arXiv:2603.19218v1 [cs.CV] 19 Mar 2026

重新思考生成式分割的向量场学习

Chaoyang Wang[1,2] **Yaobo Liang**[1] **Boci Peng**[1] **Fan Duan**[2]
JingdongWang[2] **YunhaiTong**[1]北京大学 [2]百度 cywang@stu.pku.edu.cn

摘要

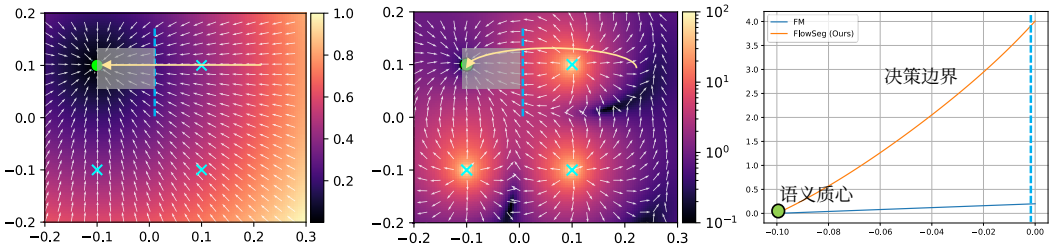
如何驾驭扩散模型以实现生成式分割，正受到越来越多的关注。现有方法主要侧重于架构调整或训练启发式方法，但对于连续流匹配目标与离散感知任务之间内在的不匹配，人们仍缺乏充分认识。在本研究中，我们从向量场学习的角度重新审视扩散分割。我们发现，常用流匹配目标存在两个关键局限：梯度消失和轨迹穿越，导致收敛缓慢且类别分离效果不佳。为解决这些问题，我们提出一种有原则的向量场重塑策略，通过添加一个分离式距离感知校正项来增强学习到的速度场。该校正项引入吸引和排斥相互作用，在保留原有扩散训练框架的同时，增强质心附近的梯度幅值。此外，我们设计了一种受克罗内克序列启发、计算效率高的拟随机类别编码方案，并将其无缝整合到端到端像素神经场框架中，以实现像素级语义对齐。大量实验一致表明，相较于原始流匹配方法，该方法取得了显著提升，大幅缩小了生成式分割与强判别式专用模型之间的性能差距。

1 引言

扩散模型 [34, 35, 16] 已成为视觉生成领域的主流方法，在复杂数据分布建模方面表现出色。它们在高质量图像合成 [32, 30, 12],视频生成 [4, 1, 15] 和图像编辑 [5] 方面的成效，推动了其向各种下游任务扩展。尽管最初是为视觉生成而提出的，近期研究 [44, 38, 37] 已开始将扩散框架应用于感知任务，旨在通过统一范式实现生成与理解。分割任务在这些任务中独具特点，因为它为每个像素分配一个离散语义标签 [29, 8, 33, 48, 11, 10], , 生成的是类别而非连续输出。这带来了一个关键挑战：扩散模型在连续空间中运行，而分割需要本质上离散的预测。这一差距给任务构建和优化带来了重大挑战。

为了将扩散模型应用于分割并利用其预训练先验，现有方法主要分为三类，如图 2 所示。1）将扩散模型用作特征提取器，并搭配分割头 [49, 42, 20, 17, 36, 3, 18]。2）采用视觉基础特征，并添加轻量级扩散模块来优化掩码 [39, 14, 13]。3）将分割视为端到端图像转换，并将掩码编码为三通道伪彩色图像 [2, 21, 37, 38, 31, 24]。有趣的是，除了模型设计之外，语义分割对于基于扩散的模型而言仍是一项颇具挑战且研究不足的任务，这主要源于两个因素：1）内在的

预印本。审稿中。



(a) 原始流匹配 $\nabla_x L \propto (x-\mu)$ (b) FlowSeg（我们的方法）$\nabla_x L \propto (x-\mu+\nabla\Phi)$ (c) 梯度范数

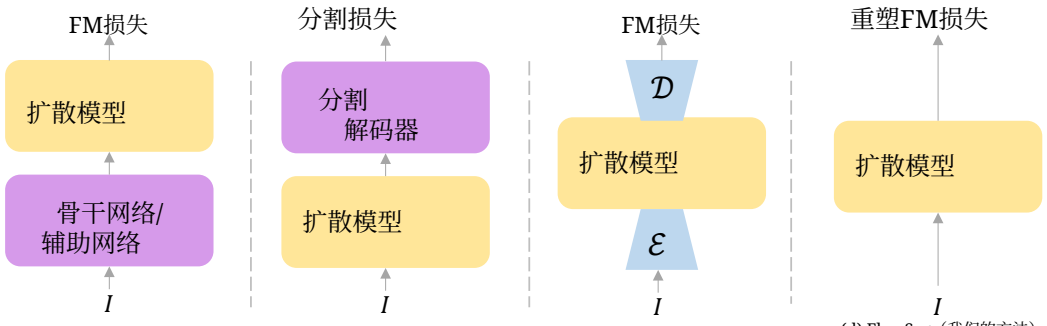
图 1：**生成式分割（x 预测）中梯度消失与轨迹穿越的可视化**。(a) 原始流匹配在语义质心 μ附近会出现梯度消失，导致轨迹缓慢且缺乏判别性，可能穿越竞争质心的邻近区域并产生错误预测。(b) FlowSeg（我们的方法）引入势函数 Φ，以增强目标质心周围的梯度，并施加远离非目标质心的排斥力，从而实现更快的收敛，并产生判别性更强、偏转更明显的轨迹。$x-\mu$项可维持轨迹在外边界附近的收敛。(c) 从质心到决策边界（(a)、(b)中的灰色区域）的梯度范数：我们的方法能够维持较强的梯度，而原始流匹配的梯度几乎消失。黄色曲线和蓝色虚线分别表示预测轨迹和决策边界；绿色圆点表示目标质心，蓝色叉号表示无关类别。

生成模型的随机性与语义分割的确定性要求不匹配；2）通常涉及大量类别。近期研究 [38, 21] 通过引入确定性流来应对第一个挑战，但类别数过多的问题依然存在。为使问题易于处理，现有方法往往会简化任务设定。例如，一些方法通过合并相似类别来缩小类别空间 [43], ，或将任务重新表述为一系列指代表达图像分割问题 [14] ，以将其拆解为更简单的子问题。这些发现表明，随机性和类别数过多是扩散分割面临的根本障碍，这说明核心难点并非模型容量，而是框架固有的优化动力学。

本文从优化动力学的角度分析了这一问题，并识别出两个关键问题： 梯度消失 和 轨迹穿越，如图 1 所示。通过考察标准流匹配所诱导的梯度 [25, 26], 我们观察到，优化信号会直接受到预测结果与语义质心之间距离的缩放。由于类别质心位于有界的低维空间中，梯度本身受到限制，并且会随着预测结果接近目标而迅速衰减，最终在接近收敛时消失，导致优化停滞。此外，基于回归的方法只能提供指向真实类别的吸引力，却缺乏来自竞争质心的显式排斥力。弱梯度与缺失的类别间排斥相结合，导致收敛速度缓慢、语义分离效果较差，并最终降低分割性能。

为克服这些局限，我们通过向原始速度中引入距离感知校正项来重塑向量场，使其同时具备吸引和排斥相互作用，并在质心附近维持非消失梯度。这增强了类别分离并加速了收敛。此外，我们提出了一种受克罗内克序列启发的准随机质心编码策略，该策略具有较高的计算效率。结合端到端像素神经场框架，我们的设计使扩散模型与像素级任务相对齐。

我们的主要贡献总结如下：1）我们分析了扩散分割收敛缓慢、性能欠佳的原因，并将其归因于原始流匹配目标造成的梯度消失和轨迹穿越。2）我们提出了一种新颖的向量场重塑策略，通过在目标速度中加入校正项来保持梯度幅值，并引入显式排斥力以改善语义分离。3）我们采用像素神经场框架进行端到端训练，并提出一种可扩展的准随机质心编码方案，无需额外优化即可确保类间几何结构均衡。4）大量实验表明，该方法取得了稳定且显著的提升，大幅缩小了生成式方法与强判别式方法之间的差距。



(a) 掩码细化器 (b) 特征提取器 (c) 图像转换器 (d) FlowSeg（我们的方法）

图 2：**扩散分割范式对比**。(a) 扩散模型主要用于掩码细化，并依赖外部骨干网络进行特征提取，或依赖辅助网络进行粗粒度分割。(b) 扩散模型充当骨干网络，后接专用分割头。(c) 分割任务被表述为图像到掩码转换，无需辅助网络，但仍依赖预训练 VAE。(d) FlowSeg（我们的方法）进行像素级端到端训练，无需额外的辅助模块，并通过重塑底层向量场来改进原始流匹配，从而实现更好的优化。为简洁起见，省略了噪声。

2 相关工作

2.1 扩散与流

扩散模型 [16] 和流模型 [25, 27] 将数据生成参数化为时变动力系统，将简单的参考分布传输至目标数据分布。它们能够在多种领域实现高保真合成，包括图像生成 [32, 30, 12], 视频生成 [4, 1, 15], 图像编辑 [5], 等。早期的扩散模型通常使用随机微分方程（SDE），在预定义的高斯扰动过程中学习分数或噪声估计。相比之下，流模型 [27, 28] 直接学习定义直线概率路径的确定性速度场，从而实现更高效的采样，并在分布传输方面提供更大的灵活性。

为降低计算开销，传统方法 [32] 主要在由变分自编码器（VAE）压缩的潜在空间中运行 [19]。然而，VAE压缩固有的信息损失往往会妨碍需要像素级精度和直接空间对齐的任务。为解决这一问题，近期出现了像素扩散模型 [46, 22, 40]，它们利用大规模分块解码，在保持可扩展性的同时避免牺牲细粒度细节。尽管参数化方式不同，这些方法遵循一个共同原则：生成建模归结为学习一个连续向量场，通过基于回归的目标函数对齐不同分布之间的概率轨迹；而这给扩展到离散感知任务带来了挑战。

2.2 生成式分割

视觉分割旨在将图像划分为语义一致区域，并赋予离散类别标签。传统方法 [29, 8, 33, 48, 11, 10] 通常将分割视为判别式密集分类任务，而生成式方法凭借其在不确定性建模和结构化预测方面的优势，近年来日益受到关注 [2, 23]。

现有的扩散分割方法通常遵循三种范式：1）基于特征的方法，将扩散骨干网络用作强大的冻结提取器，为任务特定分割头提供特征 [49, 42,20, 17, 36, 3, 18]；2）基于细化的方法，利用扩散模块迭代优化由基础模型生成的掩码 [39, 14, 13]；3）统一生成式方法，将掩码预测视为端到端图像转换过程 [2, 21, 37, 38, 31, 24]。
在统一生成式表述中，分割掩码通常会嵌入低维连续流形，以适应生成动态过程的连续特性。尽管这种嵌入对于生成建模至关重要，但仅采用原始的基于回归的流目标，无法考虑语义标签的离散特性。这会导致优化

不匹配：向量场被学习为通用密度估计器，忽略了分割中固有的类别约束。

3 优化动力学分析

本节将正式分析，为什么在生成任务的流匹配中常用的标准均方误差（MSE）目标，在用于生成式分割时会导致次优的优化动力学和性能瓶颈。

3.1 预备知识

令 $I \in \mathbb{R}^{H\times W\times 3}$ 表示一幅输入图像。我们将分割任务定义为源分布 $p_0(x_0|I)$ 与目标分布 $p_1(x_1)$ 之间的流。为消除高斯噪声的随机性并利用分割的确定性，我们将 x_0 定义为特征依赖分布（例如输入图像本身或编码后的图像特征），并将 x_1 定义为目标掩码表示。

在类别式或多通道分割场景中，每个像素或区域都映射到有界连续空间中的一个特定目标质心（例如 $\mathcal{C} \subset [-1,1]^d$）。令 $\mathcal{P} = \{\mu_1, \mu_2, \ldots, \mu_N\}$为表示 N 个不同类别或实例的目标质心集合。在整流流 [27],中，概率路径 x_t 构造如下：

$$x_t = (1-t)x_0 + tx_1, \quad t \in [0,1] \tag{1}$$

相应的真实速度（向量场）为：

$$v_{gt}(x_t) = \frac{dx_t}{dt} = x_1 - x_0 \tag{2}$$

标准训练目标是学习由神经网络参数化的速度场 $v_\theta(x_t,\ t,\ I)$，该速度场通过均方误差损失逼近真实值：

$$\mathcal{L}_{FM} = \mathbb{E}_{t,x_0,x_1} \left\| v_\theta(x_t,t,I) - v_{gt} \right\|^2 \tag{3}$$

3.2 梯度分析

为理解优化挑战，我们分析 $\mathcal{L}_{FM}$ 相对于网络预测的梯度。速度 v_θ 隐式定义了一个预测目标 $\hat{x}_1$，其可估计为：

$$\hat{x}_1 = x_t + (1-t)v_\theta \tag{4}$$

将 v_θ 和 v_{gt} 代入损失函数，我们可以将训练目标改写为预测状态与目标质心之间距离的形式：

$$\mathcal{L}_{FM} \propto \left\| \frac{\hat{x}_1 - x_t}{1-t} - \frac{x_1 - x_t}{1-t} \right\|^2 = \frac{1}{(1-t)^2} \left\| \hat{x}_1 - x_1 \right\|^2 \tag{5}$$

通过考察该损失相对于预测目标 $\hat{x}_1$的梯度，可以发现与优化动态相关的若干关键问题：

质心附近的梯度消失。 对于固定的 t，梯度幅值与预测值和目标之间的欧几里得距离成正比：

$$\nabla_{\hat{x}_1} \mathcal{L}_{FM} \propto (\hat{x}_1 - x_1) \tag{6}$$

当预测值 $\hat{x}_1$ 接近目标质心 $x_1 \in \mathcal{P}$时，梯度会消失。在训练后期，模型缺乏足够的驱动力，无法精确收敛到质心的准确坐标，从而导致分割边界模糊。

有界空间中的信号尺度有限。 鉴于质心 $\mathcal{P}$ 通常嵌入在诸如 $[-1,1]^d$ 这样的紧致空间中，最大可能距离 $\|\hat{x}_1 - x_1\|$ 本身就很小。对于类别基数 N 较高的任务，质心排列得十分密集，而基于均方误差的梯度为模型区分相邻类别锚点提供的信号非常微弱。

排斥力缺失。 均方误差损失是一个严格的单峰吸引目标函数。它只惩罚与目标质心之间的距离，而不考虑竞争质心。另一方面，

从经典对比学习的角度来看，鲁棒分类器不仅应将预测拉向正类别，还应将其推离负类别。另一方面，从流匹配的角度来看，若缺少排斥机制，目标导向轨迹可能会无意间穿过竞争类别的邻近区域，从而在生成过程中引发语义歧义。

4 方法

本节详细介绍所提出的框架，旨在纠正生成式分割的优化动力学。首先，我们介绍一种用于类别编码的确定性准随机方案。然后，我们提出向量场重塑机制，旨在将判别力引入生成式训练目标。最后，我们提出使用像素神经场解码进行端到端生成式分割训练。

4.1 类别编码

为了将 N 语义类别映射到有界的三维颜色空间 $\mathcal{C} \subset [-1,1]^3$，我们采用一种基于代数无关增量的克罗内克型序列。这是一种兼具确定性和低计算成本的准随机序列。

令 $\mathcal{V} = \{\sqrt{2}, \sqrt{3}, \sqrt{5}\}$ 为前三个素数的平方根集合。这些值在有理数域 $\mathbb{Q}$上线性无关，确保生成的序列不会坍缩到低维流形上，也不会呈现周期相关性。对于每个类别索引 $k \in \{0, \ldots, N-1\}$，首先计算其质心 μ_k：

$$\hat{c}_k = (k \cdot \mathcal{V}) \pmod 1 \tag{7}$$

随后对所得坐标进行归一化和拉伸，使其占据整个 $[-1,1]^3$立方体：

$$\mu_k = 2 \cdot \mathrm{Norm}(\hat{c}_k) - 1 \tag{8}$$

这种方法确保质心分布相对均匀，且点间距离的最小值较大，从而为向量场学习提供稳定且确定性的几何基。

4.2 向量场重塑

如第3节所述，原始流匹配目标无法提供排斥力，导致路径穿越和梯度消失。为解决这一问题，我们通过在真实速度 v_{gt} 中加入判别性校正项，重塑流匹配目标。

基于势能的向量校正。为了使流场具备判别意识，我们在质心空间上构建势场 Φ 。令 $\hat{x}_1$ 表示预测估计值，且$d_k = \|\hat{x}_1 - \mu_k\|^2$ 表示到目标质心 $\mu_k \in \mathcal{P}$的欧几里得距离平方。我们引入变换算子 $\mathcal{T}(\cdot)$，将这些几何距离映射到语义嵌入空间。在我们的主要形式中，将该算子定义为 $\mathcal{T}(d) = -\log(d+\epsilon)$，其中 ϵ 用于避免除以零。 $\mathcal{T}$可灵活适配各种距离变换策略。

通过对变换后的距离应用温度缩放的 Softmax，将距离转换为软分配分布 p：

$$p_k = \frac{\exp(\mathcal{T}(d_k)/\tau)}{\sum_j \exp(\mathcal{T}(d_j)/\tau)} \tag{9}$$

其中， τ 是控制类别分配尖锐程度的温度超参数。为指导优化，我们将判别势能 Φ 定义为当前分配 p与真实独热分布 y之间的散度：

$$\Phi(\hat{x}_1, y) = -\sum_{k=1}^{N} y_k \log p_k \tag{10}$$

当预测 $\hat{x}_1$ 与目标质心 μ_{gt} 完美对齐，同时与竞争者之间的间隔最大化时，该势函数达到全局最小值。校正项 $\nabla\Phi$ 通过以下方式获得

：计算势函数相对于预测 $\hat{x}_1$的梯度。利用性质$\mathcal{T}'(d) = -(d+\epsilon)^{-1}$，梯度表示为：

$$\nabla\Phi = \frac{\partial\Phi}{\partial\hat{x}_1} = \frac{2}{\tau}\sum_{k=1}^{N}(y_k - p_k)\frac{\hat{x}_1 - \mu_k}{d_k + \epsilon} \tag{11}$$

目标重构。 重塑后的目标速度 $\tilde{v}_t$ 定义为将校正项积分到原始速度中：

$$\tilde{v}_t = v_t - \nabla\Phi \tag{12}$$

为保持概率路径的稳定性，并防止网络直接优化势场，我们对重塑后的目标应用停止梯度算子。最终的训练目标定义为：

$$\mathcal{L}_{res} = \mathbb{E}_{t,x_0,x_1}\left\|v_\theta(x_t,t,I) - \text{sg}[\tilde{v}_t]\right\|^2 \tag{13}$$

其中 sg[·]表示分离操作。

梯度动力学比较。 为了进一步阐明流重塑的影响，我们比较标准流匹配损失与所提出目标的梯度。为便于说明，令 $g_\theta = \nabla_\theta v_\theta$ 表示网络相对于其参数的雅可比矩阵。

在标准流匹配框架中，参数更新由以下因素驱动：

$$\nabla_\theta\mathcal{L}_{FM} = g_\theta^\top\left(v_\theta - (x_1 - x_0)\right) \tag{14}$$

随着 v_θ 趋近于常数目标 $v_{gt} = x_1 - x_0$，误差信号会逐渐减弱。该信号纯粹用于重建，无法提供任何有关预测流与错误类别质心之间接近程度的信息。

相比之下，我们的重塑目标 $\mathcal{L}_{res}$ 产生如下梯度：

$$\nabla_\theta\mathcal{L}_{res} = g_\theta^\top\left(v_\theta - (x_1 - x_0 - \text{sg}[\nabla\Phi])\right) \tag{15}$$

通过分解项 $\nabla\Phi$，我们发现判别偏置会改变网络的有效目标：

$$\text{Effective Target} = \underbrace{(x_1 - x_0)}_{\text{Reconstructive}} + \underbrace{\frac{2}{\tau}\sum_{k=1}^{N}(p_k - y_k)\frac{\hat{x}_1 - \mu_k}{\|\hat{x}_1 - \mu_k\|^2 + \epsilon}}_{\text{Discriminative Correction}} \tag{16}$$

讨论。 重塑后的向量场具有以下优势：

1) 自适应排斥：与目标为静态向量的标准目标不同，我们的校正项是指向远离所有质心方向的向量的加权和。当流 v_θ 无意中将预测引向错误的质心 $\mu_j (j \neq gt)$时，项 $p_j \frac{x_1-\mu_j}{d_j}$ 会增大，从而有效地扭曲目标速度，引导轨迹远离竞争质心。

2) 缓解梯度消失：如第 3.2 节所述，标准流匹配梯度在目标质心附近会消失。在我们的公式中，项 $(p_{gt} - 1)\frac{x_1-\mu_{gt}}{d_{gt}}$ 提供了归一化的吸引力。由于 $p_{gt} < 1$, 一项 $(p_{gt} - 1)$即使在预测接近质心时仍能保持显著的幅度，确保网络继续优化。

3）优化动力学：通过引入 sg[·] operator，我们避免了势场二阶梯度所带来的不稳定性。网络学习近似一种重塑后的速度场，该速度场能够内在地感知质心空间中的决策边界，从而将流从简单的插值转变为判别式传输。

4.3 端到端像素解码

传统的基于VAE的潜在空间通常不适用于高精度分割，这是由于流形失真和优化解耦所致。前者会引入潜在伪影，破坏像素级细节；后者则会阻碍扩散过程直接与分割目标对齐。此外，VAE的高内存开销使得完整的端到端优化在计算上难以实现。

为克服这些局限，受 PixNerd 启发 [40], ，我们提出使用像素神经场解码进行端到端生成式分割训练。我们不再通过简单的线性投影来解码图块特征$\mathbf{X}^n \in \mathbb{R}^D$，而是将每个图块视为一个连续神经场。Transformer 骨干网络充当参数生成器，用于预测局部多层感知机（MLP）的权重。对于第 n个图块，网络会动态预测$\{\mathbf{W}_1^n$和$\mathbf{W}_2^n\}$这两个轻量级 MLP 的权重矩阵，这些矩阵由图块特征生成。权重通过非线性投影生成，随后进行 L_2归一化以确保数值稳定性，从而得到¯ ¯归一化权重$\mathbf{W}_1^n$和$\mathbf{W}_2^n$:

$$\mathbf{W}_1^n, \mathbf{W}_2^n = \text{Linear}(\mathbf{X}^n)$$

为了查询图像块内任意空间坐标（i, j）处的速度场，我们通过拼接离散余弦变换（DCT）位置编码与含噪像素状态$\mathbf{x}_t(i, j)$，构造输入上下文$\mathbf{h}_{i,j}$：

$$\mathbf{h}_{i,j} = \text{Concat}([\text{DCT}(i,j), x_t(i,j)])$$

The 最终的逐像素速度 $v^n(i, j)$随后通过应用动态生成的多层感知机

$$v^n(i,j) = \bar{\mathbf{W}}_2^n \sigma(\bar{\mathbf{W}}_1^n \mathbf{h}_{i,j})$$

其中 σ 表示 SiLU 激活函数。

5 实验

本节首先介绍实验设置，然后从定量和定性两方面将其与基线方法进行比较。最后，我们针对不同建模策略开展消融研究，并进行收敛性比较和生成属性分析。

5.1 实验设置

数据集。 我们在两个高基数数据集上评估我们的方法：ADE20K [50] 和 COCO- Stuff [6]。ADE20K 是一个场景解析数据集，涵盖 150 个细粒度语义概念，共包含 20k 张图像。COCO-Stuff 涵盖 171 个标签，共包含 164k 张图像。

实现。 我们的模型使用 PixNerd [40] 权重初始化。所有图像和分割掩码都调整为统一分辨率。训练分两个渐进阶段进行，并采用 AdamW 优化器和表征对齐（REPA） [45] 来加速收敛。阶段 1 使用 256×256 分辨率，批量大小为 64；随后阶段 2 使用 512×512 分辨率，批量大小为 16。默认温度为 1.0。按照标准做法，我们使用二值掩码过滤未标注区域；为简洁起见，记法中省略了这些区域。

基线方法。 我们在实验中考察判别式专用模型和扩散模型。对于判别式专用模型，我们纳入 DeeplabV3+ [9]、OCRNet [47], 、SegFormer [41] 和 Mask- Former [11]。对于扩散模型，我们纳入 InstructDiffusion [14], 、PixWizard [24] 和 Symm- Flow [7]。请注意，一些基于扩散的方法采用了类似的任务，即指代表达图像分割。它与语义分割相似，但会将一个高基数分类问题转化为多个二元分类问题，并借助文本提示进行辅助，因此相较于标准语义分割更具优势。

5.2 主要结果

定量结果。 我们在表 1a 和表1b中分别将我们的方法与判别式专用模型和基于扩散的生成式分割模型进行了比较。总体而言，扩散模型的表现显著逊于判别式模型，即使其以 Stable Diffusion 初始化 [32]，（该模型在大规模文本-图像对数据集上进行了预训练，并具备丰富的世界知识）。这一差距凸显了扩散框架进行生成式分割所面临的固有挑战，与我们在第 3 节中的分析一致。具体而言，在 ADE20K 数据集上，InstructDiffusion 和 PixWizard 分别以 10.5 和 11.3 的平均交并比落后于 DeepLabV3+ 。在 COCO-Stuff 数据集上也存在类似趋势：即使以强大的 SD2.1 初始化，SymmFlow 的平均交并比仍比 SegFormer 低 5。得益于重新设计的向量场学习，我们提出的 FlowSeg 取得了可喜的结果，在 ADE20K 和

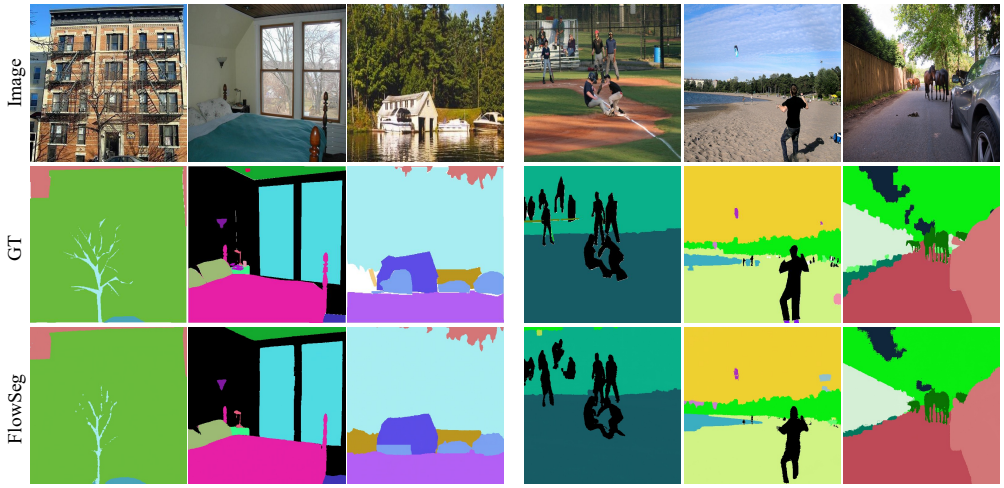
表 1：**在 ADE20K 和 COCO-Stuff 数据集上的比较。** † 表示其执行指代表达图像分割，需要为每个语义类别预测单独的掩码。LSTI 和 (·) 分别表示大规模文本-图像数据集和初始化权重。**粗体**和下划线表示排名第一和第二的结果。

(a) ADE20K 数据集上的性能

	方法	骨干网络	预训练数据	mIoU
判别式专用模型	DeeplabV3+	ResNet101	IN-1k	44.1
	SegFormer	MiT-B2	IN-1k	46.5
	MaskFormer	Swin-T	IN-1k	<u>46.7</u>
扩散模型	InstructDiffusion†	(SD1.5)	LSTI	33.6
	PixWizard	(Lumina-Next-T2I)	LSTI	32.8
	FlowSeg（我们的方法）	(PixNerd)	IN-1k	**47.1**

(b) 在 COCO-Stuff 数据集上的性能

	方法	骨干网络	预训练数据	mIoU
判别式专用模型	DeeplabV3+	ResNet50	IN-1k	38.4
	OCRNet	HRNet-W48	IN-1k	42.3
	SegFormer	MiT-B2	IN-1k	<u>44.6</u>
扩散模型	SymmFlow	(SD2.1)	LSTI	39.6
	FlowSeg（我们的方法）	（PixNerd）	IN-1k	**44.9**



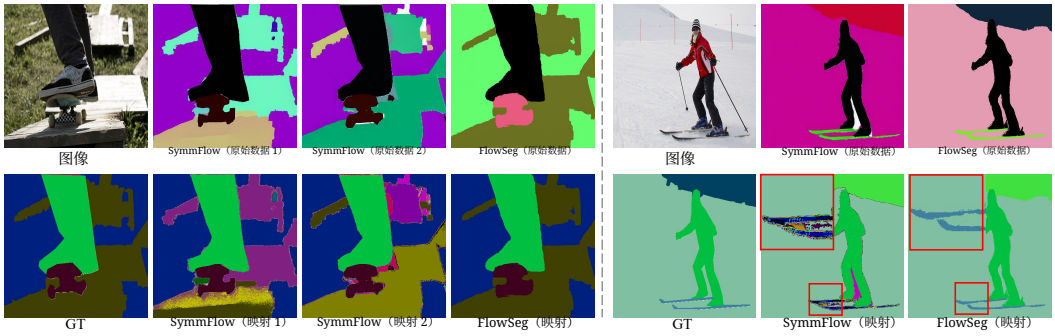
(a) ADE20K (b) COCO-Stuff

图3：在 (a) ADE20K 和 (b) COCO-Stuff 数据集上的分割结果可视化。真实标注（GT）中的白色表示忽略区域。由于 ADE20K 和 COCO-Stuff 数据集具有不同的类别基数，(a) 和 (b) 中的相同颜色不一定表示相同的语义类别。

COCO-Stuff 数据集上分别达到 47.1 和 44.9 的平均交并比，仅凭 ImageNet-1k 预训练便超越了强大的判别式基线。

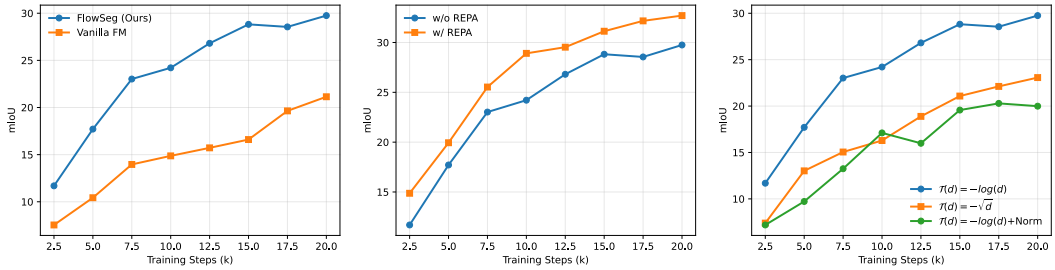
定性结果。如图 3所示，我们展示了多种具有代表性的场景，涵盖从室内到室外的环境，以及不同尺度的物体。这些可视化结果表明，FlowSeg 能够在各种上下文中稳健地处理复杂的分割任务。

此外，我们将 FlowSeg 与 SymmFlow 进行了比较，后者对从高斯噪声与图像的联合分布到分割掩码的传输过程进行建模。如图 4（a）所示，SymmFlow 无法产生确定性预测，在定性表现上劣于我们的方法。



（a）确定性传输与随机传输　　（b）VAE 错位

图 4：**FlowSeg（我们的方法）与 SymmFlow（基线）的视觉对比**。扩散模型首先预测伪掩码（**原始数据**），然后将其映射到最近的语义质心，从而得到最终掩码（**映射**）。（a）确定性建模（我们的方法）与随机建模的比较：SymmFlow 的预测会随随机种子而变化，而我们的方法始终保持一致。（b）基于 VAE 的潜在空间建模会生成颜色相似的掩码，但由于与像素级质心的对齐不完善，这些掩码可能无法对应正确的语义类别。



（a）与流匹配的比较　　（b）表征对齐　　（c）变换算子

图 5：不同训练方案的收敛情况对比。(a) FlowSeg 与原始流匹配。(b) 使用 REPA 与不使用 REPA 进行训练。(c) 不同的变换算子 $\mathcal{T}$。

5.3 消融研究

建模策略。 我们开展了两大类比较。从宏观角度来看，我们的方法作为确定性流运行，并在不使用变分自编码器的情况下进行端到端训练。如图 4（a）所示，SymmFlow 引入高斯噪声，导致不同随机种子下的预测不稳定，这与语义分割的要求相悖。相比之下，我们的 FlowSeg 凭借其确定性，始终能够获得稳定的结果。图 4（b）进一步展示了像素级端到端建模的优势：由于传统的基于变分自编码器的方法无法直接处理像素级任务，因此无法区分潜在空间中细微的逐像素 差异，导致原始数据输出看起来相似，但预测类别却高度不稳定。

关于训练目标设计，表 2 中的消融结果表明，使用退火训练目标$\tilde{v}_t = (1-t)v_t + t(v_t - \nabla\Phi)$本意是平滑训练，但实际上会降低性能，这可能是因为插值削弱了判别势能。此外，掩蔽忽略区域效果显著，因为这可以避免语义模糊区域影响优化。最后，x 预测和 v 预测都可用于模型预测，二者之间的性能差异较小。

收敛性比较。 我们从三个角度分析收敛情况。首先，图 5(a) 表明，FlowSeg 从一开始就实现了显著更快的收敛，并在整个训练过程中始终保持这一优势。在此架构的基础上，我们发现，加入 REPA 可进一步提升性能；如图 5(b) 所示，采用 REPA 后，平均交并比从 29.8 提高到 32.7。此外，我们在图 5(c) 中评估了多种变换算子。结果表明，平方根缩放或势梯度范数缩减等温和算子，无法提供足够的吸引力，使结果趋向目标质心。

表 2：**训练设计选项的消融研究**。静态 $\tilde{v}_t$ 表示原始公式12，而退火 $\tilde{v}_t$ 表示随时间 $\nabla\Phi$ 进行缩放 t。掩码表示在优化过程中排除未标注区域。所有实验均采用v损失。

预测	$\tilde{v}_t$	Mask	mIoU
v预测	退火		24.65
v预测	静态		28.15
v预测	静态	✓	29.75
x预测	静态	✓	29.06



图 6：采样过程中间状态的可视化。

表 3：ADE20K 数据集上采用不同步数进行采样。使用欧拉采样器。

Step	1	2	5	10	20	30	50	100
mIoU	43.47	45.85	46.82	**47.1**	46.78	46.59	46.13	45.78

生成属性分析。FlowSeg 继承了生成模型的内在特性。我们基于流匹配的常微分方程框架，采用欧拉求解器评估不同采样步数下的性能（表 3）。在 ADE20K 数据集上，FlowSeg 在 10 步时达到最佳性能；然而，效率与准确率之间存在权衡：步数过少会导致离散化误差，而步数过多则可能累积近似误差。如图 6所示，中间采样状态在图像与掩码分布之间呈现出平滑的演化轨迹，证实了我们的模型保留了基于流的生成所具有的连续映射特性。

6 结论

本文探讨了通过流匹配实现生成式分割的方法，尤其关注高类别数分割场景。我们发现，以往方法面临两大关键障碍：梯度消失和轨迹穿越。然而，经典的潜空间建模进一步掩盖并加剧了这些问题。为解决这些问题，我们提出了一种基于像素神经场的端到端生成式分割流程。随后，我们提出一种新颖的向量场重塑方法，在同时解决这两个问题的同时，维持语义质心附近较高的梯度，并在速度场中引入排斥力。实证研究和全面的消融实验表明，与经典流匹配训练方案相比，该方法显著提升了收敛速度和最终性能，且表现接近强大的判别式专用模型。我们希望本研究能够启发学界重新思考生成式分割。

参考文献

[1] Niket Agarwal、Arslan Ali、Maciej Bala、Yogesh Balaji、Erik Barker、Tiffany Cai、Prithvijit Chattopadhyay、Yongxin Chen、Yin Cui、Yifan Ding 等。用于物理人工智能的 Cosmos 世界基础模型平台。*arXiv*

预印本 arXiv: *2501 03575*, 2025 1, 3 。Tomer Amit、Tal Shaharbany、Eliya Nachmani 和 Lior Wolf。Segdiff：使用扩散概率模型进行图像分割。arXiv 预印本 arXiv: *2112 00390*, 2021 1, 3 。Dmitry Baranchuk、Ivan Rubachev、Andrey Voynov、Valentin Khrulkov 和 Artem Babenko。利用扩散模型实现高效的语义分割。在 ICLR, 2022 年。1, 3[4] Andreas Blattmann、Tim Dockhorn、Sumith Kulal、Daniel Mendelevitch、Maciej Kilian、Dominik Lorenz、Yam Levi、Zion English、Vikram Voleti、Adam Letts 等。稳定视频扩散：将潜在视频扩散模型扩展至大型数据集。arXiv 预印本 arXiv: *2311 15127*, 2023 1, 3 。Tim Brooks、Aleksander Holynski 和 Alexei A Efros。Instructpix pix：学习遵循图像编辑指令。在 CVPR，2023 年。1, 3[6] Holger Caesar、Jasper Uijlings 和 Vittorio Ferrari。COCO-Stuff：情境中的物体与背景类别。在 CVPR，2018 年。7[7] Francisco Caetano、Christiaan Viviers、Peter HN De With 和 Fons van der Sommen。对称流匹配：基于分数的生成模型统一图像生成、分割和分类。在 AAAI，2026 年。7[8] Liang-Chieh Chen、George Papandreou、Iasonas Kokkinos、Kevin Murphy 和 Alan L Yuille。Deeplab：采用深度卷积网络、空洞卷积和全连接条件随机场的语义图像分割。TPAMI，2017 年。1, 3[9] Liang-Chieh Chen、Yukun Zhu、George Papandreou、Florian Schroff 和 Hartwig Adam。采用空洞可分离卷积的语义图像分割编码器-解码器。在 ECCV, 2018 7 。Bowen Cheng、Ishan Misra、Alexander G Schwing、Alexander Kirillov 和 Rohit Girdhar。用于通用图像分割的掩码注意力掩码变换器。在 CVPR, 2022 1, 3 。Bowen Cheng、Alex Schwing 和 Alexander Kirillov。逐像素分类并非语义分割所需的一切。在 NeurIPS, 2021 年。1, 3, 7[12] Patrick Esser、Sumith Kulal、Andreas Blattmann、Rahim Entezari、Jonas Müller、Harry Saini、Yam Levi、Dominik Lorenz、Axel Sauer、Frederic Boesel 等。扩展整流流变换器，实现高分辨率图像合成。arXiv 预印本 arXiv:2403.03206，2024 年。1, 3[13] Yue Fan、Yongqin Xian、Xiaohua Zhai、Alexander Kolesnikov、Muhammad Ferjad Naeem、Bernt Schiele 和 Federico Tombari。迈向面向密集视觉任务的基于扩散的通用模型。arXiv 预印本 arXiv:2407.00503，2024 年。1, 3[14] Zigang Geng、Binxin Yang、Tiankai Hang、Chen Li、Shuyang Gu、Ting Zhang、Jianmin Bao、Zheng Zhang、Han Hu、Dong Chen 等。InstructDiffusion：面向视觉任务的通用建模接口。在 CVPR，2024 年。1, 2, 3, 7[15] Yoav HaCohen、Nisan Chiprut、Benny Brazowski、Daniel Shalem、Dudu Moshe、Eitan Richardson、Eran Levin、Guy Shiran、Nir Zabari、Ori Gordon 等。Ltx-video：实时视频潜在扩散。arXiv 预印本 arXiv: *2501 00103*, 2024 1, 3 。Jonathan Ho、Ajay Jain 和 Pieter Abbeel。去噪扩散概率模型。在 NeurIPS，2020 年。1, 3[17] Yuanfeng Ji、Zhe Chen、Enze Xie、Lanqing Hong、Xihui Liu、Zhaoqiang Liu、Tong Lu、Zhenguo Li 和 Ping Luo。DDP：用于密集视觉预测的扩散模型。在 CVPR, 2023 1, 3 。Aliasghar Khani、Saeid Asgari Taghanaki、Aditya Sanghi、Ali Mahdavi Amiri 和 Ghassan Hamarneh。Slime：像我一样分割。在 ICLR, 2024 1, 3 。Diederik P Kingma 和 Max Welling。自编码变分贝叶斯。arXiv 预印本 arXiv:1312 6114，2013 年 3 月。[20] Neehar Kondapaneni、Markus Marks、Manuel Knott、Rogério Guimaraes 和 Pietro Perona。面向基于扩散的感知的文本-图像对齐。在 CVPR, 2024 1, 3 。Hsin-Ying Lee、Hung-Yu Tseng 和 Ming-Hsuan Yang。利用扩散先验实现可泛化密集预测。在 CVPR, 2024 1, 2, 3 。Tianhong Li 和 Kaiming He。回归基础：让去噪生成模型执行去噪。arXiv 预印本 arXiv:2511.13720，2025 年。3[23] Chen Liang、Wenguan Wang、Jiaxu Miao 和 Yi Yang。Gmmseg：基于高斯混合的生成式语义分割模型。在 NeurIPS，2022 年。3[24] Weifeng Lin、Xinyu Wei、Renrui Zhang、Le Zhuo、Shitian Zhao、Siyuan Huang、Huan Teng、Junlin Xie、Yu Qiao、Peng Gao 等。PixWizard：支持开放语言指令的多功能图像到图像视觉助手。在 ICLR，2025 年。1, 3, 7[25] Yaron Lipman、Ricky TQ Chen、Heli Ben-Hamu、Maximilian Nickel 和 Matt Le。用于生成建模的流匹配。在 ICLR, 2023 2, 3 。Qiang Liu。整流流：一种保持边缘分布的最优传输方法。arXiv 预印本 arXiv:2209.14577，2022 年。2[27] Xingchao Liu、Chengyue Gong 和 Qiang Liu。流动更直、更快：利用整流流学习生成和迁移数据。在 ICLR, 2023 年。3, 4[28] Xingchao Liu、Xiwen Zhang、Jianzhu Ma、Jian Peng 等。Instaflow：一步即可实现高质量的基于扩散的文本到图像生成。在 ICLR，2024 年。3

[29] Jonathan Long、Evan Shelhamer 和 Trevor Darrell。用于语义分割的全卷积网络。发表于 CVPR，2015年。1、3[30] Dustin Podell、Zion English、Kyle Lacey、Andreas Blattmann、Tim Dockhorn、Jonas Müller、Joe Penna 和 Robin Rombach。Sdxl：改进用于高分辨率图像合成的潜扩散模型。发表于 ICLR，2024年。1、3[31] Lu Qi、Lehan Yang、Weidong Guo、Yu Xu、Bo Du、Varun Jampani 和 Ming-Hsuan Yang。Unigs：用于图像生成与分割的统一表示。发表于 CVPR，2024 1, 3 。Robin Rombach、Andreas Blattmann[32]、Dominik Lorenz、Patrick Esser 和 Björn Ommer。利用潜扩散模型进行高分辨率图像合成。发表于 CVPR，2022 1, 3, 7 。Olaf Ronneberger、Philipp Fischer 和 Thomas Brox[33]。U-net：用于生物医学图像分割的卷积网络。发表于 MICCAI，2015 1, 3 。Jiaming Song、Chenlin Meng 和 Stefano Ermon[34]。去噪扩散隐式模型。发表于 ICLR，2021 1 。Yang Song、Jascha Sohl-Dickstein[35]、Diederik P Kingma、Abhishek Kumar、Stefano Ermon 和 Ben Poole。通过随机微分方程实现基于分数的生成建模。发表于 ICLR，2021 1 。Wouter Van Gansbeke 和 Bert De Brabandere[36]。一种用于全景分割和掩码修复的简单潜扩散方法。arXiv 预印本 arXiv:2401.10227，2024年。1、3[37] Chaoyang Wang、Xiangtai Li、Henghui Ding、Lu Qi、Jiangning Zhang、Yunhai Tong、Chen Change Loy 和 Shuicheng Yan。通过潜扩散模型探索上下文分割。arXiv 预印本 arXiv:2403.09616，2024年。1、3[38] Chaoyang Wang、Xiangtai Li、Lu Qi、Henghui Ding、Yunhai Tong 和 Ming-Hsuan Yang。Semflow：通过整流流将语义分割与图像合成相结合。发表于 NeurIPS，2024 1, 2, 3 。Mengyu Wang、Henghui Ding[39]、Jun Hao Liew、Jiajun Liu、Yao Zhao 和 Yunchao Wei。Segrefiner：利用离散扩散过程实现与模型无关的分割细化。发表于 NeurIPS，2023 1, 3 。Shuai Wang、Ziteng Gao[40]、Chenhui Zhu、Weilin Huang 和 Limin Wang。Pixnerd：像素神经场扩散。arXiv 预印本 arXiv: *2507 23268*, 2025 3, 7 。. Enze Xie、Wenhai Wang、Zhiding Yu、Anima Anandkumar[41]、Jose M Alvarez 和 Ping Luo。Segformer：采用 Transformer 的简洁高效语义分割设计。发表于 NeurIPS，2021年。7[42] Guangkai Xu、Yongtao Ge、Mingyu Liu、Chengxiang Fan、Kangyang Xie、Zhiyue Zhao、Hao Chen 和 Chunhua Shen。将扩散模型改用于通用密集感知任务时，哪些因素至关重要？arXiv 预印本 arXiv: *2403 06090*, 2024 1, 3 。. Yifeng Xu、Zhenliang He、Meina Kan、Shiguang Shan 和 Xilin Chen。Jodi[43]：通过联合建模统一视觉生成与理解。arXiv 预印本 arXiv: *2505 19084*, 2025 2 。. Lehan Yang、Jincen Song、Tianlong Wang[44]、Daiqing Qi、Weili Shi、Yuheng Liu 和 Sheng Li。Vrmdiff：扩散模型生成的文本引导视频指代抠图。arXiv 预印本 arXiv:2503.10678，2025年。1[45] Sihyun Yu、Sangkyung Kwak、Huiwon Jang、Jongheon Jeong、Jonathan Huang、Jinwoo Shin 和 Saining Xie。用于生成的表征对齐：训练扩散 Transformer 比你想象的更容易。发表于 ICLR，2025年。7[46] Yongsheng Yu、Wei Xiong、Weili Nie、Yichen Sheng、Shiqiu Liu 和 Jiebo Luo。PixelDiT：用于图像生成的像素扩散 Transformer。arXiv 预印本 arXiv: *2511 20645*, 2025 3 。. Yuhui Yuan、Xilin Chen 和 Jingdong Wang[47]。用于语义分割的目标上下文表示。发表于 ECCV，2020年。7[48] Hengshuang Zhao、Jianping Shi、Xiaojuan Qi、Xiaogang Wang 和 Jiaya Jia。金字塔场景解析网络。发表于 CVPR，2017年。1、3[49] Wenliang Zhao、Yongming Rao、Zuyan Liu、Benlin Liu、Jie Zhou 和 Jiwen Lu。释放文本到图像扩散模型在视觉感知中的潜力。发表于 ICCV，2023 1, 3 。Bolei Zhou、Hang Zhao、Xavier Puig、Sanja Fidler、Adela Barriuso[50] 和 Antonio Torralba。通过 ADE20K 数据集进行场景解析。发表于 CVPR，2017年。7