arXiv:2302.02394v3 [cs.CV] 5 Oct 2023

通过双循环扩散消除语义图像编辑中的上下文先验偏差

Zuopeng Yang、Tianshu Chu、Xin Lin、Erdun Gao、Daqing Liu、Jie Yang、*IEEE*高级会员和 Chaoyue Wang

摘要——近年来，文本到图像生成扩散模型取得了成功，也彻底革新了语义图像编辑，使人们能够根据查询／目标文本操控图像。尽管取得了这些进展，图像编辑过程中预训练模型可能引入上下文先验偏差，仍是一项重大挑战，例如**对不恰当的区域进行意外修改。为了解决这一问题，我们提出了一种名为双循环扩散的新方法，该方法生成无偏掩码来引导图像编辑。所提模型包含一个偏差消除循环，该循环由正向路径和反向路径组成，每条路径都包含一个结构一致性循环，以确保编辑过程中保留图像内容。正向路径利用预训练模型生成编辑后的图像，而反向路径则将结果还原为源图像。通过比较处理后的源图像与编辑后的图像之间的差异，生成无偏掩码，以确保二者符合相同的分布。实验表明，所提方法效果显著，可将D-CLIP 分数从**0。272 **提升至**0。283。**代码将在 https://github.com/JohnDreamer/DualCycleDiffsion 发布**。

索引词—**上下文先验偏差、语义图像编辑、双循环扩散**

I. 引言

语EMANTIC义图像编辑是图像处理领域一项重要且具有挑战性的问题，其目标是根据指定的文本变换指令修改现有图像。与传统图像编辑技术不同， [42], 例如使用深度 [38], 骨架 [40], 边缘 [39], 热力图 [1], 或分割 [3]‚[37], , 利用文本 [4]–[6], [43] 引导编辑过程，为实现所需的图像修改提供了一种更灵活、易用的方法。如图 1 所示，给定一张猫的图像和指令“猫坐着并戴着围巾的照片”，目标是在保持其他部分完整的同时，在猫的脖子上添加一条围巾

稿件收到日期：2023年2月5日；修订日期：2023年4月11日；录用日期2023年6月11日。发布日期为2022年6月13日；当前版本日期为2023年6月13日。本研究部分受中国国家自然科学基金（编号：61876107、U1803261）资助。本文由副编辑 L. Jiang 推荐。（通讯作者：Jie Yang。）

Zuopeng Yang、Tianshu Chu 和 Jie Yang 任职于中国上海 200240 上海交通大学自动化系图像处理与模式识别研究所（电子邮箱：{yzpeng、chutianshu、jieyang}@sjtu.edu.cn）。Xin Lin 任职于中国广州 510006 广州大学人工智能与区块链研究所（电子邮箱：linxj68@gmail.com）。Erdun Gao 任职于澳大利亚维多利亚州墨尔本 3010 墨尔本大学数学与统计学院（电子邮箱：erdun.gao@student.unimelb.edu.au）。Daqing Liu 和 Chaoyue Wang 任职于中国北京 100176 京东探索研究院（电子邮箱：li-udq.ustc@gmail.com、chaoyue.wang@outlook.com）。

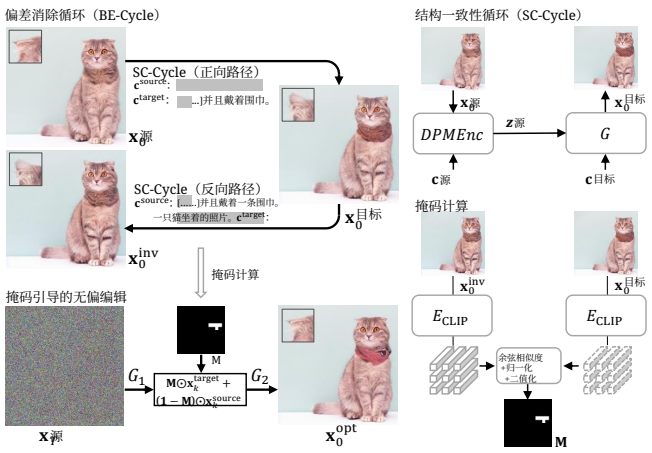


图1. 所提出的用于语义图像编辑任务的双循环扩散框架概览。左侧展示了偏差消除循环（BE-Cycle）和掩码引导的无偏编辑的流程及细节。**对猫耳朵的修改**是由**预训练模型产生的上下文先验偏差**导致的，相关示例展示在图像左上角。给定一幅源图像、一段源文本和一段目标文本，我们首先利用 BE-Cycle 生成无偏掩码，再用该掩码引导图像编辑。右侧展示了结构一致性循环（SC- Cycle）的细节 [7]以及掩码计算过程。 ⊙ 是逐元素乘积。

图像中的区域。因此，这项任务可视为文本条件图像生成的扩展，能够精准且直观地辨别需要修改和应予保留的区域。

近年来，文本条件生成技术取得了长足进展，DALL-E [19], 、Make-A-Scene [20], 、Imagen [21],、Parti [25], 和 DALL-E2 [22] 等模型相继问世。这些模型在从互联网收集的大量数据上进行训练，因此在文本语义理解和图像合成能力方面均取得了显著提升。在这些模型中，基于扩散的模型因其能够通过迭代优化，将随机高斯噪声转化为逼真的图像而备受关注。

作为一种开创性方法，SDEdit [13]通过随机微分方程（SDE）进行迭代去噪来实现编辑，但往往会修改整幅图像。Prompt2prompt [24] 通过控制交叉注意力图，将编辑局限在特定区域，从而避免仅在合成图像上出现意外修改。DiffEdit [23] 通过对比以不同文本提示为条件的扩散模型预测结果，自动生成掩码。然而，据观察，它无法生成聚焦于需要编辑区域的掩码

。此外，为了保留图像的完整性，CycleDiffusion [7]提出了DPMEncoder，用于编码潜在代码。

尽管利用预训练模型进行语义图像编辑具有诸多优势，但训练图像与目标编辑图像之间的分布偏移会引入先验偏差，从而带来一项重大挑战。本文特别关注上下文先验偏差问题，该问题往往会导致不同区域之间出现虚假相关性，进而使原本不打算修改、却仍符合目标文本上下文要求的区域遭到不必要的改动。例如，如图 1 左上角所示，尝试给猫的脖子添加围巾时，预训练模型未能保持耳朵的形状。

据我们所知，本文是首批致力于消除文本引导语义图像编辑中上下文先验偏差的研究之一。我们提出双循环扩散，这是一种通过生成无偏掩码来引导图像编辑、从而有效消除上下文先验偏差的方法。该方法包含偏差消除循环（BE-Cycle），由正向路径和反向路径组成；每条路径都包含一个结构一致性循环（SC-Cycle），以在编辑过程中保持图像内容的完整性。给定目标文本后，正向路径利用预训练模型合成编辑后的图像。随后，反向路径使用描述原始图像的源文本，将编辑后的图像还原为源图像。

此外，该方法既可编辑真实图像，也可编辑由其他模型生成的合成图像；只要源文本能够充分描述图像的主要内容，就不必是生成原始图像时所使用的文本。具体而言，源文本可以通过现成的图像描述模型 [2] 自动生成，也可以通过对目标文本稍作修改来获得，例如更改、添加或删除若干词语。如图 1 所示，一个例子是删除短语“并戴着围巾”。处理后的源图像与编辑后的图像符合相同的分布，因此可以通过识别两者特征之间的差异，自动生成无偏掩码。为此，我们采用改进的掩码计算流程，利用 CLIP [10] 的编码器提取两幅图像的特征。随后，生成的掩码将用于引导无偏语义图像编辑。全面的实验表明，所提方法具有优越性，能够显著提升定量指标和定性视觉效果。

II. 所提方法

A. 预备知识

扩散概率模型（DPMs）。DPMs [8], [26]–[28] 是一类生成模型，旨在通过逐步对服从正态分布的变量进行去噪，来估计底层数据分布 $p(\mathbf{x})$。这可以看作学习长度为 T 的固定马尔可夫链的逆过程。在语义图像编辑是文本到图像生成任务扩展这一背景下，我们当前仅关注文本条件扩散概率模型。在推理阶段，当给定文本查询 $\mathbf{c}$ 时，DPMs 利用参数化均值估计器μ_θ，逐步对从白高斯噪声 $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ 中采样得到的合成图像进行去噪。利用重参数化技巧，我们可以对 $\mathbf{x}_{t-1} \sim \mathcal{N}(\mu_\theta(\mathbf{x}_t, \mathbf{c}, t), \sigma_t^2 \boldsymbol{I})$ 进行采样，如下所示:

$$\mathbf{x}_{t-1} = \mu_\theta(\mathbf{x}_t, \mathbf{c}, t) + \sigma_t \boldsymbol{\epsilon}_t,\ t = 1, \cdots, T, \tag{1}$$

其中，σ_t 是第 t 步中的标准差，且 $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$。注意，图像生成方向是从 $\mathbf{x}_T$ 到 $\mathbf{x}_0$。为简化符号表示，本文始终将 $\mathbf{z} := \mathbf{x}_T \oplus \boldsymbol{\epsilon}_{1:T}$用作潜在代码，其中 $\oplus$ 表示拼接。因此，从 $\mathbf{x}_T$ 到 $\mathbf{x}_0$ 的映射可表示为 $\mathbf{x}_0 = G(\mathbf{z}, \mathbf{c})$。

DPM-Encoder。为获取潜在代码 $\mathbf{z}$，Wu 等人 [7]提出了DPM-Encoder，这是一种用于随机扩散概率模型的可逆编码器。具体而言，随机扩散概率模型在正向扩散过程中定义了后验分布 $q(\mathbf{x}_{1:T}|\mathbf{x}_0)$ [8], [9] 。随后，DPM-Encoder 根据$q(\mathbf{x}_{1:T}|\mathbf{x}_0)$生成带噪图像 $\mathbf{x}_1, \cdots, \mathbf{x}_T$，并计算 $\boldsymbol{\epsilon}_t$。形式上，DPM-Encoder 的整体采样过程 $\mathbf{z} \sim q_{\text{DPMEnc}}(\mathbf{z}|\mathbf{x}_0, \mathbf{c}, \mu_\theta)$可定义为:

$$\begin{aligned} &\mathbf{x}_1, \cdots, \mathbf{x}_T \sim q(\mathbf{x}_{1:T}|\mathbf{x}_0), \\ &\boldsymbol{\epsilon}_t = (\mathbf{x}_{t-1} - \mu_\theta(\mathbf{x}_t, \mathbf{c}, t))/\sigma_t,\ t = 1, \cdots, T, \\ &\mathbf{z} := \mathbf{x}_T \oplus \boldsymbol{\epsilon}_{1:T}. \end{aligned} \tag{2}$$

B. 结构一致性循环

SC-Cycle [7] 的目标是在编辑过程中保留图像内容的完整性。该框架如图 1 右上角所示。给定源图像$\mathbf{x}_0^{\text{source}}$和源描述$\mathbf{c}^{\text{source}}$，SC-Cycle 首先通过 DPM-Encoder 对潜在代码$\mathbf{z}^{\text{source}}$ 进行编码。然后，利用目标文本$\mathbf{c}^{\text{target}}$应用映射 G，生成编辑后的图像$\mathbf{x}_0^{\text{target}}$：

$$\begin{aligned} \mathbf{z}^{\text{source}} &\sim q_{\text{DPMEnc}}(\mathbf{z}|\mathbf{x}_0^{\text{source}}, \mathbf{c}^{\text{source}}, \mu_\theta), \\ \mathbf{x}_0^{\text{target}} &= G(\mathbf{z}^{\text{source}}, \mathbf{c}^{\text{target}}) = G(\mathbf{x}_T^{\text{source}} \oplus \boldsymbol{\epsilon}_{1:T}^{\text{source}}, \mathbf{c}^{\text{target}}). \end{aligned} \tag{3}$$

SC-Cycle [7] 合成的编辑后的图像$\mathbf{x}_0^{\text{target}}$ 利用潜在代码$\mathbf{z}^{\text{source}}$保留了源图像$\mathbf{x}_0^{\text{source}}$的特有细节。这一特性使$\mathbf{x}_0^{\text{source}}$与$\mathbf{x}_0^{\text{target}}$ 在结构上具有相似性。

C. 偏差消除循环

BE-Cycle 旨在通过对比正向路径和反向路径生成的图像，自动生成无偏掩码。如图 1 左上角所示，每条路径都包含一个 SC-Cycle。在正向路径中，$\mathbf{x}_0^{\text{target}}$ 按照公式 (3) 合成。然后，在反向路径中，我们尝试以源文本$\mathbf{c}^{\text{source}}$为条件，从$\mathbf{x}_0^{\text{target}}$ 恢复源图像:

$$\begin{aligned} \mathbf{z}^{\text{target}} &\sim q_{\text{DPMEnc}}(\mathbf{z}|\mathbf{x}_0^{\text{target}}, \mathbf{c}^{\text{target}}, \mu_\theta), \\ \mathbf{x}_0^{\text{inv}} &= G(\mathbf{z}^{\text{target}}, \mathbf{c}^{\text{source}}) = G(\mathbf{x}_T^{\text{target}} \oplus \boldsymbol{\epsilon}_{1:T}^{\text{target}}, \mathbf{c}^{\text{source}}). \end{aligned} \tag{4}$$

BE-Cycle 的正向路径和反向路径都依赖于同一个预训练模型，因此生成图像$\mathbf{x}_0^{\text{target}}$ 和$\mathbf{x}_0^{\text{inv}}$中存在相似的上下文先验偏差。换言之，两幅图像都符合相同的分布。因此，上下文先验偏差不会影响通过对比$\mathbf{x}_0^{\text{target}}$ 和$\mathbf{x}_0^{\text{inv}}$来识别

通过对比$\mathbf{x}_0^{\text{target}}$和$\mathbf{x}_0^{\text{inv}}$，确定需要编辑的区域。

为生成掩码，首先使用预训练 CLIP 模型 [10]中的视觉编码器 E_{CLIP}，提取$\mathbf{x}_0^{\text{target}}$和$\mathbf{x}_0^{\text{inv}}$的视觉特征：

$$\mathbf{F} = E_{\text{CLIP}}(\mathbf{x}_0),\ \mathbf{F} \in \mathbb{R}^{m\times n\times d}, \tag{5}$$

其中，$\mathbb{R}^{m\times n}$ 表示空间域，d 表示特征维度，$\mathbf{F}_{i,j,:}$ 表示坐标为 (i,j) 的图像网格特征，其中 $i \in \{1,2,\cdots,m\}$ 且 $j \in \{1,2,\cdots,n\}$。为衡量各区域的变化程度，通过以下方式计算$\mathbf{F}_i^{\text{target}}{}_{,j,:}$与$\mathbf{F}_i^{\text{inv}}{}_{,j,:}$之间的余弦相似度：

$$\mathbf{S}_{i,j} = \cos\left\langle \mathbf{F}_{i,j,:}^{\text{target}}, \mathbf{F}_{i,j,:}^{\text{inv}} \right\rangle. \tag{6}$$

最后，我们可以通过以下方式得到掩码$\mathbf{M}$：

$$\mathbf{M}_{i,j} = \begin{cases} 1, & \text{if } \frac{\text{abs}(\mathbf{S}_{i,j}) - \min(\text{abs}(\mathbf{S}))}{\max(\text{abs}(\mathbf{S})) - \min(\text{abs}(\mathbf{S}))} > \delta, \\ 0, & \text{else}, \end{cases} \tag{7}$$

其中，δ 是用于控制掩码二值化的阈值；max(·) 和 min(·) 分别获取矩阵的最大值和最小值，而 abs(·) 则返回矩阵中每个元素或标量输入的绝对值。实际应用中，我们设定 $\delta = 0.5$。为提高掩码的分辨率，将图像划分为 2×2 个网格，每个网格分别生成一个掩码。随后，根据各网格的空间位置将所有网格掩码拼合，得到最终掩码。最后，将掩码调整大小，使其与$\mathbf{x}_0$的空间尺寸相匹配。

D. 掩码引导的无偏编辑

在最后阶段，生成的无偏掩码用于引导图像编辑过程。整个流程如图 1 左下角所示。正如 [11], 所指出的，文本条件扩散概率模型在采样初期主要依赖文本提示来引导采样过程，随着生成的进行，则逐渐转向依赖视觉特征。因此，我们利用文本提示引导掩码内的生成，同时尽可能保持掩码外的区域与源图像相似。为方便起见，我们将映射 G 分为两个步骤：$G(\cdot) = G_2(G_1(\cdot))$。在第一步中，我们获得受目标文本强烈影响的图像：

$$\mathbf{x}_k^{\text{target}} = G_1(\mathbf{x}_T^{\text{source}} \oplus \boldsymbol{\epsilon}_{(k-1):T}^{\text{source}}, \mathbf{c}^{\text{target}}), \tag{8}$$

其中，$k \in \{1,2,\cdots,T\}$ 表示扩散过程的第 k步。然后，我们从后验分布 $q(\mathbf{x}_k|\mathbf{x}_0^{\text{source}})$中采样$\mathbf{x}_k^{\text{source}}$，并利用掩码优化图像：

$$\mathbf{x}_k^{\text{opt}} = \mathbf{M} \odot \mathbf{x}_k^{\text{target}} + (\mathbf{1} - \mathbf{M}) \odot \mathbf{x}_k^{\text{source}}, \tag{9}$$

其中，$\odot$ 是逐元素乘积。最后，通过以下方式合成不受上下文先验偏差影响的编辑后的图像：

$$\mathbf{x}_0^{\text{opt}} = G_2(\mathbf{x}_k^{\text{opt}} \oplus \boldsymbol{\epsilon}_{1:k}^{\text{source}}, \mathbf{c}^{\text{target}}). \tag{10}$$

无偏掩码仅关注需要编辑的区域，从而避免对最终编辑图像$\mathbf{x}_0^{\text{opt}}$ 造成意外修改，例如猫耳朵。

表 I Q零样本-语义图像编辑数据集上的定量结果 [7]。 *表示每种超参数组合仅进行了 1 次试验，而其余实验均进行了 15 次试验。最佳和次佳结果以高亮显示。第二结果以高亮显示。

	方法	$\mathcal{S}_{\text{CLIP}}\uparrow$	$\mathcal{S}_{\text{D-CLIP}}\uparrow$	峰值信噪比↑	结构相似性指数↑
LDM-400M	SDEdit [13]	0.332	0.264	13.68	0.390
	DiffEdit [23]	0.323	0.208	18.50	0.588
	CycleDiffusion [7]	0.333	0.275	18.72	0.625
	Ours	0.333	0.281	19.12	0.635
SD-v1-1	SDEdit [13]	0.339	0.248	15.50	0.498
	DiffEdit [23]	0.316	0.175	22.47	0.729
	CycleDiffusion [7]	0.331	0.262	21.98	0.731
	Ours	0.332	0.265	22.21	0.734
SD-v1-4	SDEdit [13]	0.344	0.258	15.93	0.512
	DiffEdit [23]	0.318	0.172	22.70	0.731
	CycleDiffusion [7]	0.334	0.272	21.92	0.731
	Ours	0.337	0.283	22.08	0.730
SD-v1-4	SDEdit* [13]	0.335	0.221	15.64	0.505
	DiffEdit* [23]	0.305	0.088	22.72	0.733
	CycleDiffusion* [7]	0.327	0.206	21.85	0.721
	我们的方法*	0.331	0.235	22.71	0.741

表 II 使用不同掩码进行的对比，比较是否使用 CLIP [10] 编码器。通过对比 BE-Cycle 中正向路径/反向路径的输入与输出，生成有偏/无偏掩码。所有实验均使用 SD-v1-4 预训练模型进行。

	使用 CLIP 编码器	$\mathcal{S}_{\text{CLIP}}\uparrow$	$\mathcal{S}_{\text{D-CLIP}}\uparrow$	峰值信噪比↑	结构相似性指数↑
使用有偏掩码	✗	0.336	0.281	21.49	0.725
	✓	0.338	0.281	21.85	0.724
使用无偏掩码	✗	0.337	0.282	21.90	0.725
	✓	0.337	0.283	22.08	0.730

III. 实验

A. 实现细节

实验是在零样本语义图像编辑数据集 [7]，上进行的，该数据集专门为语义图像编辑任务收集。数据集包含 150 个三元组，每个三元组由一张源图像、一段源文本和一段目标文本组成。为确保公平比较，DiffEdit [23]采用其原论文中的设置，其余实验则遵循 CycleDiffusion [7] 的设置：使用 100 步的 DDIM采样器（$\eta = 0.1$），将编码过程的无分类器引导尺度设为 1，并将解码步骤的无分类器引导尺度遍历为 $\{1, 1.5, 2, 3, 4, 5\}$。为保留图像内容，编辑从一张未完全加噪的图像开始。因此，我们遍历了添加噪声的步骤数 $\{85, 80, 75, 70, 60, 50\}$。[1] 随后，从 $\{85, 80, 75, 70, 60, 50\}$中选择优化步骤 k。每种超参数组合均进行了 15 次试验。为消除掩码计算过程中随机噪声的影响，我们对使用不同超参数组合生成的所有掩码取平均值。最终结果根据方向性CLIP分数 $\mathcal{S}_{\text{D-CLIP}}$自动选出。

指标： 为评估所有对比方法及我们的双循环扩散的编辑性能，我们采用四项指标来评估编辑后的图像对源图像的忠实度及其对目标文本的真实性。它们是**结构相似性指数**、**峰值信噪比**、CLIP分数 $\mathcal{S}_{\text{CLIP}}$ [10]和方向性CLIP分数$\mathcal{S}_{\text{D-CLIP}}$ [12]。**结构相似性指数** 用于衡量两幅图像之间的相似性

[1]这与 [7] 中 $\{15, 20, 25, 30, 40, 50\}$ 的提前停止步骤相同。

，而**峰值信噪比**用于量化图像质量。$\mathcal{S}_{\text{CLIP}}$ 可评估生成图像与目标文本之间的对齐程度，而 $\mathcal{S}_{\text{D-CLIP}}$ 可评估图像变化与文本变化之间的相似性。此外，我们还开展了一项用户研究，以进一步评估该模型消除上下文先验偏差的效果。参与者根据目标文本和源图像，投票选出意外修改较少的图像；每种方法之间的比较获得了1000张选票。结果见表III。

B. 与现有方法的比较

为验证所提出的双循环扩散的有效性，我们将模型与其他基于扩散模型的方法进行了比较，包括 SDEdit [13], DiffEdit [23], 和 CycleDiffusion [7]。为探究数据规模、数据质量及训练细节对模型性能的影响，我们采用了若干预训练文本到图像扩散模型：（1）LDM-400M，一种包含 1.45B 个参数的潜在扩散模型 [17]，在 LAION-400M [15], 上训练；（2）SD-v1-1，一种包含 0.98B 个参数的 Stable Diffusion 模型 [17]，在 LAION-5B [16],上训练；（3）SD-v1-4，由 SD-v1-1 微调而来，以提升美学效果和无分类器引导采样性能。

表 I 展示了各参评方法的定量结果。在三种预训练模型中，所提方法在所有指标上均取得了最佳或具有竞争力的得分，表明其能够同时提升编辑后的图像对源图像的忠实度以及对目标文本的真实性。对 LDM-400M 和 SD-v1-4 结果的比较表明，较大的训练数据规模有助于理解图像内容，从而在使用 SD-v1-4 时更好地保留图像内容。此外，通过比较 SD-v1-1 和 SD-v1-4 的结果可以观察到，改进的无分类器引导采样有助于合成对目标文本更具真实性的编辑后的图像。总体而言，使用 SD-v1-4 作为预训练模型取得了最佳性能。因此，后续所有实验和比较都将基于 SD-v1-4 模型展开。

在图 2 中，我们展示了由不同方法生成的编辑结果的若干视觉比较。从这些视觉比较中可以观察到，所提方法不仅能够根据文本完成修改，

而且图像内容保留效果更好。例如，如图 2 第一行所示，目标是将绵羊替换为老虎。SDEdit 修改了整幅图像，而 DiffEdit 和 CycleDiffusion 均能保留源图像的大部分内容。但它们仍意外地在汽车上添加了一个车门把手，黄色方框标出了这一点。这一现象也可以证明上下文先验偏差的存在。此外，第二行的目标是添加一个苹果。CycleDiffusion 未能保留源图像中背包和苹果的形状，而我们的方法成功做到了这一点。DiffEdit 甚至未能生成另一个苹果。第三行展示了更多由所提方法生成的结果。

表III 用户研究结果。参与者按照目标文本和源图像，投票选出意外修改较少的图像。

	CycleDiffusion	我们的方法（使用有偏掩码）	我们的方法（使用无偏掩码）
投票数（%）	42.7	46.8	参考

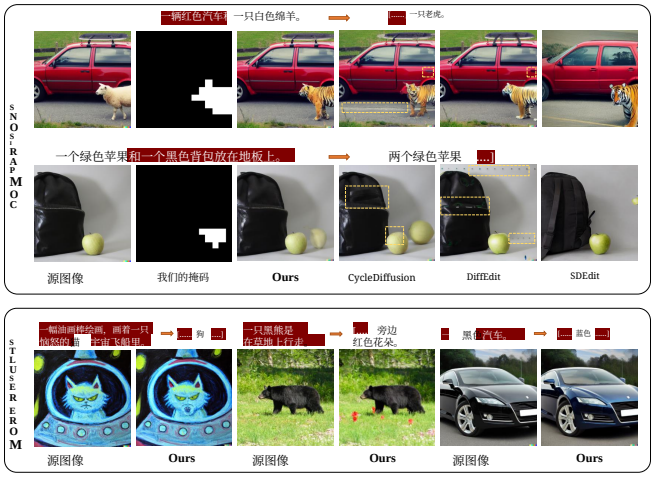


图 2. 我们所提方法生成的掩码和编辑样本。前两行展示了与两个最具代表性的基线模型的比较：SDEdit [13], DiffEdit [23], 和 CycleDiffusion [7]。在所有样本中，我们的方法在图像内容保留方面优于其他模型。黄色方框标示了意外修改，而掩码则显示了具体的编辑区域。最后一行展示了所提方法生成的更多结果。



图 3. BE-Cycle 的正向路径和反向路径的掩码、输出及最终编辑结果。具体而言，通过比较正向或反向路径的输入与输出生成掩码。为消除掩码计算过程中随机噪声的影响，我们对使用不同超参数组合生成的所有掩码取平均值。通过比较，我们可以利用无偏掩码获得与源图像相似的冰淇淋。

C. 消融研究

本节首先旨在探究 BE-Cycle 和改进的掩码计算流程的有效性，进而进一步展示上下文先验偏差的存在。为此，我们基于有偏掩码开展了额外实验：使用相同的流程对比 $\mathbf{x}_0^{\text{source}}$ 和 $\mathbf{x}_0^{\text{target}}$，以计算无偏掩码。此外，我们还展示了不使用 CLIP 编码器时的结果。这里，我们使用 S D-v1-4 作为预训练模型。定量结果见表 II。从结果中可以看出，所有模型都获得了相近的 $\mathcal{S}_{\text{CLIP}}$ 和 $\mathcal{S}_{\text{D-CLIP}}$ 得分。然而，使用 CLIP 编码器的模型在图像内容保留方面表现更好，这证明了所提出的掩码计算流程是有效的。接下来，我们通过比较使用不同掩码的模型，讨论掩码的影响

。从使用 CLIP 编码器的结果可以看出，使用反向路径生成的无偏掩码的模型，在 **结构相似性指数**和 **峰值信噪比** 得分上取得了显著提升。这意味着反向路径提高了识别需要编辑区域的准确性。换言之，我们能够以更少的修改获得更好的效果。使用有偏掩码导致性能较差的原因在于上下文先验偏差的存在，这会造成意外修改。总之，BE-Cycle 能有效消除源自预训练文本到图像模型的上下文先验偏差。

此外，我们还在图3中展示了 BE-Cycle 的路径所生成的掩码和输出的视觉对比，并提供了使用不同掩码得到的最终编辑图像。此处，我们对使用不同超参数组合生成的所有掩码取平均，以消除掩码计算过程中噪声的影响（即男孩张开的嘴）。从正向路径的结果可以看出，使用预训练模型会改变部分图像内容，例如冰淇淋的形状。这一变化也反映在掩码中。因此，使用有偏掩码时，模型会缩短最终编辑图像中冰淇淋的长度。相比之下，由反向路径生成的无偏掩码能够聚焦于圣诞帽所在的区域，从而合成出尽可能接近源图像的冰淇淋。总之，所提方法能够在不受上下文先验偏差影响的情况下实现图像编辑。

IV. 结论

本文针对语义图像编辑中的上下文先验偏差问题，提出了双循环扩散，通过生成无偏掩码来引导图像编辑，从而消除其影响。我们的方法采用 SC-Cycle 和 BE-Cycle，消除预训练模型生成的源图像与生成图像之间的分布偏移。借助 CLIP 的编码器提取图像视觉内容，我们的模型能够通过识别各个区域的内容差异来生成无偏掩码。这样，模型便可以专注于需要编辑的区域，不受上下文先验偏差的影响。通过大量实验和消融研究，我们验证了双循环扩散优于其他基于扩散的语义图像编辑方法。

参考文献

[1] H. Yan、H. Zhang、J. Shi 和 J. Ma，《用于时尚灵感迁移的纹理笔刷：一种带热力图引导语义解耦的生成对抗网络》，*IEEE TCSVT*，2022年。[2] R. Mokady、A. Hertz 和 A. H. Bermano，《Clipcap：用于图像描述生成的 CLIP 前缀》，*arXiv* 预印本 *arXiv:2111.09734*，2021年。[3] S. Gu、J. Bao、H. Yang、D. Chen、F. Wen 和 L. Yuan，《掩码引导的人像编辑与条件生成对抗网络》，载于*CVPR* 论文集，2019年，第 3436–3445 页。[4] Z. Xu、T. Lin、H. Tang、F. Li、D. He、N. Sebe、R. Timofte、L. Van Gool 和 E. Ding，《预测、预防与评估：由预训练视觉语言模型赋能的解耦文本驱动图像操作》，载于*CVPR* 论文集，2022年，第 18 229–18 238 页。[5] B. Li、X. Qi、T. Lukasiewicz 和 P. H. Torr，《ManiGAN：文本引导的图像操作》，载于 *CVPR* 论文集，2020年，第 7880–7889 页。[6] T. Zhang、H.-Y. Tseng、L. Jiang、W. Yang、H. Lee 和 I. Essa，《文本作为神经算子：通过文本指令进行图像操作》，载于第 *29* 届 *ACM* 多媒体大会论文集，2021年，第 1893–1902 页。

[7] C. H. Wu 和 F. De la Torre，《统一扩散模型的潜在空间，并将其应用于 CycleDiffusion 和引导》，arXiv 预印本 arXiv:2210.05559，2022年。[8] J. Ho、A. Jain 和 P. Abbeel，《去噪扩散概率模型》，NeurIPS，第 33 卷，第 6840–6851 页，2020年。[9] J. Song、C. Meng 和 S. Ermon，《去噪扩散隐式模型》，载于 ICLR，2021年。[10] A. Radford、J. W. Kim、C. Hallacy、A. Ramesh、G. Goh、S. Agarwal、G. Sastry、A. Askell、P. Mishkin、J. Clark 等，《通过自然语言监督学习可迁移视觉模型》，载于 ICML。PMLR，2021年，第 8748–8763 页。[11] Y. Balaji、S. Nah、X. Huang、A. Vahdat、J. Song、K. Kreis、M. Aittala、T. Aila、S. Laine、B. Catanzaro 等，《ediffi：采用专家去噪器集成的文本到图像扩散模型》，arXiv 预印本 arXiv:2211.01324，2022年。[12] O. Patashnik、Z. Wu、E. Shechtman、D. Cohen-Or 和 D. Lischinski，《StyleCLIP：文本驱动的 StyleGAN 图像操控》，载于 ICCV 论文集，2021年，第 2085–2094 页。[13] C. Meng、Y. He、Y. Song、J. Song、J. Wu、J.-Y. Zhu 和 S. Ermon，《SDEdit：利用随机微分方程进行引导式图像合成与编辑》，载于 ICLR，2021年。[14] X. Su、J. Song、C. Meng 和 S. Ermon，《用于图像到图像转换的双重扩散隐式桥》，arXiv 预印本 arXiv:2203.08382，2022年。[15] C. Schuhmann、R. Kaczmarczyk、A. Komatsuzaki、A. Katta、R. Vencu、R. Beaumont、J. Jitsev、T. Coombes 和 C. Mullis，《LAION-400M：经 CLIP 筛选的 4 亿组图文对开放数据集》，载于 NeurIPS 数据中心 AI 研讨会，第 FZJ-2022-00923 号。J¨ulich 超级计算中心，2021年。[16] C. Schuhmann、R. Beaumont、R. Vencu、C. W. Gordon、R. Wightman、M. Cherti、T. Coombes、A. Katta、C. Mullis、M. Wortsman 等，《LAION-5B：用于训练下一代图文模型的开放大规模数据集》，载于第 36 届 NeurIPS 数据集与基准赛道会议，2022 年。[17] R. Rombach 、 A. Blattmann、D. Lorenz、P. Esser 和 B. Ommer，《利用潜在扩散模型进行高分辨率图像合成》，载于 CVPR 论文集，2022年，第 10 684–10 695 页。[18] J. Zhang、P. Yang、W. Wang、Y. Hong 和 L. Zhang，《通过分割引导的自注意力网络进行图像编辑》，SPL，第 27 卷，第 1605–1609 页，2020年。[19] A. Ramesh、M. Pavlov、G. Goh、S. Gray、C. Voss、A. Radford、M. Chen 和 I. Sutskever，《零样本文本到图像生成》，载于 ICML。PMLR，2021年，第 8821–8831 页。[20] O. Gafni、A. Polyak、O. Ashual、S. Sheynin、D. Parikh 和 Y. Taig-man，《Make-A-Scene：具有人类先验的基于场景的文本到图像生成》，载于 ECCV，2022年。[21] C. Saharia、W. Chan、S. Saxena、L. Li、J. Whang、E. Denton、S. K. S. Ghasemipour、B. K. Ayan、S. S. Mahdavi、R. G. Lopes 等，《通过深度语言理解实现照片级真实感的文本到图像扩散模型》，arXiv 预印本 arXiv:2205.11487，2022年。[22] A. Ramesh、P. Dhariwal、A. Nichol、C. Chu 和 M. Chen，《利用 CLIP 潜变量进行分层文本条件图像生成》，arXiv 预印本 arXiv:2204.06125，2022年。[23] G. Couairon、J. Verbeek、H. Schwenk 和 M. Cord，《DiffEdit：基于扩散并通过掩码引导进行语义图像编辑》，载于 ICLR，2023年。[24] A. Hertz、R. Mokady、J. Tenenbaum、K. Aberman、Y. Pritch 和 D. Cohen-Or，《通过交叉注意力控制实现提示词到提示词图像编辑》，arXiv 预印本 arXiv:2208.01626，2022年。[25] J. Yu、Y. Xu、J. Y. Koh、T. Luong、G. Baid、Z. Wang、V. Vasudevan、A. Ku、Y. Yang、B. K. Ayan 等，《扩展自回归模型以实现内容丰富的文本到图像生成》，TMLR，2022年。[26] Y. Song 和 S. Ermon，《通过估计数据分布的梯度进行生成建模》，NeurIPS，第 32 卷，2019年。[27] Y. Song、J. Sohl-Dickstein、D. P. Kingma、A. Kumar、S. Ermon 和 B. Poole，《通过随机微分方程进行基于得分的生成建模》，载于 ICLR，2020年。[28] D. Watson、W. Chan、J. Ho 和 M. Norouzi，《通过对样本质量求导来学习扩散模型的快速采样器》，载于 ICLR，2021年。[29] C. Lu、Y. Zhou、F. Bao、J. Chen、C. Li 和 J. Zhu，《DPM-Solver：一种用于扩散概率模型采样的快速常微分方程求解器，可在约 10 步内完成》，载于 NeurIPS，2022年。[30] Q. Zhang 和 Y. Chen，《利用指数积分器快速采样扩散模型》，arXiv 预印本 arXiv:2204.13902，2022年。[31] T. Karras、M. Aittala、T. Aila 和 S. Laine，《阐明基于扩散的生成模型的设计空间》，载于 NeurIPS，2022年。[32] M. Chen 和 Z. Liu，《EDBGAN：通过边缘感知双分支生成对抗网络进行图像修复》，SPL，第 28 卷，第 842–846 页，2021年。

[33] S. S. Phutke 和 S. Murala，《Fasnet：用于图像修复的特征聚合与共享网络》，*SPL*，第 29 卷，第 1664–1668 页，2022年。[34] C. Han 和 J. Wang，《利用进化生成器进行人脸图像修复》，*SPL*，第 28 卷，第 190–193 页，2021年。[35] S. S. Phutke 和 S. Murala，《用于图像修复的基于多样化感受野的对抗式并发编码器网络》，*SPL*，第 28 卷，第 1873–1877 页，2021年。[36] X. Xu、Y.-C. Chen、X. Tao 和 J. Jia，《通过图文共享空间实现文本引导的人像图像编辑》，*IEEE TPAMI*，2021年。[37] H. Zheng、Z. Lin、J. Lu、S. Cohen、J. Zhang、N. Xu 和 J. Luo，《利用高分辨率稀疏注意力进行语义布局编辑》，*IEEE TPAMI*，2022年。[38] G. Luo、Y. Zhu、Z. Weng 和 Z. Li，《用于基于深度的视图合成的遮挡区域修复框架》，*IEEETPAMI*，第 42 卷，第 6 期，第 1289–1302 页，2019年。[39] S. Xu、D. Liu 和 Z. Xiong，《E2i：从边缘到图像的生成式图像修复》，*IEEE TCSVT*，第 31 卷，第 4 期，第 1308–1322 页，2020年。[40] P. Zhang、L. Yang、X. Xie 和 J. Lai，《用于姿态引导人像生成的轻量级纹理相关网络》，*IEEE TCSVT*，2021年。[41] L. Zhang、H. Yang、T. Qiu 和 L. Li，《Ap-gan：提升视频人脸替换中的属性保持能力》，*IEEE TCSVT*，第 32 卷，第 4 期，第 2226–2237 页，2021年。[42] Y. Liu、Q. Li、Q. Deng 和 Z. Sun，《利用预训练 StyleGAN 对人脸图像进行空间解耦操控》，*IEEE TCSVT*，2022年。[43] O. Avrahami、D. Lischinski 和 O. Fried，《用于自然图像文本驱动编辑的混合扩散》，载于*CVPR* 论文集，2022年，第 18 208–18 218 页。