ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 40,No. 4,Jul. 2025,pp. 1035－1045
DOI:10. 16337/j. 1004-9037. 2025. 04. 016

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于映射融合嵌入扩散模型的文本引导图像编辑方法

吴　飞[1]，马永恒[1]，邓哲颖[1]，王银杰[1]，季一木[2]，荆晓远[3]

(1. 南京邮电大学人工智能学院,南京 210023; 2. 南京邮电大学计算机学院,南京 210023; 3. 武汉大学计算机学院,武汉 430072)

摘　要：在只有图像和目标文本提示作为输入的情况下,对真实图像进行基于文本引导的编辑是一项极具挑战性的任务。以往基于微调大型预训练扩散模型的方法,往往对源文本特征和目标文本特征进行简单的插值组合,用于引导图像生成过程,这限制了其编辑能力,同时微调大型扩散模型极易出现过拟合且耗时长的问题。提出了一种基于映射融合嵌入扩散模型的文本引导图像编辑方法(Text-guided image editing method based on diffusion model with mapping-fusion embedding, MFE-Diffusion)。该方法由两部分组成:(1)大型预训练扩散模型与源文本特征向量联合学习框架,使模型可以快速学习以重建给定的原图像;(2)特征映射融合模块,深度融合目标文本与原图像的特征信息,生成条件嵌入,用于引导图像编辑过程。在具有挑战性的文本引导图像编辑基准TEdBench上进行实验验证,结果表明所提方法在图像编辑性能上具有优势。

关键词：文本引导图像编辑;扩散模型;图像生成;图像编辑;特征映射融合

中图分类号：TP391　　**文献标志码**:A

Text-Guided Image Editing Method Based on Diffusion Model with Mapping-Fusion Embedding

WU Fei[1], MA Yongheng[1], DENG Zheying[1], WANG Yinjie[1], JI Yimu[2], JING Xiaoyuan[3]

(1. College of Artificial Intelligence, Nanjing University of Posts and Telecommunications, Nanjing 210023, China; 2. School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China; 3. School of Computer Science, Wuhan University, Wuhan 430072, China)

Abstract: Text-guided editing of real images with only images and target text prompts as input is an extremely challenging problem. Previous approaches based on fine-tuning large pre-trained diffusion models often simply interpolate and combine source and target text features to guide the image generation process, which limits their editing capabilities, while fine-tuning large diffusion models is highly susceptible to overfitting and time-consuming problems. In this paper, we propose a text-guided image editing method based on diffusion model with mapping-fusion embedding (MFE-Diffusion). The method consists of the following two components: (1) A large pre-trained diffusion model and source text feature vectors joint learning framework, which enables the model to quickly learn to reconstruct the original image. (2) A feature mapping-fusion module, which deeply fuses the feature information of the target text and the

基金项目:国家自然科学基金(62076139);信息系统工程全国重点实验室开放基金(05202305);南京邮电大学1311人才计划。
收稿日期:2024-06-19;**修订日期**:2024-10-29

original image to generate conditional embedding that is used to guide the image editing process. Experimental validation on the challenging text-guided image editing benchmark TEdBench shows that the proposed method has advantages in image editing performance.

Key words: text-guided image editing; diffusion model; image generation; image editing; feature mapping-fusion

引　言

近年来,图像生成技术取得了显著进步,已成为计算机视觉领域的关键分支之一。这一领域不只涉及到图像的自动生成,还包括多种图像编辑任务。这些编辑任务可通过多种引导方式来实现,包括图像掩码[1]、草图[2]、风格迁移[3]及文本描述[4]等。在这些方法中,文本描述作为引导方式最为直观,易于理解,同样难度也是最高的,因此以文本为引导的图像编辑尤其受到关注。

本文旨在探索解决具有挑战性的文本引导的图像编辑问题,即仅使用原始图像和目标文本描述,这代表了文本引导图像编辑领域对输入的最基本要求。图像编辑任务又可以按照编辑效果分为两类,一类是外观调整型编辑,主要关注于改变图像的外观、特征和风格,而不涉及图像内容的实质性改变;另一类是内容变换型编辑,涉及到对图像中某些部分的替换、添加或删除,从而改变图像的内容或含义。根据基础的结构,又可以将目前主流的文本引导图像编辑方法分为基于生成对抗网络(Generative adversarial networks, GAN)的方法和基于扩散模型的方法。基于GAN的方法,如ManiTrans[5]、DE-NET[6],因其不可解释性、训练难收敛、可控性差,以及生成图像的分辨率通常比较低等问题逐渐被扩散模型取代[7]。基于扩散模型的方法,如ControlNet[8]、SDEdit[1]、PnP Diffusion[9]、instruct pix2pix[10]等,都可以进行效果不错的外观调整型编辑,但不足以在保留原本图像特征的同时进行复杂的内容变换型编辑,同时引导编辑过程往往不局限于文本描述,例如ControlNet可以引入深度图、草稿图、姿态图等多种引导。Imagic[11]是一种三阶段文本引导图像编辑方法,首先通过微调过程使其可以重建原图像,并在采样过程融入目标文本信息以达到图像编辑效果。将目标文本描述视为描述原始图像的伪源提示,在第一阶段,Imagic将目标文本嵌入视为可学习的网络参数并对其进行微调,使其可以作为源文本嵌入的替代。第二阶段,Imagic微调UNet网络,冻结其他参数所有的参数值,使模型可以更好地重建原图像。在第三阶段,Imagic对微调后的伪源文本嵌入和目标文本嵌入进行插值,并利用插值后的文本嵌入引导文本到图像的生成,从而完成编辑。搭配大型预训练图像生成模型Imagen[12]的Imagic是目前最先进的文本引导图像编辑方法。然而,其简单地对微调后的伪源文本嵌入和目标文本嵌入进行插值,这种融合方式不能有效地在保留语义无关部分(图像中不需要编辑的部分)信息的同时,融合语义相关部分(图像中需要编辑的部分)的信息,限制了其模型的编辑能力。并因其多阶段的微调过程,需要花费很长时间并耗费大量计算资源,同时存在过度拟合的问题。

针对以上问题,提出基于映射融合嵌入扩散模型的文本引导图像编辑方法(Text-guided image editing method based on diffusion model with mapping-fusion embedding, MFE-Diffusion),该方法分为两个阶段,分别是参数微调和采样推理。在参数微调阶段,使用Stable Diffusion-v1.5模型[13]作为基础图像生成模型。首先使用ChatGPT4-Vision模型生成多段关于原图像的描述,这些文本描述尽可能从不同的角度概括原图像,将其嵌入的平均值作为源文本嵌入,并联合扩散模型的UNet网络进行统一微调,得到优化后的文本嵌入和微调后的UNet网络,使模型可以快速重建原图像,大大缓解了先前方法在微调扩散模型时耗时长且容易出现过拟合的问题。在采样推理阶段,通过特征映射融合模块,深度融合目标文本和原图像的特征信息,得到引导扩散模型编辑过程所需的条件嵌入,有效地缓解了先前方法

中存在的编辑对象不准确、编辑效果差的问题。

综上所述,本文的主要贡献可以概括如下:

(1)提出了一个大型预训练扩散模型与源文本特征嵌入联合学习框架,用于联合微调大型预训练扩散模型与源目标文本特征嵌入,使模型可以快速学习重建给定原图像。

(2)提出了特征映射融合模块,用于深度融合目标文本与原图像的特征信息,精确引导图像生成过程。

(3)所提方法在目前最具挑战性的文本引导图像编辑基准TEdBench[11]上进行实验验证,结果表明所提方法在图像编辑性能上具有显著优势。

1　基于扩散模型的文本引导图像编辑方法

1.1　任务定义

给定原图像I和目标文本T,文本引导图像编辑任务的目标为:在仅给定目标文本和图像的情况下,文本引导图像编辑根据目标文本的描述对图像进行编辑,这不仅要求需要准确编辑需要被修改的部分,而且需要保持其他与文本描述无关的部分不变。

1.2　扩散模型基本原理

去噪扩散概率模型(Denoising diffusion probabilistic models, DDPM)[14]从给定图像I_0开始,然后在每个时间步t中依次添加高斯噪声$\epsilon_t\sim\mathcal{N}(0,1)$得到$I_t$,在这样的扩散过程中,每个时间步$t\in0,1,\cdots,T$的$I_t$可以直接通过式(1)计算得出。

$$I_t=\sqrt{\alpha_t}\,I_0+\sqrt{1-\alpha_t}\,\epsilon_t \tag{1}$$

式中:α_t作为扩散强度系数,取值范围为$\alpha_0\to1>a_1>\cdots>a_{t-1}>a_t\to0$。

给定I_t和文本嵌入e以及时间序列$t\in0,1,\cdots,T$,可以通过扩散模型中的UNet网络得到预测的噪声$\epsilon_\theta(I_t,t,e)$,在模型能力绝对理想情况下,ϵ_θ和添加到I_{t-1}中的随机噪声ϵ_t应该是相同的。通过模型预测得到的ϵ_θ,经过去噪扩散隐式模型(Denoising diffusion implicit models, DDIM)[15]过程,即可得到I_{t-1}

$$I_{t-1}=\frac{\sqrt{\alpha_{t-1}}}{\sqrt{\alpha_t}}\Big(I_t-\sqrt{1-\alpha_t}\,\epsilon_\theta(I_t,t,e)\Big)+\sqrt{1-\alpha_{t-1}}\,\epsilon_\theta(I_t,t,e) \tag{2}$$

UNet模型整体训练损失如下

$$L=\mathbb{E}_{I,\epsilon\sim\mathcal{N}(0,1),t,e}\left[\|\epsilon_t-\epsilon_\theta(I_t,t,e)\|_2^2\right] \tag{3}$$

在Stable Diffusion[13]中通过VAE编码首先对原图像I_0进行特征提取得到Z_0,训练损失为

$$L=\mathbb{E}_{z,\epsilon\sim\mathcal{N}(0,1),t,e}\left[\|\epsilon_t-\epsilon_\theta(Z_t,t,e)\|_2^2\right] \tag{4}$$

式中:$t\in0,1,\cdots,T$表示时间序列,Z_t表示在t时刻下的潜空间特征,e表示文本嵌入,$\epsilon_\theta(Z_t,t,e)$表示通过UNet网络预测得到的噪声,ϵ_t表示在前向扩散过程中添加的标签噪声。

本文使用Stable Diffusion作为基础预训练扩散模型,首先通过微调其Unet网络使其可以快速重建原图像,并通过特征映射融合模块将目标文本特征注入采样过程中,以达到图像编辑的目的。

1.3　模型框架

提出基于映射融合嵌入扩散模型的文本引导图像编辑方法(MFE-Diffusion),框架如图1所示,模型基于Stable Diffusion[13],分为参数微调和采样推理两部分。

在参数微调阶段,首先使用ChatGPT4-Vision生成n段关于原图像的描述,称之为源文本描述。然

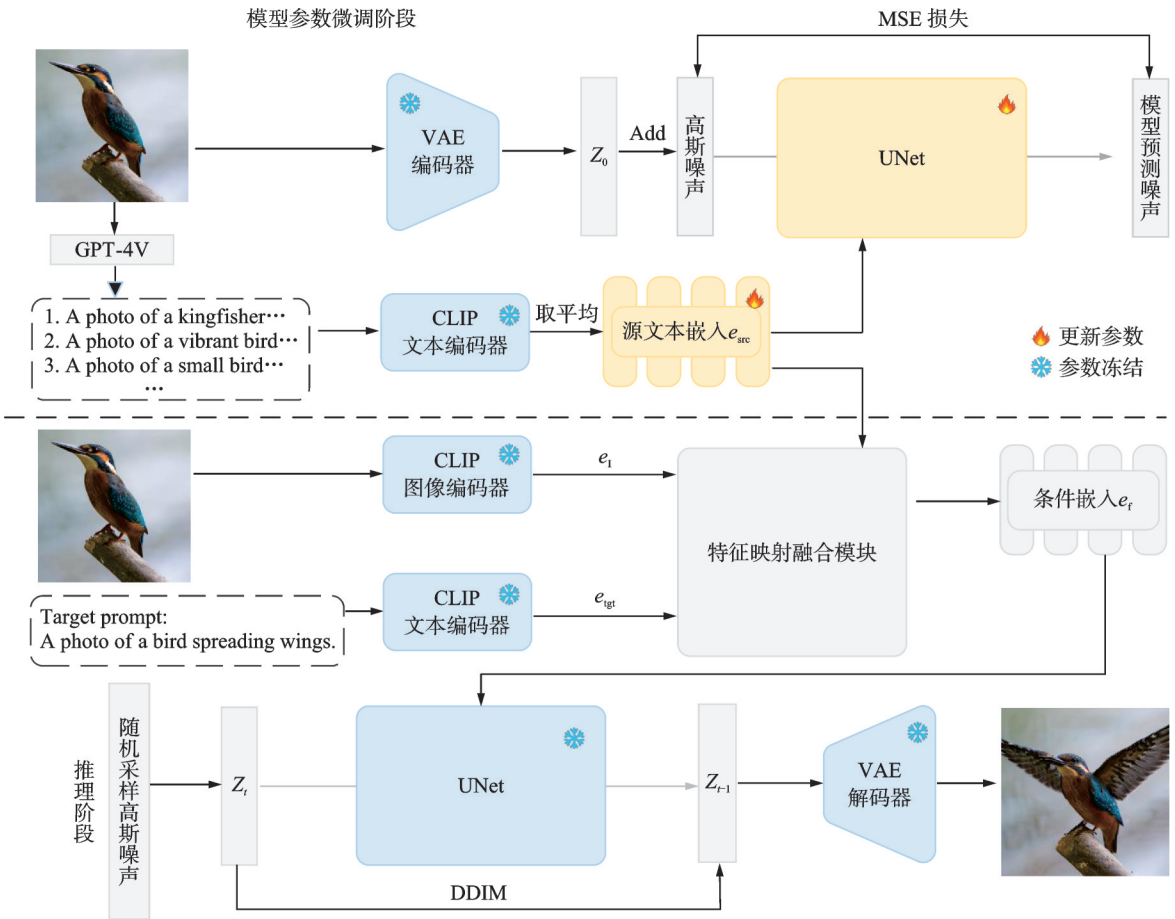

图1 本文方法总体框架

Fig.1 The overall framework of the proposed method

后将源文本描述输入Stable Diffusion的对比语言-图像预训练(Contrastive language-image pre-training, CLIP)文本编码器,得到n个对应的文本嵌入组成的文本嵌入集合$E_t=\{e_{t_1}, e_{t_2}, \cdots, e_{t_n}\}$,接着在数值上对$E_t$取平均,得到源文本嵌入$e_{src}$,随后将$e_{src}$视作可学习的网络参数,并使用不同的学习率对$e_{src}$与UNet联合优化得到优化后的文本嵌入$e_{opt}$和优化后的UNet模型。

在采样推理阶段,对文本嵌入集合$E_t=\{e_{t_1}, e_{t_2}, \cdots, e_{t_n}\}$进行格拉姆-施密特正交化,构建语义子空间$S$。目标文本$T$和原图像$I$分别经过CLIP文本编码器和CLIP图像编码器得到目标文本嵌入e_{tgt}和原图像嵌入e_I,随后将两者同时映射到语义子空间S中,并取二者在语义子空间中的映射与优化后的文本嵌入e_{opt}进行融合,得到最终融合后的条件嵌入e_f,用于引导编辑过程。

1.4 预训练扩散模型与特征向量联合学习框架

考虑到需要被编辑的图片极大可能不在预训练扩散模型原本的训练集中,为了使模型可以快速重建原图像,需要使用原图像对模型进行微调,以实现高质量的重建和语义理解。

如图1所示,为了充分地捕捉原图像的特征细节,使模型可以尽可能好地重建原图像,首先使用ChatGPT4-Vision生成n段关于原图像的描述,通过大模型提示词工程,如表1所示,使ChatGPT4-Vision模型尽可能地从不同角度描述原图像[16],然后将原图像的文本描述输入Stable Diffusion的CLIP文

表1　ChatGPT4-Vision图像描述提示词

Table 1　ChatGPT4-Vision prompt for image description

提示词	Describe this picture from five different perspectives, requiring: 1. Use only one sentence for each description. Each sentence should not be too long, about 15 words. 2. Begin each sentence with "A photo of". 3. Description content as close as possible to the source image, as faithful as possible to the image, do not have unnecessary associations. 4. Describe the image from different perspectives as much as possible.

本编码器中,得到n个对应的文本嵌入组成的文本嵌入集合$E_t=\{e_{t_1},e_{t_2},\cdots,e_{t_n}\}$,接着在数值上对$E_t$取平均,得到源文本嵌入$e_{\text{src}}$。

先前的三阶段文本引导的图像编辑方法Imagic将目标文本嵌入视为e_{src}。本文通过实验发现,使用上述方法生成的文本描述作为源文本描述,而不是像Imagic那样使用目标文本描述作为伪源文本描述,可以有效地缓解过度拟合问题。本文还发现,使用ChatGPT4-Vision模型从不同角度生成的源文本描述可以更好地涵盖原图像的特征信息,可以更好地与给定图像进行语义对齐,从而获得更好的重建原图像。

与Imagic类似,本文也将源文本嵌入视为可学习的网络参数。但与Imagic不同的是,本文联合优化源文本嵌入e_{src}和Stable Diffusion的UNet参数。实验发现这样做除了可以提高整个微调过程的速度,还可以提高模型对原图像的重构质量。考虑到UNet和源文本嵌入e_{src}的参数规模不同,本文使用较小的学习率微调UNet,使用较大的学习率微调e_{src}。

使用均方误差损失来衡量Unet模型的预测噪声与标签噪声之间的差异,采用源文本嵌入e_{src}作为源文本描述,并通过监控损失值的收敛情况,当最终损失下降为0.02时,停止训练来确保模型适时收敛,防止训练过度。损失函数如下

$$L=\mathbb{E}_{z,\epsilon\sim\mathcal{N}(0,1),t,e_{\text{src}}}\left[\left\|\epsilon_t-\epsilon_{\theta,e_{\text{src}}}(Z_t,t,e_{\text{src}})\right\|_2^2\right] \tag{5}$$

式中:$t\in 0,1,\cdots,T$表示时间序列,Z_t表示在t时刻下的潜空间特征,e_{src}表示源文本嵌入,$\epsilon_\theta(Z_t,t,e_{\text{src}})$表示通过UNet网络预测得到的噪声,$\epsilon_t$表示在前向扩散过程中添加的标签噪声。

与式(3)不同的是,e_{src}也被看作可学习的参数进行优化,通过联合扩散模型与文本特征向量进行微调,得到优化后的文本嵌入e_{opt}。

1.5　特征映射融合模块

在采样推理阶段,为了使生成的图像可以在保留原本不需要被编辑的语义特征的同时,深度融合需要被编辑的语义特征,本文提出特征映射融合模块,结构如图2所示。首先对文本嵌入集合$E_t=\{e_{t_1},e_{t_2},\cdots,e_{t_n}\}$进行格拉姆-施密特正交化,构建语义子空间$S$。并将原图像和目标文本描述分别送入CLIP图像编码器和CLIP文本编码器,得到原图像嵌入e_I和目标文本嵌入e_{tgt},将两者分别映射到子空间S中,得到映射后的原图像嵌入m_I和目标文本嵌入m_{tgt}。最终与优化后的原始文本嵌入e_{opt}进行融合,得

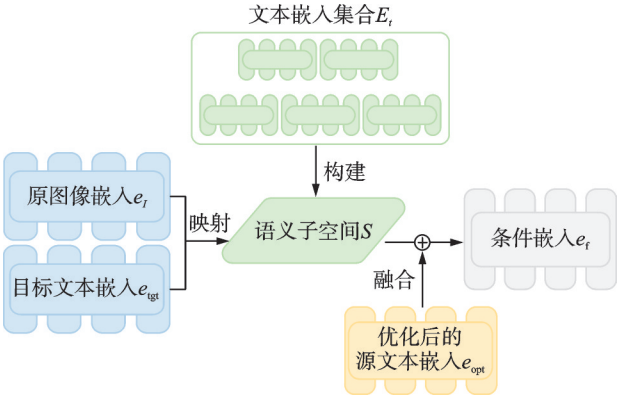

图2　特征映射融合模块结构图

Fig.2　Structure diagram of feature mapping fusion module

到条件嵌入e_{f},用于引导编辑过程。相较于直接融合原图像和文本特征,先进行子空间映射再融合具有更大优势。子空间是由ChatGPT4-Vision通过生成多段原图像描述,将该描述输入Stable Diffusion并进行一系列处理所得。与对原始特征直接融合相比,其确保了对图像内容更深刻、更精确的描述。

1.5.1 语义子空间构建

Zhou等[17]发现,由于CLIP联合空间是一个向量空间,因此可以通过一组相关的文本描述作为基向量来构建一个子空间,如果原始图像嵌入的某个属性发生变化,子空间中相应的属性也会发生变化。受这一发现的启发,在参数微调阶段,首先通过ChatGPT4-Vision生成n段关于原图像的描述,称之为源文本描述,然后将源文本描述输入Stable Diffusion的文本编码器,得到由n个对应的文本嵌入$\boldsymbol{e}_{t_i}$组成的文本嵌入集合$\boldsymbol{E}_t=\{\boldsymbol{e}_{t_1},\boldsymbol{e}_{t_2},\cdots,\boldsymbol{e}_{t_n}\}$,对每个集合中的每个向量进行归一化处理后,形成一组基向量$\boldsymbol{B}=\{\boldsymbol{b}_k\}_N$。最后为了保留更多的语义信息,对$\boldsymbol{B}$进行格拉姆-施密特正交化处理,得到一组相互正交的基向量,也就是语义子空间$S=\{\hat{\boldsymbol{b}}_k\}_N$。

1.5.2 映射过程

构建语义子空间S后,在本文中将原图像和目标文本描述分别送入CLIP图像编码器和CLIP文本编码器,得到原图像嵌入e_I和目标文本嵌入e_{tgt},并将两者分别映射到子空间S中,分别得到映射后的原图像嵌入m_I和映射后的目标文本嵌入m_{tgt},即

$$\begin{cases} m_I=\sum_k^N\left(\hat{b}_k'e_I\right)\hat{b}_k \\ m_{\mathrm{tgt}}=\sum_k^N\left(\hat{b}_k'e_{\mathrm{tgt}}\right)\hat{b}_k \end{cases} \tag{6}$$

式中$(\cdot)'$表示转置操作。

1.5.3 融合过程

为了使最终融合后的条件嵌入e_{f}可以在保留原始图像特征中的语义无关部分的同时,也就是保留不需要被修改的特征信息,又可以深度融合语义相关部分,也就是需要被修改的特征信息,从而准确引导编辑过程,本文设计增强了文本相关成分对条件嵌入e_{f}的影响,削弱了文本无关成分对条件嵌入e_{f}的影响。

具体来说,由于e_I和e_{tgt}被映射到同一个由一组标准正交基构成的子空间S中,分别得到m_I和m_{tgt},所以在该子空间中m_I和m_{tgt}以及其权重系数α_k和β_k又可以被表示为如下形式

$$m_I=\sum_k^N\alpha_k\hat{b}_k,m_{\mathrm{tgt}}=\sum_k^N\beta_k\hat{b}_k \tag{7}$$

$$\alpha_k=\hat{b}_k'm_I,\beta_k=\hat{b}_k'm_{\mathrm{tgt}} \tag{8}$$

本文设计了以下融合公式

$$e_{\mathrm{f}}=\sum_{k=1}^N(1-\varepsilon)\alpha_k\hat{b}_k+(1-\varepsilon)e_{\mathrm{opt}}+\sum_{k=1}^N\varepsilon\beta_k\hat{b}_k \tag{9}$$

式中:超参数ε可以控制融合强度,精确地调节目标文本特征在最终生成图像中的权重。在式(9)中,首先通过ε对第一项$\sum_{k=1}^N(1-\varepsilon)\alpha_k\hat{b}_k$和第二项$(1-\varepsilon)e_{\mathrm{opt}}$进行削弱,削弱$m_I$和$e_{\mathrm{opt}}$中所有成分的影响。第一项削弱$m_I$的所有成分,用于减少文本无关属性的影响,与此同时,文本相关属性的影响也被减小。第二项融入优化后的原始文本嵌入,起到了进一步平衡作用,确保原图像的无关部分不会对最终结果产生过强的影响。第三项$\sum_{k=1}^N\varepsilon\beta_k\hat{b}_k$融入映射后目标文本嵌入$m_{\mathrm{tgt}}$,并对前两项被削弱的部分进行补偿,弥补前两项编辑任务相关属性的损失,最终得到融合后的特征嵌入e_{f},用于引导图像生成过程。

2　实验及结果分析

2.1　实验数据集与评价指标

TEdBench数据集[11]在Imagic中首次被提出,被认为是目前公开的最难的文本引导图像编辑基准之一。该数据集中共有100个编辑目标,每个编辑目标有1句目标文本描述和1张原图像。这些目标文本描述通常简洁明了、多种多样又高度概括编辑需求,既包含了外观调整型编辑,也包含了内容变换型编辑。特别是对于许多文本引导图像编辑方法来说,内容变换型编辑是非常困难的。

在定量评估方面,利用CLIP Score[18]来衡量编辑后的图像与目标文本描述之间的语义一致性,并利用IS(Inception score)[19]来衡量编辑后图像的真实性。

CLIP Score是基于CLIP模型[20]的图像-文本语义匹配指标。作为一种多模态模型,CLIP能够同时处理图像和文本两个模态,并将其投射到相同的特征空间中。将编辑后的图像和目标文本输入CLIP模型,分别提取两者的特征嵌入,通过计算图像和文本特征向量的余弦相似度得到CLIP Score,以衡量图像和文本描述之间的相似性。CLIP Score越高,表示编辑后的图像与目标文本的语义相关性越强。基于计算图像和文本特征的余弦相似度,可以得出CLIP Score。具体计算公式为

$$\text{CLIP Score}=\frac{\boldsymbol{v}_{\mathrm{I}}\cdot\boldsymbol{v}_{\mathrm{T}}}{\|\boldsymbol{v}_{\mathrm{I}}\|\|\boldsymbol{v}_{\mathrm{T}}\|} \tag{10}$$

式中:$\boldsymbol{v}_{\mathrm{I}}$为图像的特征向量,$\boldsymbol{v}_{\mathrm{T}}$为文本的特征向量。该分数反映了生成的图像与目标文本的匹配度,CLIP Score越高,说明图像与文本描述越一致。

IS使用Inception网络通过多样性和清晰度两个方面来评估生成图像的真实性。具体计算公式为

$$\text{IS}=\exp(\mathbb{E}_{x\sim P_g}D_{\mathrm{KL}}(p(y|x)/\!/p(y))) \tag{11}$$

式中:$x\sim P_g$表示x从P_g中生成的图像样本,$p(y|x)$为对图像x生成的条件类别分布,$p(y)$表示边缘分布,D_{KL}为KL散度。IS值越高,意味着生成图像的多样性越高,且质量越逼真。

2.2　实验设置

使用单张NVIDIA A6000 GPU,单次微调时间约为2 min,推理阶段使用DDIM进行采样,单张图像采样步数设为40,推理时间约为1.5 s。在微调阶段使用Adam优化器微调扩散模型中的UNet部分和源文本嵌入e_{src},其中UNet部分的学习率设置为5e−6,源文本嵌入e_{src}的学习率设置为2e−3,对于Adam优化器,本文使用$\beta_1=0.9$,$\beta_2=0.999$微调模型参数。式(9)的超参数ε设置为0.6,用于控制特征融合的强度,并在2.5节介绍设置不同大小超参数ε对于编辑结果的影响。

2.3　对比实验

为了验证MFE-Diffusion的先进性,与SDEdit[1]、SDXL img2img[21]和Imagic[11]进行定量比较。此外,与SDXL img2img和Imagic进行定性比较,从与文本相关的特征处理和与文本无关的特征保留两方面评价编辑后的图像质量。

2.3.1　定量分析

定量实验结果如表2所示,在TEdBench基准数据集上获得了超越SDEdit、SDXL img2img和Imagic的IS分数和CLIP Score。MFE-Diffusion通过特征映射融合模块和联合学习框架,既有效融合了目标文本和原图像的特征信息,提高了CLIP Score,又避免了过拟合从而提高IS分数。

表2　定量实验结果

Table 2　Quantitative experimental results

方法	IS ↑	CLIP Score ↑
SDEdit	5.551	0.214
SDXL img2img	5.659	0.258
Imagic	5.819	0.274
MFE-Diffusion	5.883	0.279

更高的IS分数表明了本方法可以生成更逼真的图像，而更高的CLIP Score说明所提方法可以更正确和准确地处理与文本相关的图像部分，并保留与文本无关的部分，从而使编辑后的图像与目标文本描述取得更好的语义一致性。

2.3.2 定性分析

如图3所示，从两个角度评估编辑图像的质量：与目标文本相关的特征处理和与目标文本无关的特征保留。首先，MFE-Diffusion在所有方法中实现了最好的文本无关部分的保留，例如第1列中的背景布帘和两个木支架、第2列中香蕉的纹理、第4列背景中的树木形状以及最后1列中花瓶里的绿叶等都属于不需要被修改的区域，MFE-Diffusion都很好地进行了保留，而对比方法都对其进行了或多或少的修改。其次，在目标文本相关的特征编辑方面，所提方法同样是最优秀的。例如第3列中SDXL img2img不能将书合上，Imagic虽然成功合上了书，但场景有了很大的变化，而所提方法可以在原场景下将书修改为闭合的状态。在第4列将网球修改为番茄的场景中，SDXL img2img和Imagic生成的番茄都不是一个合理的红番茄。同样在第5列将小熊修改为站立状态的场景中，SDXL img2img依旧无法准确地达成编辑目标，Imagic虽然成功将小熊修改为站立状态，但小熊身后的场景已经出现了很大的变化。总之，MFE-Diffusion不仅可以准确根据目标文本描述修改原图像，而且可以精准地保留与目标文本描述无关的特征。

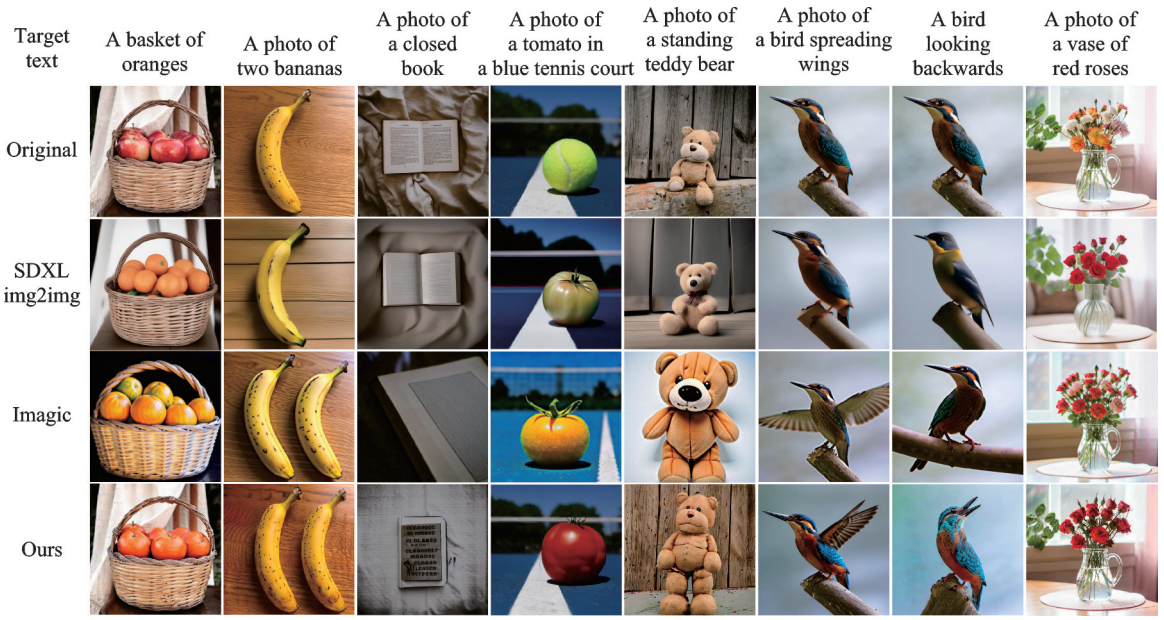

图3 不同方法在TEdBench数据集上的定性比较

Fig.3 Qualitative comparison between different methods on TEdBench dataset

2.4 消融实验

本节将对所提方法中的主要组件的重要性进行评估。首先将不使用预训练扩散模型与特征向量联合学习的方法版本称为Ours-J，其次将不使用特征映射融合模块的方法版本称为Ours-M。具体来说，在Ours-J中，该版本方法在原来的框架中不对扩散模型进行微调，但保留推理采样阶段的原本操作，直接进行图像编辑过程；在Ours-M中，该版本方法不使用特征映射融合模块，仅在推理阶段对50%的源文本嵌入与50%的目标文本嵌入进行融合引导图像编辑过程。

消融实验结果如图4所示,与MFE-Diffusion相比,Ours-J不能有效地重建原图像,如第1行中的场景花瓶底座和窗户都发生了很大变化,第2行的网球场场景发生了很大的变化,甚至图像的风格也无法复现。同样与MFE-Diffusion相比,Ours-M虽然可以有效地重建原图像,但不能准确地融合目标文本特征信息,例如在第2行中,生成的番茄更像是替换了颜色的网球,第3列小鸟的翅膀也不能很好地展开。

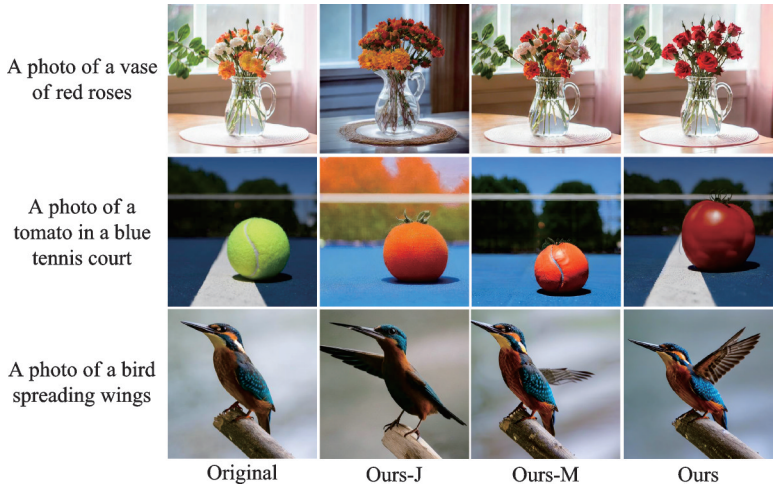

图4　消融实验比较结果

Fig.4　Ablation experiment results

综上所述,预训练扩散模型与特征向量联合学习框架和特征映射融合模块都有助于文本引导图像编辑任务。

2.5　关键参数效果分析

在本节中,将对MFE-Diffusion中的主要参数ε在不同数值设定下对实验结果的影响进行评估。如图5所示,ε作为特征映射融合模块中控制特征融合强度的参数,在ε较小时,编辑后的图片会保留更多的原图像和源文本特征;在ε较大时,编辑后的图像会更贴近目标文本描述,但会损失一些原图像中语义无关部分的保留效果。值得一提的是,ε的取值在0.4～0.7时,图像的编辑效果都是相当可观的,既可以根据目标文本描述操作原图像,又可以准确地保留与目标文本描述无关的特征,这也表明了所提方法具有一定的鲁棒性。

图5　关键参数ε效果分析

Fig.5　Effectiveness analysis of key parameter ε

本文还对ε进行了定量分析,考查了不同的ε取值(0.2、0.4、0.6、0.8),用来探索不同特征融合强度下模型性能的变化。将在不同ε情况下得到的定量指标IS分数和CLIP Score绘制折线图,如图6所示。

根据图6分析得到，较低的ε会导致模型保留了过多原图像信息，不能充分反映目标文本描述，而较高的ε则会导致生成图像的质量下降（如图中ε=0.8时，IS降低）。由于ε=0.6时，IS和CLIP Score同时达到峰值，即表明ε=0.6能够在图像质量和文本匹配度之间取得最优的平衡，因此在式(9)中选择0.6作为超参数ε值。

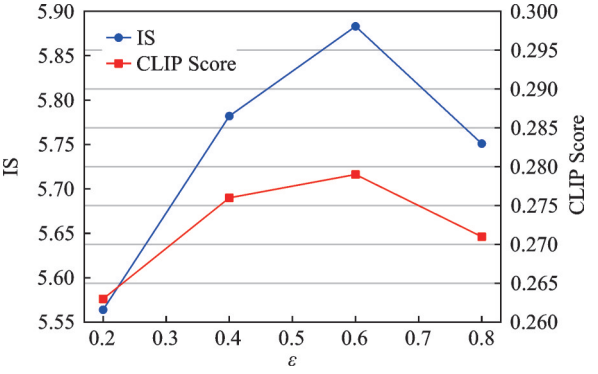

图6 不同ε下的定量指标分析

Fig.6 Analysis of quantitative indicators under different ε

3 结束语

针对以往基于微调大型预训练扩散模型的文本引导图像编辑方法，存在的既耗时又容易出现过拟合，同时编辑不准确的问题，提出了一种基于映射融合嵌入扩散模型的文本引导图像编辑方法(MFE-Diffusion)，设计了一个预训练扩散模型与特征向量联合学习框架和一个特征映射融合模块，使模型可以在快速重建给定的原图像的同时深度融合目标文本的特征信息。MFE-Diffusion在具有挑战性的文本引导图像编辑基准TEdBench上进行实验验证，在IS分数上超越对比方法Imagic 6.4%，在CLIP Score上超越对比方法Imagic 0.5%，结果表明所提方法在图像编辑性能上具有优势。下一步工作重点在于提高基座大型预训练扩散模型生成图像的质量和速度，并优化模型结构，减少时间复杂度，提高图像生成与编辑的效率。

参考文献：

[1] MENG C, HE Y, SONG Y, et al. SDEdit: Guided image synthesis and editing with stochastic differential equations[C]//Proceedings of International Conference on Learning Representations. [S.l.]: [s.n.], 2022.

[2] YANG S, WANG Z, LIU J, et al. Controllable sketch-to-image translation for robust face synthesis[J]. IEEE Transactions on Image Processing, 2021, 30: 8797-8810.

[3] ZHANG Y, HUANG N, TANG F, et al. Inversion-based style transfer with diffusion models[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2023: 10146-10156.

[4] 吴福祥，程俊. 基于自编码器生成对抗网络的可配置文本图像编辑[J]. 软件学报, 2022, 33(9): 3139-3151.
WU Fuxiang, CHENG Jun. Configurable text-based image editing by autoencoder-based generative adversarial networks[J]. Journal of Software, 2022, 33(9): 3139-3151.

[5] WANG J, LU G, XU H, et al. ManiTrans: Entity-level text-guided image manipulation via token-wise semantic alignment and generation[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2022: 10707-10717.

[6] TAO M, BAO BK, TANG H, et al. DE-Net: Dynamic text-guided image editing adversarial networks[C]//Proceedings of AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2023: 9971-9979.

[7] 刘雨生，肖学中. 基于扩散模型微调的高保真图像编辑[J]. 计算机应用, 2024(11): 3574-3580.
LIU Yusheng, XIAO Xuezhong. High-fidelity image editing based on fine-tuning of diffusion models[J]. Journal of Computer Applications, 2024(11): 3574-3580.

[8] ZHANG L, RAO A, AGRAWALA M. Adding conditional control to text-to-image diffusion models[C]//Proceedings of IEEE International Conference on Computer Vision. [S.l.]: IEEE, 2023: 3836-3847.

[9] TUMANYAN N, GEYER M, BAGON S, et al. Plug-and-play diffusion features for text-driven image-to-image translation [C]//Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2023: 1921-1930.

[10] BROOKS T, HOLYNSKI A, EFROS AA. Instructpix2pix: Learning to follow image editing instructions[C]//Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2023: 18392-18402.
[11] KAWAR B, ZADA S, LANG O, et al. Imagic: Text-based real image editing with diffusion models[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2023: 6007-6017.
[12] SAHARIA C, CHAN W, SAXENA S, et al. Photorealistic text-to-image diffusion models with deep language understanding [C]//Proceedings of Advances in Neural Information Processing Systems. [S.l.]: [s.n.], 2022, 36479-36494.
[13] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2022: 10684-10695.
[14] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[C]//Proceedings of Advances in Neural Information Processing Systems. [S.l.]: [s.n.], 2020: 6840-6851.
[15] SONG J, MENG C, ERMON S. Denoising diffusion implicit models[C]//Proceedings of International Conference on Learning Representations. [S.l.]: [s.n.], 2021.
[16] 赵睿卓,曲紫畅,陈国英,等. 大语言模型评估技术研究进展[J]. 数据采集与处理, 2024, 39(3): 502-523.
ZHAO Ruizhuo, QU Zichang, CHEN Guoying, et al. Research progress in evaluation techniques for large language models[J]. Journal of Data Acquisition and Processing, 2024, 39(3): 502-523.
[17] ZHOU C, ZHONG F, ÖZTIRELI C. CLIP-PAE: Projection-augmentation embedding to extract relevant features for a disentangled, interpretable and controllable text-guided face manipulation[C]//Proceedings of ACM SIGGRAPH 2023 Conference. [S.l.]: ACM, 2023: 1-9.
[18] HESSEL J, HOLTZMAN A, FORBES M, et al. CLIPScore: A reference-free evaluation metric for image captioning[EB/OL]. (2022-03-23). https://doi.org/10.48550/arXiv.2104.08718v3.
[19] SALIMANS T, GOODFELLOW I. Improved techniques for training GANs[C]//Proceedings of Advances in Neural Information Processing Systems. [S.l.]: [s.n.], 2016: 2234-2242.
[20] RADFORD A, KIM JW, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//Proceedings of ACM International Conference on Machine Learning. [S.l.]: ACM, 2021: 8748-8763.
[21] PODELL D, ENGLISH Z, LACEY K, et al. SDXL: Improving latent diffusion models for high-resolution image synthesis [EB/OL]. (2023-07-04). https://doi.org/10.48550/arXiv.2307.01952.

作者简介:

吴飞(1989-),**通信作者**,男,教授,研究方向:模式识别与机器学习,E-mail: wufei_8888@126.com。

马永恒(1999-),男,硕士研究生,研究方向:模式识别与人工智能,E-mail: 1003777847@qq.com。

邓哲颖(2000-),男,硕士研究生,研究方向:模式识别与人工智能。

王银杰(2004-),女,硕士研究生,研究方向:模式识别与智能控制。

季一木(1978-),男,教授,研究方向:模式识别与人工智能。

荆晓远(1971-),男,教授,研究方向:模式识别与人工智能。

(编辑:夏道家)