Ke Liang, Hao Yu, Siwei Wang, and Xinwang Liu. High-order topology for deep single-cell multi-view fuzzy clustering. *IEEE Transactions on Fuzzy Systems*, 2024.

[53] Fan Zhang, Gongguan Chen, Hua Wang, and Caiming Zhang. Cf-dan: Facial-expression recognition based on cross-fusion dual-attention network. *Computational Visual Media*, pages 1–16, 2024.

[54] Fan Zhang, Gongguan Chen, Hua Wang, Jinjiang Li, and Caiming Zhang. Multi-scale video super-resolution transformer with polynomial approximation. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):4496–4506, 2023.

[55] Dapeng Zhang, Peng Zhi, Binbin Yong, Jin-Qiang Wang, Yufeng Hou, Lan Guo, Qingguo Zhou, and Rui Zhou. Ehss: An efficient hybrid-supervised symmetric stereo matching network. *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1044–1051, 2023.

[56] Dapeng Zhang, Dayu Chen, Peng Zhi, Yinda Chen, Zhenlong Yuan, Chenyang Li, Sunjing, Rui Zhou, and Qingguo Zhou. Mapexpert: Online hd map construction with simple and efficient sparse map element expert, 2024.

[57] Zhenlong Yuan, Cong Liu, Fei Shen, Zhaoxin Li, Jinguo Luo, Tianlu Mao, and Zhaoqi Wang. MSP-MVS: Multi-Granularity Segmentation Prior Guided Multi-View Stereo, December 2024.

[58] Zhenlong Yuan, Jinguo Luo, Fei Shen, Zhaoxin Li, Cong Liu, Tianlu Mao, and Zhaoqi Wang. DVP-MVS: Synergize Depth-Edge and Visibility Prior for Multi-View Stereo, December 2024.

[59] Zhenlong Yuan, Jiakai Cao, Zhaoxin Li, Hao Jiang, and Zhaoqi Wang. SD-MVS: Segmentation-Driven Deformation Multi-View Stereo with Spherical Refinement and EM Optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6871–6880, 2024.

附录

A MaskFactory 算法的伪代码

MaskFactory 是一种用于为 DIS 任务生成高质量合成数据集的两阶段方法。在第一阶段，现有的真值掩码经过刚性与非刚性编辑来生成多样的合成掩码。刚性编辑利用扩散模型中的几何先验进行精确的视角变换，而非刚性编辑采用对抗训练与自注意力机制，在保持拓扑的同时进行复杂的形状修改。

在第二阶段，生成的掩码及其对应的 Canny 边缘作为条件，与类别提示词一起使用多条件可控生成方法引导高分辨率 RGB 图像的生成。该过程确保了生成图像与掩码之间的一致性，同时提升了数据集的真实感与多样性。我们的伪代码见算法 1。

算法 1: MaskFactory 算法

输入: $D = \{(I_i^r, I_i^m)\}_{i=1}^N$ - 原始数据集
输出: $G = \{(g_i^r, g_i^m)\}_{i=1}^M$ - 合成数据集

阶段一：掩码编辑
对于 $i \leftarrow 1$ **至** N **执行**
  步骤 1.1：刚性掩码编辑
  $I_i^{m'} \leftarrow \text{Invert}(I_i^m)$ // 对源掩码取反
  $g_i^m \leftarrow \psi_\theta(I_i^{m'}, T_i)$ // 应用视角变换
  步骤 1.2：非刚性掩码编辑
  $E_s \leftarrow E(I_i^m)$ // 从源掩码提取边缘图
  $\mathcal{V} \leftarrow \{v_j\}_{j=1}^{N_v}$ // 从边缘图获取关键点
  $\mathcal{T}_s \leftarrow (V, E_s)$ // 构建源掩码的结构图
  $g_i^m \leftarrow G_\theta(z, P_i, I_i^m)$ // 生成合成掩码
  $T_g = (\mathcal{V}, \mathcal{E}_s)$ // 构建合成掩码的结构图
  $\mathcal{L}_{GAN} \leftarrow \mathbb{E}_{T_s \sim p_{data}(T_s)}[\log D_\phi(T_s)] + \mathbb{E}_{T_g \sim p_{gen}(T_g)}[\log(1 - D_\phi(T_g))]$
  $\mathcal{L}_{content} \leftarrow \|g_i^m - I_i^m\|_1$
  $\mathcal{L}_{structure} \leftarrow \|T_g - T_s\|_1$
  $\mathcal{L}_{total} \leftarrow \mathcal{L}_{GAN} + \lambda_1 \mathcal{L}_{content} + \lambda_2 \mathcal{L}_{structure}$
  更新 G_θ 和 D_ϕ，以最小化 $\mathcal{L}_{total}$
阶段二：图像生成
对于 $i \leftarrow 1$ **至** M **执行**
  $c_i^s \leftarrow g_i^m$ // 分割条件
  $c_i^y \leftarrow \text{Canny}(g_i^m)$ // Canny 条件
  $z \sim \mathcal{N}(0, 1)$ // 采样高斯噪声
  $g_i^r \leftarrow M_\theta(z, B_\theta(c_i^s), B_\theta(c_i^y))$ // 生成 RGB 图像
返回 $G = \{(g_i^r, g_i^m)\}_{i=1}^M$

B 数据集详情

我们在 DIS5K 数据集上开展实验，该数据集包含 5,479 张高分辨率图像，涵盖各种背景中的伪装物体、显著物体或精细物体。DIS5K 数据集划分为三个子集：用于训练的 DIS-TR（3,000 张图像）、用于验证的 DIS-VD（470 张图像）以及用于测试的 DIS-TE（2,000 张图像）。对于数据增强，我们利用训练子集（DIS-TR）中的掩码部分。

为评估我们的模型，我们采用一组多样化的指标以确保全面的性能评估。这些指标包括：max F_1，它在精确率与召回率之间取得平衡，给出反映整体准确性的调和平均值；F_β^ω，一种补偿类别不平衡的加权 F 度量，取值范围为 0 到 1，值越高表示性能越优；M（平均绝对误差），计算预测掩码与真值掩码之间的平均绝对差异，值越低表示准确性越好；S_α，一种结构相似性度量，评估图像中重要结构的保持情况，值越接近 1 表示性能越好；以及 E_M^ϕ，一种同时考虑像素级与图像级信息以进行更全面评估的增强度量，值越高表示性能越好。总的来说，这些指标为评估我们分割模型的有效性与可靠性提供了稳健的框架。

C 数学细节

C.1 基于扩散的图像生成

MaskFactory 中用于图像生成的扩散模型遵循 Ho 等人 [40] 提出的公式。给定数据分布 $x_0 \sim q(x_0)$，前向扩散过程定义为逐渐向数据添加高斯噪声的马尔可夫链：

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}), \tag{9}$$

其中 $\beta_t \in (0,1)$ 为方差调度。反向过程由神经网络 ϵ_θ 学习，该网络预测每一步添加的噪声：

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t,t), \sigma_t^2\mathbf{I}), \tag{10}$$

其中 $\mu_\theta(x_t,t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t,t)\right)$，$\alpha_t = 1-\beta_t$，$\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$。
目标是最大化 **变分下界**：

$$\mathcal{L} = \mathbb{E}_{q(x_0)}\mathbb{E}_{q(x_1,\ldots,x_T|x_0)}\left[\sum_{t=1}^T \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t,x_0)}\right]. \tag{11}$$

C.2 拓扑保持对抗训练

MaskFactory 中的**拓扑保持对抗训练**涉及生成器 G_θ 和判别器 D_ϕ。生成器旨在最小化对抗损失：

$$\mathcal{L}_{\text{GAN}}(G) = \mathbb{E}_{T_g \sim p_{\text{gen}}(T_g)}[\log(1 - D_\phi(T_g))], \tag{12}$$

而判别器试图最大化对抗损失：

$$\mathcal{L}_{\text{GAN}}(D) = \mathbb{E}_{T_s \sim p_{\text{data}}(T_s)}[\log D_\phi(T_s)] + \mathbb{E}_{T_g \sim p_{\text{gen}}(T_g)}[\log(1 - D_\phi(T_g))]. \tag{13}$$

生成器与 **判别器**交替更新以达到平衡。

C.3 结构图构建

为在编辑过程中保持源掩码的拓扑结构，**MaskFactory** 分别为源掩码与合成掩码构建结构图 T_s 与 T_g，结构图 $T = (V, E)$ 由表示关键点的顶点集合 $V = \{v_j\}_{j=1}^{N_v}$ 与表示其连接关系的边集合 E 组成。

结构保持损失定义为结构图之间的 L1 距离：

$$\mathcal{L}_{\text{structure}} = \|T_g - T_s\|_1 = \sum_{(i,j)\in E} \|T_g(i,j) - T_s(i,j)\|, \tag{14}$$

其中 $T_g(i,j)$ 与 $T_s(i,j)$ 分别表示合成掩码与源掩码结构图中顶点 i 与 j 之间的边权重。

通过将**结构保持损失**与对抗损失、内容损失一同最小化，**MaskFactory** 确保编辑后的掩码保持源掩码的拓扑结构。

D 使用 MaskFactory 两种编辑方法的结果可视化

在本节中，我们展示 MaskFactory 对常见物体与细粒度物体进行掩码编辑的能力。

D.1 常见物体掩码编辑的可视化

我们选取了桌子、椅子、包和乐器等常见家居物品，使用 MaskFactory 进行编辑。非刚性编辑使用不同的提示词进行，刚性编辑从不同视角进行。编辑结果及对应的 RGB 图像生成如图 5 所示。MaskFactory 展现出强大的拓扑结构保持能力，并产生多样的编辑结果，这从对刚性与非刚性物体所做的多种修改中得到印证。

D.2 细粒度物体掩码编辑的可视化

我们从 MaskFactory 生成的掩码中选取了细粒度物体，例如华丽的欧式吊灯、铁门、鸟笼和海马，以展示 MaskFactory 的精细编辑能力。即使面对复杂的几何结构，MaskFactory 也能够在不丢失原始掩码语义信息的情况下进行拓扑保持编辑。这些复杂结构在分割指标中起着至关重要的作用。如图 6 所示，可视化结果表明模型能够在保持原始掩码结构完整性的同时处理细粒度细节。

D.3 带 Canny 约束的可视化结果

在引入 Canny 边缘检测作为约束后，MaskFactory 的可视化结果在边界精度上表现出显著提升。Canny 边缘有效地引导了生成过程，使图像具有更清晰、更准确的边界细节，避免了模糊或不明确的过渡区域。可视化结果表明，Canny 边缘不仅更好地约束了生成图像的轮廓，还提升了输出的整体保真度与视觉质量。如图 7 所示，与没有边缘约束的模型相比，我们的方法生成了更细致、结构更连贯的图像。

D.4 拓扑结构可视化

我们对选定样本进行了拓扑结构可视化，以评估模型在编辑过程中保持拓扑的能力。这些可视化清晰地展示了我们的方法如何在必要时允许有效操控的同时保留原始数据的拓扑结构。如图 8 所示，无论是处理复杂的几何形状还是细微的结构修改，拓扑可视化都表明我们的方法在整个编辑过程中可靠地保持了几何一致性与拓扑特征。

E 未来工作

目前，我们的工作仅探索了生成方法 MaskFactory 在 DIS 分割任务中的性能，这可以视为一种数据增强形式。然而，该方法实际上具有更广泛应用于更多种类任务的潜力，例如姿态估计 [41, 42, 43, 44]、超高分辨率图像分割 [45, 46, 47]、遥感图像增强 [48, 49, 50]、图像拓扑增强 [51, 52]、图像超分辨率 [53, 54]、自动驾驶 [55, 56] 以及多视图估计 [57, 58, 59]。此外，基于生成的方法不可避免地会引入一些噪声数据。因此，如何筛选生成的数据，或研究如何从噪声数据中学习并评估生成数据的质量，也是一个值得关注的问题。

(a) 原始图像 (b) 刚性 1 (c) 刚性 2 (d) 非刚性 1 (e) 非刚性 2 (f) 非刚性 3

图 5: 常见物体掩码编辑的可视化结果。模型在刚性与非刚性编辑中均展现出强大的拓扑结构保持能力与多样的编辑结果。

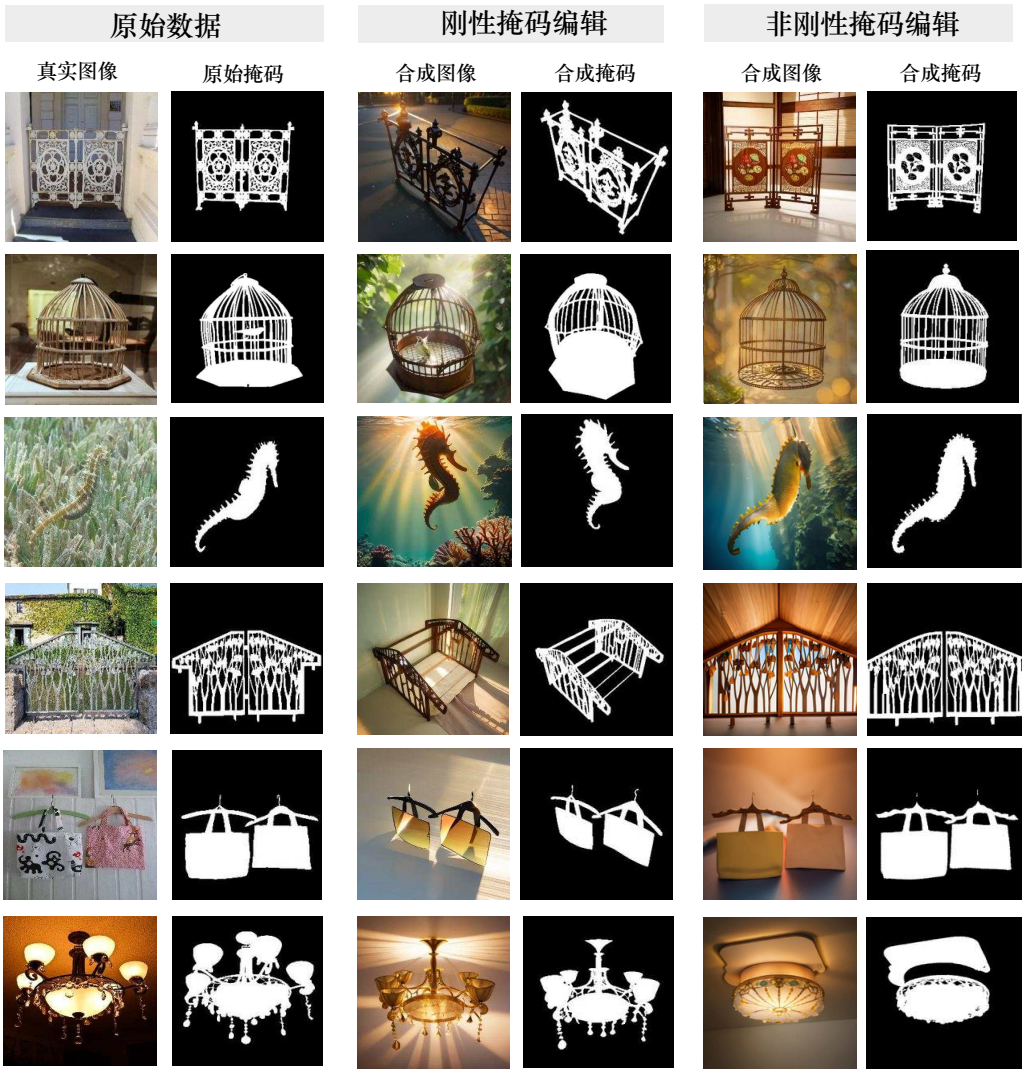


图 6: 细粒度物体掩码编辑的可视化结果。模型成功地编辑了复杂结构，而不损害原始掩码的语义信息。

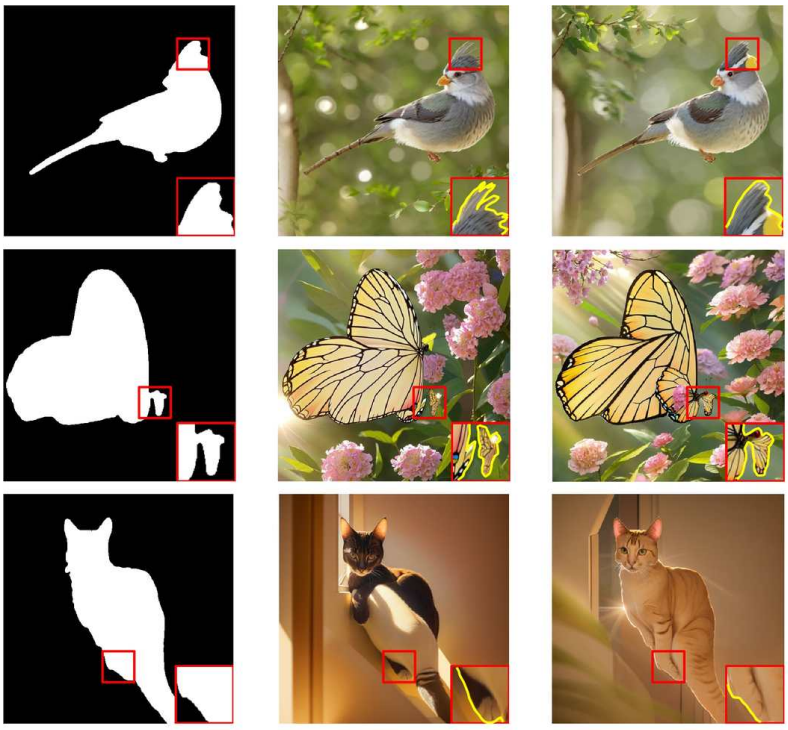


图 7: Canny 条件的可视化结果。生成的图像表现出更高的边界精度与更好的结构连贯性。

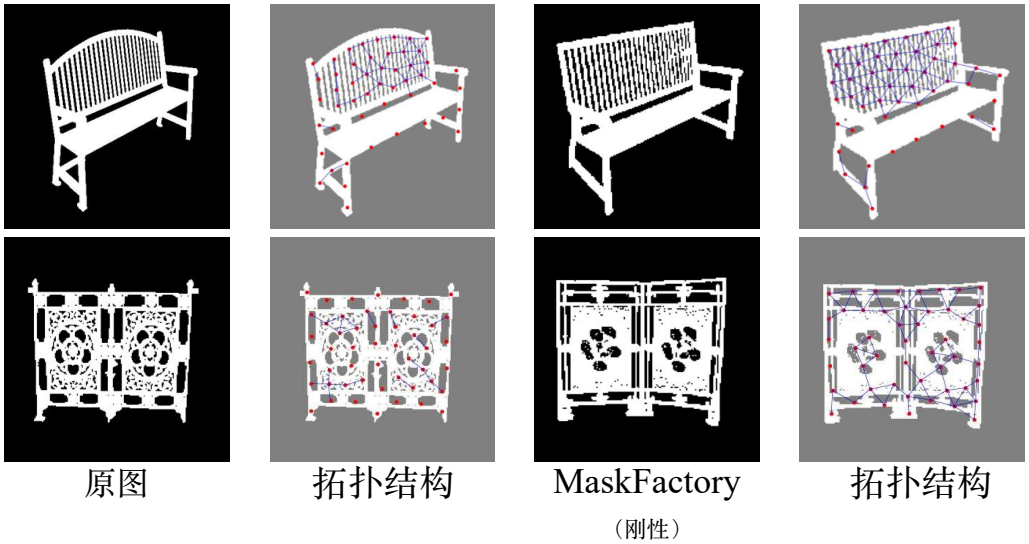


图 8: 拓扑结构可视化。可视化结果展示了模型在编辑过程中保持与操控拓扑的能力。