软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2022,33(9):3139−3151 [doi: 10.13328/j.cnki.jos.006622]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于自编码器生成对抗网络的可配置文本图像编辑*

吴福祥, 程 俊

(中国科学院 深圳先进技术研究院 广东省机器人与智能系统重点实验室, 广东 深圳 518055)

通信作者: 程俊, E-mail: jun.cheng@siat.ac.cn

摘 要: 基于文本的图像编辑是多媒体领域的一个研究热点并具有重要的应用价值. 由于它是根据给定的文本编辑源图像, 而文本和图像的跨模态差异很大, 因此它是一项很具有挑战的任务. 在对编辑过程的直接控制和修正上, 目前方法难以有效地实现, 但图像编辑是用户喜好导向的, 提高可控性可以绕过或强化某些编辑模块以获得用户偏爱的结果. 针对该问题, 提出一种基于自动编码器的文本图像编辑模型. 为了提供便捷且直接的交互配置和编辑接口, 该模型在多层级生成对抗网络中引入自动编码器, 该自动编码器统一多层级间高维特征空间为颜色空间, 从而可以对该颜色空间下的中间编辑结果进行直接修正. 其次, 为了增强编辑图像细节及提高可控性, 构造了对称细节修正模块, 它以源图像和编辑图像为对称可交换输入, 融合文本特征以对前面输入编辑图像进行修正. 在 MS-COCO 和 CUB200 数据集上的实验表明, 该模型可以有效地基于语言描述自动编辑图像, 同时可以便捷且友好地修正编辑效果.

关键词: 基于文本的图像编辑; 生成对抗网络; 交互编辑

中图法分类号: TP391

中文引用格式: 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. 软件学报, 2022, 33(9): 3139–3151. http://www.jos.org.cn/1000-9825/6622.htm

英文引用格式: Wu FX, Cheng J. Configurable Text-based Image Editing by Autoencoder-based Generative Adversarial Networks. Ruan Jian Xue Bao/Journal of Software, 2022, 33(9): 3139–3151 (in Chinese). http://www.jos.org.cn/1000-9825/6622.htm

Configurable Text-based Image Editing by Autoencoder-based Generative Adversarial Networks

WU Fu-Xiang, CHENG Jun

(Guangdong Provincial Key Laboratory of Robotics and Intelligent System, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China)

Abstract: Text-based image editing is popular in multimedia and is of great application value, which is also a challenging task as the source image is edited on the basis of a given text, and there is a large cross-modal difference between the image and text. The existing methods can hardly achieve effective direct control and correction of the editing process, but image editing is user preference-oriented, and some editing modules can be bypassed or enhanced by controllability improvement to obtain the results of user preference. Therefore, this study proposes a novel autoencoder-based image editing model according to text descriptions. In this model, an autoencoder is first introduced in stacked generative adversarial networks (SGANs) to provide convenient and direct interactive configuration and editing interfaces. The autoencoder can transform high-dimension feature space between multiple layers into color space and directly correct the intermediate editing results under the color space. Then, a symmetrical detail correction module is constructed to enhance the detail of the edited image and improve controllability, which takes the source image and the edited image as symmetrical exchangeable input to correct the previously input edited image by the fusion of text features. Experiments on the MS-COCO and CUB200 datasets demonstrate that the proposed model can effectively and automatically edit images on the basis of linguistic descriptions while providing user-friendly and

* 基金项目: 国家自然科学基金 (U21A20487); 深圳市基础研究项目 (JCYJ20200109113416531, JCYJ20180507182610734); 中国科学院关键技术人才项目

本文由“融合媒体环境下的媒体内容分析与信息服务技术”专题特约编辑汪萌教授、张勇东教授、俞俊教授以及张伟高级工程师推荐.

收稿时间: 2021-06-30; 修改时间: 2021-08-15; 采用时间: 2022-01-14; jos 在线出版时间: 2022-02-22

convenient corrections to the editing.

Key words: text-based image editing; generative adversarial networks (GANs); interactive editing

1 引 言

自动图像编辑[1−6]是基于给定条件对源图像在语义层面和几何层面进行编辑, 该任务具有广泛的应用前景. 如在融媒体分析下语言相关的图像处理, 可以提升大数据环境下融媒体热点信息采集处理效率, 此外, 也可以使用基于语言的图像编辑给文本配对相关图像, 能够增强融媒体的传播效果. 使用主流图像编辑工具处理图像要求用户有专业知识, 并且编辑图像需要非常繁琐且复杂细节操作, 这就给图像编辑带来较大的难度. 而自动图像编辑能够降低该难度, 从而吸引了很多研究者的目光, 且随着条件生成对抗网 (GANs)[7−10]的发展, 该任务已经取得很大的进步. 由于自然语言是一种用户友好的、便捷的、最自然的操作方式, 因此基于自然语言的自动图像编辑具有自然且高效的特点, 特别适用于便携设备等手动输入不便的场合.

目前基于文本的图像编辑模型 (流程如图 1) 一般使用堆栈式多层级 GANs [3,11−14], 其中每个 GAN 以前一个 GAN 输出的高维中间特征作为输入, 逐步编辑更高分辨率图像. 在该框架下, 很难对高维中间特征进行直接修改以修正和增强自动编辑效果, 然而由于文本模态和图像模态的巨大差距, 自动编辑效果很难直接满足用户需求, 因此在自动编辑模型中对中间编辑结果的直接修正具有非常重要的作用. 此外, 文本具有含义抽象且多义的特征, 因此会有很多图像符合文本描述并接近于源图像, 这些图像都作为理想的编辑结果, 这就需要用户介入以选择更加适合的结果. 同时, 由于跨模态编辑任务的复杂性, 自动编辑模型难以准确地根据文本语义编辑源图像, 这需要生成文本相关的特征并保留文本无关的特征, 这也需要用户交互编辑配置以辅助自动编辑模型生成更好结果. 因此, 在自动编辑过程中引入直接且便利的控制途径有益于生成更加高质量的编辑图像.

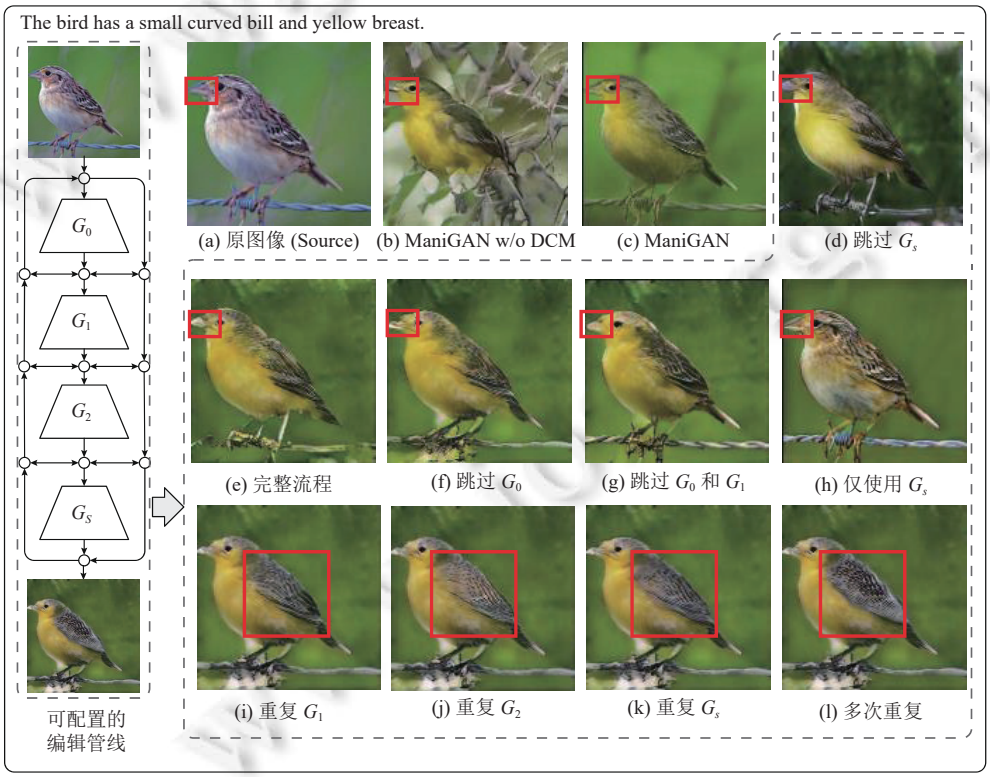


图 1 基于文本的可配置图像编辑流程

在图 1 中, 目标文本没有涉及鸟的翅膀, 因此编辑后图像中鸟的翅膀应该和源图像 (a) 一样, 具有清晰的灰暗色调纹理. 在 (b) 和 (c) 中, ManiGAN[3]编辑图像的翅膀比较模糊, 这是由于 ManiGAN 中有些模块对该源图像编辑性能不佳. 如图 1 左侧, 借助于自动编码器, 本文模型中的生成器可以划分为 4 个子生成器 G_0、G_1、G_2 和 G_s, 因此, 该模型可以重新配置处理流程, 根据模块对源图像的处理效果跳过或重复应用该模块, 其中, 圆圈表示输入图像上的用户定义的操作符. 如图 (e)–(h) 所示, 通过配置不同处理流程使编辑后的翅膀保留或增强其源图像的暗灰色调. 此外, 在图 1 示例中, 目标文本没有限定修改程度, 编辑结果 (d)–(o) 都是合理的, 因为它们都具有黄色的胸部和小嘴, 因此, 它们都可以提供给用户以选择其最喜欢的编辑结果.

为了提供高效的基于文本的自动图像编辑和直接修正该中间编辑结果的能力, 本文提出一种新颖的可配置图像编辑流程, 在多层级堆叠 GANs 上利用自动编码器统一其内部高维特征, 并隔离不同的生成阶段以支持重新执行或跳过某编辑模块. 本文模型主要使用仿射组合模块 (ACM) 构造生成器以根据文本特征修正视觉特征. 为了进一步提高可控性, 本文提出一个对称细节修正模块 (SDCM), 它以源图像和编辑后图像为对称输入, 融合文本特征对输入编辑图像进行修改及补充细节. 综上所述, 我们的工作贡献主要是提出了一个集成自动编码器的新型图像编辑框架, 以支持在每个编辑阶段直接对中间编辑结果进行修正和增强, 并可以构造一个可配置的辅助编辑流程. 在 MS-COCO 和 CUB200 数据集上的实验验证了本文模型能够有效地编辑图像, 并可以直接且友好地修正和增强编辑效果.

2 国内外研究现状

近年来, 条件生成对抗网络[9]的研究推动了自动图像编辑的发展. 在本节中, 我们回顾最近相关工作, 并讨论本文工作与这些工作间的区别. 在文本生成图像任务中, 由于高分辨率图像生成是一个复杂的任务, 为了降低该任务的复杂度, 构造多层级堆叠 GANs[15,16]将该高分辨生成任务分解为几个相对容易的任务: 生成低分辨图像与逐步细化以生成更高分辨图像. 在多模态细粒度特征关联上, 目前常用双线性注意力模型[17−21]能够同时考虑两个模态间的特征交互, 从而提高了多模态特征提取效果. 考虑到图像不同部分会和句子中不同词关联, Xu 等人[11]和 Zhu 等人[22]利用细粒度注意力模型来提高生成图像的质量. 为了加强对文本描述的理解, Chen 等人[23]提出了 RiFeGAN 来补充给定的文本描述的信息. Qiao 等人[24]通过使用生成图像重新生成文本描述, 引入 MirrorGAN 来提高语义一致性.

在基于属性的编辑任务上, Collins 等人[25]提出基于 StyleGAN 学习解耦对象特征的语义编辑方法以高质量地修改图像. Liu 等人[1]提出了基于图像到图像转换的无监督方法以分开图像的内容与属性, 并通过命令式文本描述改变属性, 进一步通过属性修改图像. Dorta 等人[26]提出一种借助平滑场和非配对样本训练的语义图像编辑模型来编辑高分辨率脸图像. Lang 等人[27]提出利用标记引导注意力和相似性约束机制来转换跨类别图像的 DesignGAN. Li 等人[28]提出使用基于如自然语言等因素来控制图像转换图像的方法. Liang 等人[29]通过将对抗对比目标函数引入对比 GAN 来支持在高层语义和低层语义上编辑图像. Chen 等人[30]提出将人脸区域分割成多个语义组件来提供有效的精细交互修改方法. Dhamo 等人[4]通过改变其语义节点或边来编辑场景图, 从而编辑其对应的图像. Tang 等人[31]提出基于 GAN 的统一图像转换模型, 可以根据源图像和可控结构来转换编辑该图像.

在基于文本的图像编辑任务上, Nam 等人[13]提出基于文本自适应判别器的 TAGAN, 它在保留源图像中与文本无关内容的同时根据文本对图像进行编辑. Chen 等人[32]介绍了一个通用的基于语言分割和着色的递归注意力框架来融合源图像的视觉特征和语言特征. Zhang 等人[33]提出基于双重多模态注意机制和图像文本匹配损失的 TDANet 以补充基于文本描述的缺失部分. Li 等人[34]提出基于空间和通道注意力驱动的生成器和词级判别器构造的 ControlGAN, 能够有效地编辑和控制图像合成. Zhou 等人[35]提出一个基于文本的姿势生成和外观特征迁移的两阶段方法来编辑人物的姿势和属性. Cheng 等人[36]建立基于神经网络状态跟踪和顺序注意力机制 SeqAttnGAN 以对图像进行交互和多轮编辑. Liu 等人[2]使用词级和指令级的指令编码器和推理判别器来提高图像和语言之间的一致性. Li 等人[3]提出基于文本图像仿射组合模块和细节修正模块的 ManiGAN 来根据给定文本编辑图像.

类似于 ManiGAN[3], 本文模型也包含主编辑模块和细节修正模块以提升高层语义编辑能力. 考虑到难以对中

间编辑效果的直接干预, 我们在编辑流程中引入自动编码器以统一不同层级间的中间高维特征, 可以构造可配置的编辑流程以支持直接精细的交互式修正. 此外, 本文进一步提出对称细节修正模块 (SDCM) 以提高可控性, 它融合文本特征对前面修改图像进行进一步修正和细节增强.

3 模　型

基于文本的图像编辑是根据目标文本描述 T 在语义和几何层面上编辑源图像 I. 在图 2 中, 我们构建了一个基于自动编码器的主编辑模块, 能够根据 T 生成一个初步的编辑图像 $\hat{I}_2$, 其中灰色方框为上采样模块, 包含一个上采样层、3×3 卷积层和实例归一化层, 而黄色方框为残差网络层. 而后, 对称细节修正模块 (SDCM) 进一步根据源图像 I 和文本描述 T 来增强 $\hat{I}_2$ 的细节.

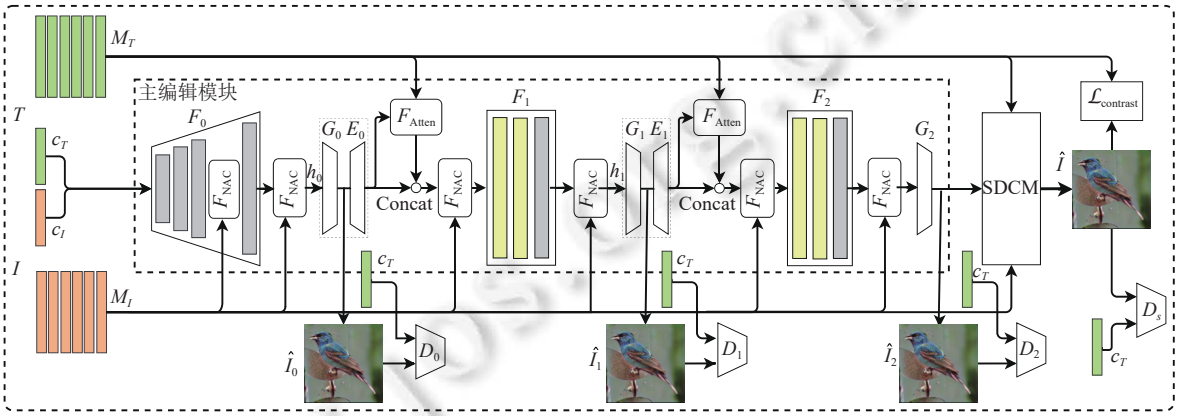


图 2　基于文本的交互可配置图像编辑模型结构

3.1 基于自动编码器的主编辑模型

在图 2 中, 类似于 AttnGAN[11]、ControlGAN[34]和 ManiGAN[3], 我们采用多层级堆栈 GANs. 给定源图像 I, 本文利用 VGG 来提取区域特征 $M_I \in R^{N_0 \times W \times H}$ 和总特征 $C_I \in R^{N_0}$, N_0 为内部特征维度. 对于目标文本描述 T, 采用 LSTM-RNN 模型来获得词嵌入 $M_T \in R^{256 \times |T|}$ 和句子特征 $c_T \in R^{256}$. 在图 2 中, 第一级生成器 F_0 以 c_T、c_I 和 M_I 为输入, 经过一个线性层获得 4×4 高维特征, 而后经过 2 个由上采样层、3×3 卷积层和实例归一化组成的上采样模块获得 16×16 高维特征, 再通过 F_{NAC} 增强内部特征, 最后再经过 2 个上采样模块生成内部高维特征 $h_{F0} \in R^{N_0 \times 64 \times 64}$. 然后, 我们进一步利用 F_{NAC} 来融合视觉特征 M_I:

$$h_i = F_{\text{NAC}}(h_{Fi}, M_I),\ i = 0, 1, 2 \tag{1}$$

其中, F_{NAC} 是一个带噪声仿射组合模块,

$$F_{\text{NAC}}(h, M) = h \odot W_{\text{rnd}}(M) + b_{\text{rnd}}(M) \tag{2}$$

其中, 函数 $W_{\text{rnd}}(M)$ 和 $b_{\text{rnd}}(M)$ 通过区域特征 $M \in R^{N_0 \times W \times H}$ 借助线性网络计算权重和偏差, 并且额外在特征中加入高斯噪声, 以使网络更多地关注文本模态, $\odot$ 是 Hadamard 元素积. 初始编辑图像 $\hat{I}_0$ 是由 G_0 生成的. 为了支持对 $\hat{I}_0$ 的直接修改来改变最终的编辑结果, 我们利用编码器 E_0 来恢复内部高维视觉特征 $h_0 \in R^{N_0 \times 64 \times 64}$. 因此, E_0 和 G_0 为一个自动编码器, 用于恢复 h_0 特征信息. 我们对每一层级都引入自动编码器 G_i 和 E_i, 其中 G_i 由 3×3 卷积层和 tanh 激活函数组成, E_i 包含 atanh 函数、3×3 卷积层、Leaky ReLU 层和实例规范化层. 此外, 在测试中, 由 G_i 和 E_i 组成的自动编码器将中间编辑后的图像带到主要生成流程中, 且隔离不同的生成阶段. 而后, 编码器 E_i 的输出用注意力模块 F_{Atten} 和 F_{NAC} 来增强, 以融合文本和图像模态中的特征, 其中 F_{Atten} 为 ManiGAN[3]中定义的空间和通道注意力模型. 编码器 E_i 生成的特征 $\hat{h}_i \in R^{N_0 \times W_i \times H_i}$, 其中通道注意力特征定义为 $\hat{h}_i \cdot \alpha$, 其中注意力 $\alpha \in R^{|T| \times W_i \times H_i}$ 算为:

$$\alpha_{i,j} = \exp(\alpha'_{i,j}) \Big/ \sum_{i=1}^{|T|} \exp(\alpha'_{i,j}) \tag{3}$$

其中, $\alpha' = M_T^{\mathrm{T}} \cdot M_W \cdot \hat{h}_i \in R^{N_0 \times W_i \times H_i}$, $M_W \in R^{256 \times N_0}$ 为变换矩阵, 将词特征变换到视觉特征空间. 类似地, 空间注意力[11]引导生成器在处理不同空间区域时注意不同的词特征.

本文在生成器的目标函数中引入了自动编码器的正则化项, 其计算为:

$$L_G^m = L_G + \lambda_0 \frac{1}{N} \sum_i^N l_i^{AE} \tag{4}$$

其中, L_G 是 ControlGAN[34]中定义的生成器目标函数, N 是多层级的级数. L_i^{AE} 为层间的自动编码器损失函数:

$$L_i^{AE} = \|E_i(G_i(h_i)) - h_i\|_2, \ i = 0, 1 \tag{5}$$

判别器 $\boldsymbol{D}^m$ 的损失函数为:

$$L_D^m = E_{\hat{I} \sim P_G(I,T)} \left(\sum_i \log D_i^U(\hat{I}) + \log D_i^F(\hat{I}|T_I) \right) - E_{I \sim P_{data}} \left(\sum_i \log D_i^F(I|T_I) + \log D_i^U(I) \right) \tag{6}$$

其中, D_i^F 包括判别器 $D_i \in \boldsymbol{D}^m$ 的有条件和无条件的判别计算, 而 D_i^U 只包含无条件的计算. 在训练过程中, 图 2 中主模块 G_{Main} 可以分成 3 部分: G_0^{Main}, 包括 $\hat{I}_0$ 之前的网络层; G_1^{Main}, 包括 $\hat{I}_0$ 和 $\hat{I}_1$ 之间的网络层; G_2^{Main}, 其余层. 由于这 3 个模块被 $\hat{I}_0$ 和 $\hat{I}_1$ 隔离, 因此给定一个源图像 I, 它们可以通过替换成相应分辨率下的真实图像而被跳过:

$$\begin{cases} G_0^{\mathrm{Main}} = G_2^{\mathrm{Main}}\left(G_1^{\mathrm{Main}}(I_0, C), C\right) \\ G_{I_1}^{\mathrm{Main}} = G_2^{\mathrm{Main}}(I_1, C) \end{cases} \tag{7}$$

其中, $C=\{c_T, c_I, M_T, M_I\}$, I_0 和 I_1 是从 I 缩放得来的低分辨率真实图像. 因此, 在训练中, 我们将随机选择一种生成器, 用真实图像替换中间生成图像, 能够减轻错误传播并加速训练.

3.2 对称细节修正模型

在图 3 中, 为了更好实现交互配置式编辑并提供额外的可控性, 细节修正模块对称和可交换地处理前面的编辑图像和给定的源图像, 以支持交互配置式的交换或替换其中的输入图像. $M_I \in R^{N_0 \times W \times H}$ 和 $M_{\hat{I}_2} \in R^{N_0 \times W \times H}$ 是 VGG 从 $\hat{I}_2$ 和 I 中提取的区域特征. 与 StyleGAN[37]类似, F_{NAC} 通过利用这些区域特征 M_I 和 $M_{\hat{I}_2}$ 作为风格特征, 从固定的随机噪声中合成视觉特征, 而 F_{WAtten} 类似于 AttnGAN 中的注意模型, 用词嵌入 M_T 增强转换后的特征. 在图 3 中, 虚线框内的 F_{NAC} 输出的两个特征 $x_I \in R^{64 \times W \times H}$ 和 $x_{\hat{I}_2} \in R^{64 \times W \times H}$ 通过一个 multiplexer 模块 F_{fuse} 融合:

$$F_{\mathrm{fuse}}\left(x_I, x_{\hat{I}_2}\right) = F_{\mathrm{residual}}\left(\frac{1}{2}\left(x_i \odot \beta_1 + x_{\hat{I}_2} \odot \beta_2\right)\right) \tag{8}$$

其中, F_{residual} 是残差网络, β_1 和 β_2 是由 multiplexer 函数[38]计算的注意力权重, 最后生成器 G_S 将融合后的特征转化为最终的图像 $\hat{I}$. 从上式可以看到, 模块 F_{fuse} 可以通过计算 β_1 和 β_2 控制输入特征中 x_I 和 $x_{\hat{I}_2}$ 的比例, 因此该模块主要是判断输出结果中需要保留基于主模块的进一步编辑结果还是基于真实图像的编辑结果, 当主模块编辑结果较差时, 该模块能够赋予较低权重, 从而缓解编辑错误传递, 进而提升了编辑效果.

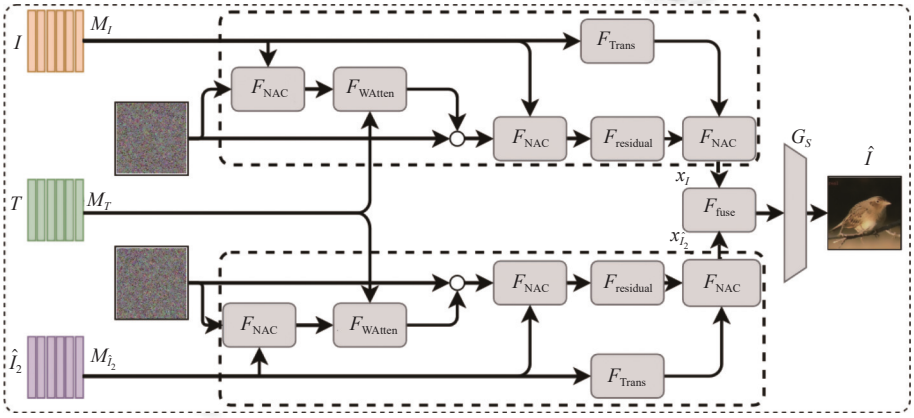


图 3 对称细节修正模块

给定源图像 I、目标文本 T 和与图像 I 相匹配的文本 T_I, 由于对应于 T 的编辑图像是未知的, 所以很难根据 T 来直接监督模型训练. 然而, 我们可以采用无条件的对抗性损失来引导编辑的合成, 要求编辑后图像尽可能逼真以增强编辑效果. 因此, 生成器 G_S 的损失为:

$$L_{Gs}=-E_{\hat{I}\sim P_{Gs}(I,T)}\left(\log D_S^U\left(\hat{I}\right)+\log D_S^F\left(\hat{I}|T_I\right)+L_{\text{constrast}}\left(\hat{I},T,T_I\right)\right)+L_{\text{ControlGAN}}+L_{\text{reg}}\left(\hat{I},I\right) \tag{9}$$

其中, D_S^F 包括判别器 D_S 的有条件和无条件的判别计算, 而 D_S^U 只包含无条件的计算, $L_{\text{ControlGAN}}$ 和 L_{reg} 类似于 ManiGAN 中的损失函数为:

$$L_{\text{ControlGAN}}=L_{\text{DAMSM}}-L_{\text{corre}}\left(\hat{I},T\right)+L_{\text{reg}}\left(\hat{I},I\right) \tag{10}$$

其中, 函数 $L_{\text{reg}}\left(\hat{I},I\right)$[34]定义负的图像 $\hat{I}$ 和 I 的像素平均偏差; 关联函数 $L_{\text{corre}}\left(\hat{I},T\right)$ [34]定义为:

$$L_{\text{corre}}\left(\hat{I},T\right)=\sum_i Sigmoid\left(\frac{(\gamma_i b_i)^{\mathrm{T}}M_{T,i}}{\|\gamma_i b_i\|\left\|M_{T,i}\right\|}\right) \tag{11}$$

其中, $b_i\in R^{256\times|T|}$ 为第 i 个词对应的视觉特征, 通过注意力方式计算为 $M'_{\hat{I}_2}\cdot\exp\left(M'^{\mathrm{T}}_{\hat{I}_2}\cdot M_T\right)\Big/C_{\text{norm}}$, 其中 $M'_{\hat{I}_2}\in R^{256\times W\times H}$ 为 $M_{\hat{I}_2}$ 通过线性层投影到词特征空间中. $L_{\text{constrast}}\left(\hat{I},T,T_I\right)$ 是语义对比函数, 目的是使合成的图像 $\hat{I}$ 相对于 T 更接近于 T_I. $L_{\text{constrast}}\left(\hat{I},T,T_I\right)$ 定义为:

$$L_{\text{constrast}}\left(\hat{I},T,T_I\right)=\max\left(L_{\text{corre}}\left(\hat{I},T\right)-L_{\text{corre}}\left(\hat{I},T_I\right)+\rho_c,0\right) \tag{12}$$

其中, ρ_c 是对比度控制阈值. 同样地, 判别器 D_S 的损失函数为:

$$L_{Ds}=E_{\hat{I}\sim P_{Gs}(I,T)}\left(\log D_S^U\left(\hat{I}\right)+\log D_S^F\left(\hat{I}|T_I\right)\right)-L_{\text{corre}}(I,T_I)+L_{\text{corre}}(I,T)-E_{I\sim P_{data}}\left(\log D_S^F(I|T_I)+\log D_S^U(I)\right) \tag{13}$$

其训练与主编辑模型的训练策略类似, 我们对模块 G_{SDCM} 随机地用 I 替换 $\hat{I}$ 以加速训练.

4 实验结果

4.1 数据集

本文采用包含复杂场景的 MS- COCO[39]和 Caltech-UCSD Birds-200-2011 (CUB200)[40]数据集进行评测. 在 MS-COCO 数据集中, 我们采用 2014 年评测的数据集划分方式, 该划分中训练部分包含大约 80 000 张图片, 测试部分则包含 40 000 张图片. 该数据集包含 80 种不同种类的物体, 且每张图片具有 5 个文本描述. 而 CUB200 数据集包含 200 个类别, 包含有 11 788 张图片, 其中每张图片有 10 个标题来描述其细粒度特征. 在 CUB200 数据集中, 类似于[3,11,16]工作, 我们使用类不交叉的数据划分方式.

4.2 评价指标

尽管 Inception Score 评分 (IS)[41]有些缺陷[42], 但由于其倾向于有意义且多样化的图像, 因此可以较好地评估合成图像的质量, 本文采用 Inception 模型来评估模型性能. 除了 Inception Score, 类似于 Li 等人[3], 我们采用编辑精度指标 (MP) 来衡量编辑图像中包含的目标文本视觉特征质量和源图像中原始细节重建质量. 例如, 高的 MP 意味着高的文本-图像相似度和低的编辑变化. 此外, 类似于 Xu 等人[11], 我们采用 R-precision 精度来衡量目标文本和相应编辑图像之间的文本-图像相似度, 用以验证编辑效果. 具体来说, 给定一个目标文本和随机选择的 99 个不匹配文本, 只有当目标文本特征和图像全局特征的余弦相似度高于其他文本时, 检索才是相关的, 而 R-precision 精度是所有检索标题中相关的比率.

4.3 实验结果

本文构建了 3 个模型: Our_{main} 表示不包含 SDCM 模块的模型; $\text{Our}_{\text{w/oDU}}$ 表示没有借助无条件 D_S^U 训练的模型, Our_{all} 则表示完整的模型. 基线模型是 TAGAN[13]和 ManiGAN[3], 此外我们也添加 SISGAN 模型[3,43]评分. 表 1 描述了在 CUB200 和 MS-COCO 数据集中的 MP 和 IS, 其中, 为了公平地和基线进行比较, 我们在测试集中随机选择了目标文本. 与 SISGAN, TAGAN 和 ManiGAN 相比, 本文模型 MP 在 CUB200 中增加了 7.53%, 在 MS-COCO

中增加了 35.36%. 在 CUB200 中, Our_{all} 的 IS 为 5.06, 比 TAGAN 的 IS 低 0.31, 这是由于编辑程度高后的鸟图像类别可能会远离测试数据对应的类别, 而 CUB200 的 IS 指标评价模型于测试类型进行微调, 因此评分会相应偏低些, 这也是比 TAGAN 效果好的模型 ManiGAN 在 IS 指标上偏低的原因. 然而, Our_{all} 的 IS 比 SISGAN 高, 且 Our_{all} 的 MP 比 TAGAN 的 MP 高 11.60%. 此外, Our_{all} 的 MP 和 IS 都高于 ManiGAN, 这表明我们的模型可以更有效地编辑图像, 同时保持其质量. 在 MS-COCO 中, Our_{all} 的 IS 高于基线, 并比 ManiGAN 增加了 2.07. Our_{all} 的 MP 为 39.29%, 与基线相比增加了 35.36%, 这表明我们的模型也可以有效地处理复杂场景下的图像.

表 1　在 CUB200 和 MS-COCO 数据集上的 MP (%)/Inception Score 评分

Dataset	SISGAN	TAGAN	ManiGAN	Our_{main}	$\text{Our}_{\text{w/oDU}}$	Our_{all}
CUB200	2.2/2.24	15.86/**5.37±0.06**	20.95/4.36±0.37	26.77/4.70±0.23	**28.49**/4.88±0.20	27.46/5.06±0.22
MS-COCO	4.2/3.44	N/A	3.93/23.78±0.53	31.01/19.29±0.36	37.02/19.35±0.17	**39.29/25.85±0.43**

在没有使用 SDCM 的情况下, 相对于 Our_{all} 和 $\text{Our}_{\text{w/oDU}}$, Our_{main} 的 MP 在 CUB200 中下降了 0.69%, 在 MS-COCO 中下降了 8.28%, 然而, 得分仍然高于基线. Our_{main} 的 IS 在 CUB200 中减少了 0.36, 在 MS-COCO 中减少了 6.56. 因此, 结果表明 SDCM 可以提高合成和编辑的质量. 此外, 在没有无条件对抗 D_S^U 的情况下, $\text{Our}_{\text{w/oDU}}$ 的 MP 在 CUB200 中比 Our_{all} 高 1.02%, 然而在 MS-COCO 中比 Our_{all} 低 2.27%. 此外, 与 Our_{all} 相比, $\text{Our}_{\text{w/oDU}}$ 的 IS 减少了 6.5, 这表明了无条件对抗损失是有益于提升编辑质量.

在表 2 中, 本文模型的 R-precision 精度高于基线模型. 相对于基线, Our_{main} 的 R-precision 精度在 CUB200 和 MS-COCO 中分别提高了 38.54% 和 90.91%. 此外, 模块 SDCM 可以在 CUB200 中提高 R-precision 精度 2.65%, 在 MS-COCO 中提高 R-precision 精度 19.93%. 此外, 在没有无条件对抗损失 D_S^U 的情况下, $\text{Our}_{\text{w/oDU}}$ 的 R-precision 精度在 CUB200 中下降了 0.63%, 在 MS-COCO 中下降了 17.73%. 因此, 结果表明本文模型具有较好的编辑能力.

表 2　在 CUB200 和 MS-COCO 数据集上的 R-precision 评分 (%)

Dataset	TAGAN	ManiGAN	Our_{main}	$\text{Our}_{\text{w/oDU}}$	Our_{all}
CUB200	9.57±0.52	57.04±0.51	92.93±0.38	94.95±0.38	**95.58±0.36**
MS-COCO	N/A	0.18±0.2	71.17±0.56	73.37±0.49	**91.10±0.46**

4.4　实验分析

在图 4 第 1 行的例子中, 给定 Source 源图像和其上的目标文本, 相对于基线, 本文模型 Our_{all} 编辑的图像包含更清晰且符合“brown wings and crown”和“white belled”的鸟. 通过将在 64×64 和 128×128 分辨率下的中间编辑图像替换为真实图像, $\text{Our}_{\text{all, r0}}$ 和 $\text{Our}_{\text{all, r1}}$ 提供更加细腻的编辑结果, 不过几何层面的修改将变弱; 而 $\text{Our}_{\text{all, r2}}$ 是 256×256 分辨率下的图像替换, 即跳过了主编辑模块, 可以看到该编辑结果保留了更多的源图像细节特征, 但编辑效果变弱了些; $\text{Our}_{\text{all, inv}}$ 交换 SDCM 的对称输入, 其结果介于 $\text{Our}_{\text{all, r2}}$ 和 $\text{Our}_{\text{all, r1}}$ 之间, 提供了更多可选编辑结果. 因此, 该图像替换也验证了图 1 中可配置流程的可行性和有效性. 在 MS-COCO 中, 给定目标文本“几个人站在森林里的一辆灰色汽车旁边”, Our_{all} 也提供了比基线更符合语义的细节.

在图 5 中, 给定同一幅 Source 源图像, 我们给定不同的目标编辑文本, 可以看到本文模型可以根据目标文本编辑图像同时保留文本无关细节. 如左上角蓝色的鸟, 可以根据目标文本“brown body”和“grey belly”等抽象特征生成相应视觉特征的图像. 此外, 直接使用 SDCM 模块, $\text{Our}_{\text{all, r2}}$ 可以轻度编辑图像, 这也说明需要主编辑模块以进一步根据目标文本修正源图像.

(1) 编辑模块复用

在图 6 中, 下标 G1、G2 和 Gs 分别表示重复 G_1^{Main}、G_2^{Main} 和 G_{SDCM} 所得到的编辑结果; “more”表示重复模块 G_{SDCM} 两次; “rp2”表示重新应用 2 次. 在图 6 的 CUB200 数据集中的鸟, 我们能够选择增强翅膀的细节, 并保留翅膀在源图像中丰富和暗色调的细节. 如通过重新应用 G_{SDCM} 模块, 鸟翅膀具有更加明显的细纹路, 而进一步重复, $\text{Our}_{\text{all, More}}$ 提供更加突出的特征. 对于 MS-COCO 数据集中的复杂场景也具有类似的情况, 且我们可以进一步应用了背景掩码来约束编辑, 从而比较自然地保留钟楼的细节.

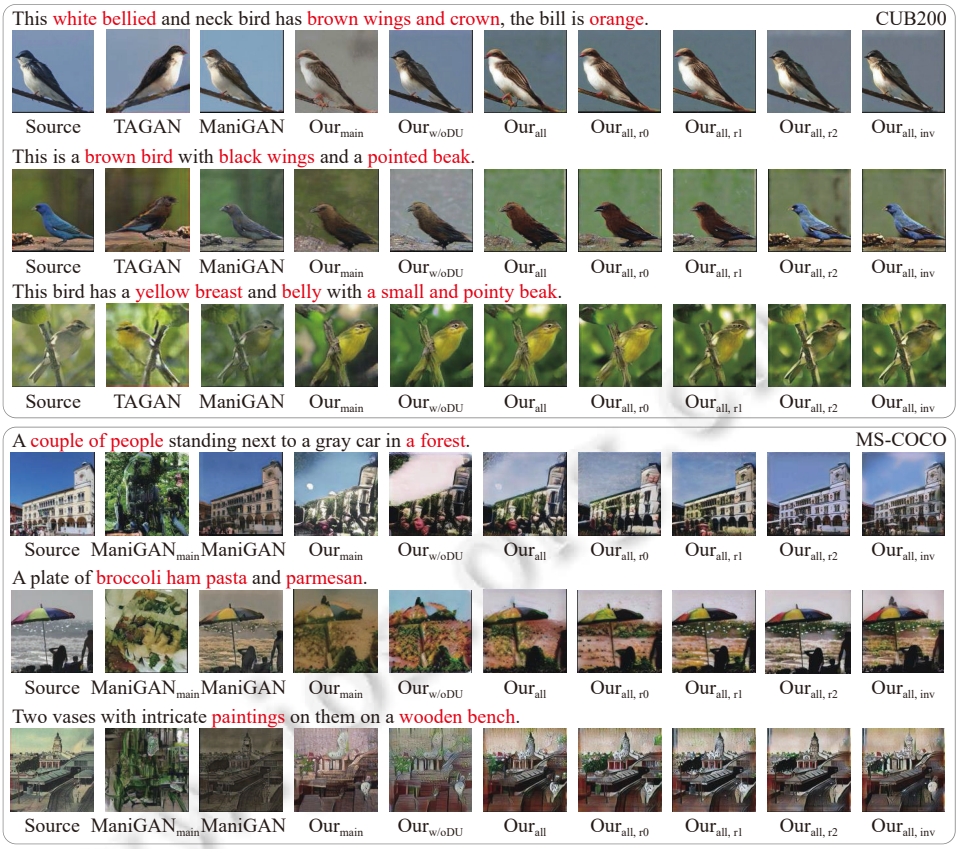


图 4 编辑示例: 图像上面是目标编辑文本

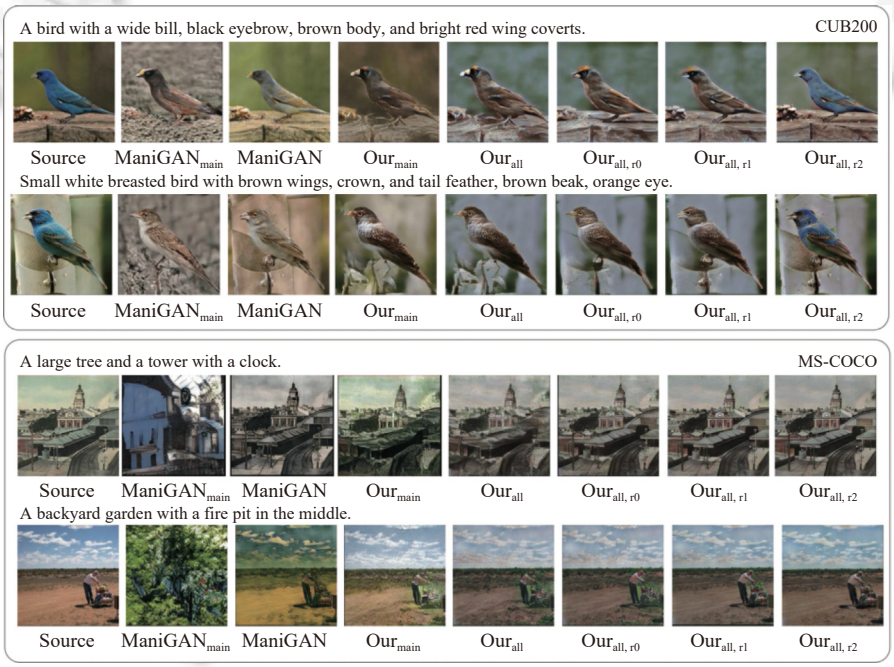


图 5 使用不同目标编辑文本的编辑示例

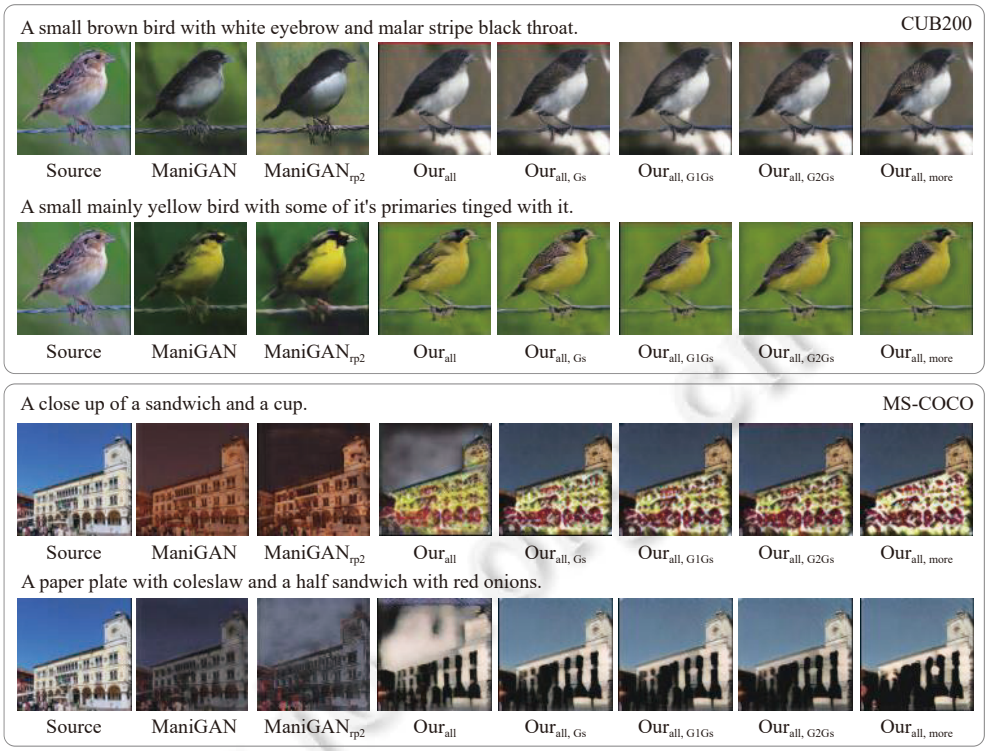


图 6　重复和跳过编辑模块的编辑示例

通过图 5 和图 6 中跳过和重复编辑模块的示例可以看出, 中间编辑图像处于主要编辑流程中, 改变它们可以影响最终编辑结果, 且该中间图像是在共同的 RGB 空间中, 这意味着编辑流程可以像决策树中的方式一样, 将训练过不同的模型 G_0^{Main} 、G_1^{Main} 、G_2^{Main} 和 G_{SDCM} 的不同模块结合起来, 以可交互方式获得更多的多样化和高质量的图像, 下面我们将进一步验证模型的可配置性.

(2) 配置编辑流程

为了进一步验证模型的可配置性, 不失一般性, 我们以对鸟翅膀和脚部的限定为例, 构造了如下的编辑流程以自动控制不同的编辑效果.

在图 7 中, 不同生成器配置代表不同的图像编辑处理流程. 其中, 噪声图像 I_{random} 每次都重新生成, 点“·”表示张量的元素乘积. 配置 (a) 是原始设置; 配置 (b*) 和 (c*) 表示在不同的约束强度下保留脚、翅膀纹理和背景细节, 同时保持语义一致性; 配置 (d*) 表示通过在中间编辑图像中简单地注入噪声来增强翅膀的细节.

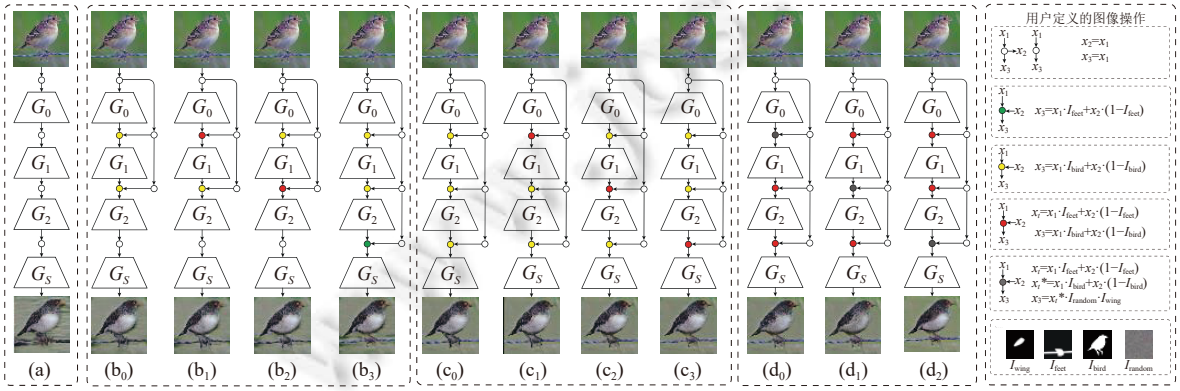


图 7　不同配置和用户定义图像运算符的示例

图 8 是通过图 7 中不同的生成器配置来自动编辑操作的示例, 其中, $Our_{all}(a)$ 表示在图 7 的配置 (a) 下, 通过使用模型 Our_{all} 并根据给定的目标文本来编辑左边的源图像. 从编辑结果可以看到, $Our_{all}(b_2)$ 右边列可以保留鸟脚部及铁丝的细节, 而最终结果可以根据用户喜好来选择不同列相应的编辑图像. 因此, 类似图 7 的配置流程可以灵活地控制编辑过程, 并能够提供更加好的编辑效果.

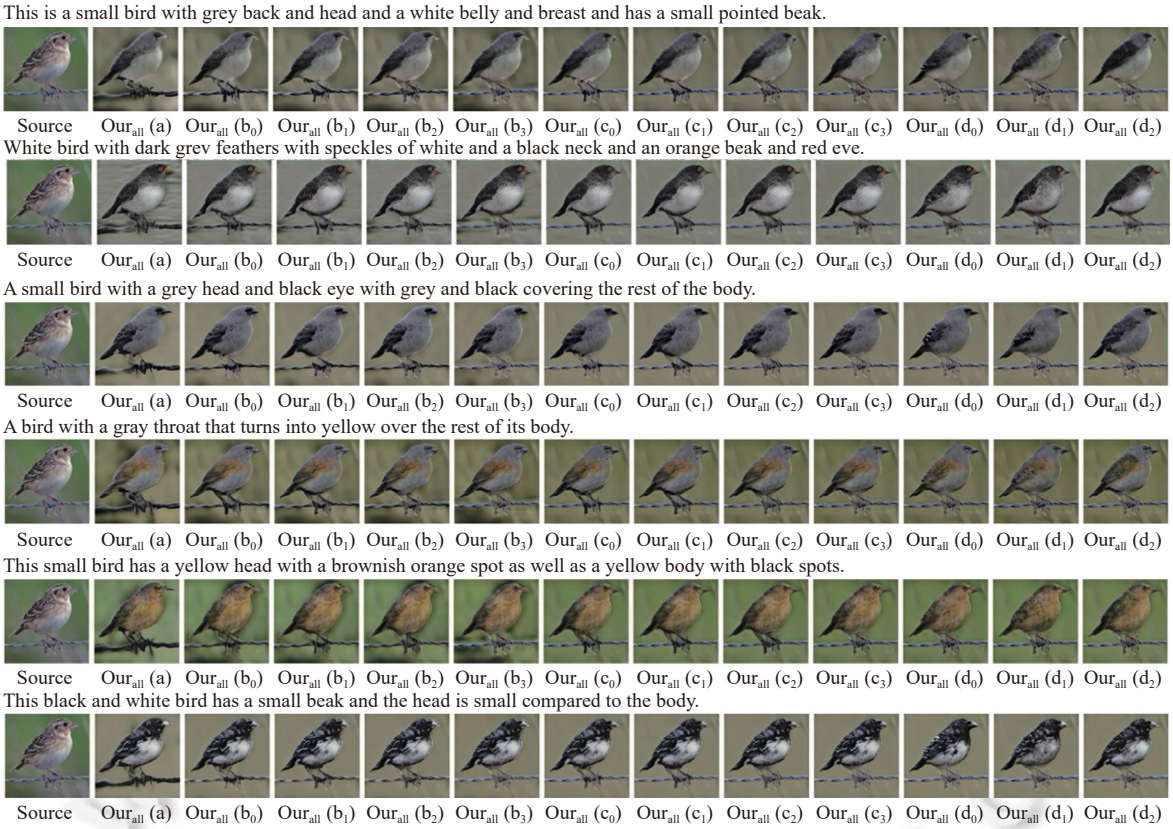


图 8　使用图 7 中配置流程的编辑示例

4.5　限制与讨论

本文构造了基于自动编码器的多层级 GANs 以将中间编辑图像引入主要的编辑流程中, 从而能够直接编辑它们以修正最终的编辑结果, 这提供了更强的和友好的编辑可控性. 在本文中, 我们利用纯自动编码器来保留输入特征的潜在信息和标准的 GAN 框架来训练模型. 但是, 我们可以利用如超分辨率算法等更好的算法和结构以生成更好的文本相关视觉特征, 从而进一步改善编辑结果. 此外, 如语义句子嵌入[44]等更好的文本编码器能够提高语义一致性, 因此, 我们可以利用这些复杂的编码器来提高编辑效果. 且如 BigGANs[45]等更复杂的模型能够提高合成质量, 我们后面将进一步改进该模型.

5　结　论

为了构造可配置且高效的基于文本图像编辑, 我们提出了一种新颖的交互配置式图像编辑模型, 该模型由基于自动编码器的多层级堆叠 GANs 构建的主要编辑模块和补充合成图像细节的对称细节修正模块组成. 通过统一不同层级间的内部高维特征为图像空间, 能够直接编辑该中间图像以修正最终的编辑效果, 因此可以支持可配置的编辑流程. 在 CUB200 和 MS-COCO 数据集中的实验表明, 本文模型能够在语义层面和几何层面有效地编辑图像, 并可支持用户友好及可配置的交互编辑修正.

References:

[1] Liu YH, De Nadai M, Cai D, Li HY, Alameda-Pineda X, Sebe N, Lepri B. Describe what to change: A text-guided unsupervised image-to-image translation approach. In: Proc. of the 28th ACM Int'l Conf. on Multimedia. Seattle: ACM, 2020. 1357–1365. [doi: 10.1145/3394171.3413505]

[2] Liu ZH, Deng JC, Li L, Cai SF, Xu QQ, Wang SH, Huang QM. IR-GAN: Image manipulation with linguistic instruction by increment reasoning. In: Proc. of the 28th ACM Int'l Conf. on Multimedia. Seattle: ACM, 2020. 322–330. [doi: 10.1145/3394171.3413777]

[3] Li BW, Qi XJ, Lukasiewicz T, Torr PHS. ManiGAN: Text-guided image manipulation. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 7877–7886. [doi: 10.1109/CVPR42600.2020.00790]

[4] Dhamo H, Farshad A, Laina I, Navab N, Hager GD, Tombari F, Rupprecht C. Semantic image manipulation using scene graphs. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 5212–5221. [doi: 10.1109/CVPR42600.2020.00526]

[5] Bau D, Strobelt H, Peebles W, Wulff J, Zhou BL, Zhu JY, Torralba A. Semantic photo manipulation with a generative image prior. ACM Trans. on Graphics, 2019, 38(4): 59. [doi: 10.1145/3306346.3323023]

[6] Liu M, Ding YK, Xia M, Liu X, Ding E, Zuo WM, Wen SL. STGAN: A unified selective transfer network for arbitrary image attribute editing. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3673–3682. [doi: 10.1109/CVPR.2019.00379]

[7] Mirza M, Osindero S. Conditional generative adversarial nets. arXiv: 1411.1784, 2014.

[8] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2014. 2672–2680.

[9] Reed S, Akata Z, Yan XC, Logeswaran L, Schiele B, Lee H. Generative adversarial text to image synthesis. In: Proc. of the 33rd Int'l Conf. on Machine Learning. New York: JMLR.org, 2016. 1060–1069.

[10] Chen FJ, Zhu F, Wu QX, Hao YM, Wang ED, Cui YG. A survey about image generation with generative adversarial nets. Chinese Journal of Computers, 2021, 44(2): 347–369 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2021.00347]

[11] Xu T, Zhang PC, Huang QY, Zhang H, Gan Z, Huang XL, He XD. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 1316–1324. [doi: 10.1109/CVPR.2018.00143]

[12] Shaham TR, Dekel T, Michaeli T. SinGAN: Learning a generative model from a single natural image. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 4569–4579. [doi: 10.1109/ICCV.2019.00467]

[13] Nam S, Kim Y, Kim SJ. Text-adaptive generative adversarial networks: Manipulating images with natural language. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montreal: Curran Associates Inc., 2018. 42–51.

[14] Yang WX, Yan Y, Chen S, Zhang XK, Wang HZ. Multi-scale generative adversarial network for person re-identification under occlusion. Ruan Jian Xue Bao/Journal of Software, 2020, 31(7): 1943–1958 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5932.htm [doi: 10.13328/j.cnki.jos.005932]

[15] Zhang ZZ, Xie YP, Yang L. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 6199–6208. [doi: 10.1109/CVPR.2018.00649]

[16] Zhang H, Xu T, Li HS, Zhang ST, Wang XG, Huang XL, Metaxas DN. StackGAN++: Realistic image synthesis with stacked generative adversarial networks. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2019, 41(8): 1947–1962. [doi: 10.1109/TPAMI.2018.2856256]

[17] Yu Z, Yu J, Xiang CC, Fan JP, Tao DC. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. IEEE Trans. on Neural Networks and Learning Systems, 2018, 29(12): 5947–5959. [doi: 10.1109/TNNLS.2018.2817340]

[18] Yu Z, Yu J, Fan JP, Tao DC. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 1839–1848. [doi: 10.1109/ICCV.2017.202]

[19] Kim JH, Jun J, Zhang BT. Bilinear attention networks. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montreal: Curran Associates Inc., 2018. 1571–1581.

[20] Fukui A, Park DH, Yang D, Rohrbach A, Darrell T, Rohrbach M. Multimodal compact bilinear pooling for visual question answering and visual grounding. In: Proc. of the 2016 Conf. on Empirical Methods in Natural Language Processing. Austin: ACL, 2016. 457–468. [doi: 10.18653/v1/D16-1044]

[21] Kim JH, On KW, Lim W, Kim J, Ha JW, Zhang BT. Hadamard product for low-rank bilinear pooling. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: OpenReview.net, 2017.

[22] Zhu MF, Pan PB, Chen W, Yang Y. DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 5795–5803. [doi: 10.1109/CVPR.2019.00595]

[23] Cheng J, Wu FX, Tian YL, Wang L, Tao DP. RiFeGAN: Rich feature generation for text-to-image synthesis from prior knowledge. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 10908–10917. [doi: 10.1109/CVPR42600.2020.01092]

[24] Qiao TT, Zhang J, Xu DQ, Tao DC. MirrorGAN: Learning text-to-image generation by redescription. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 1505–1514. [doi: 10.1109/CVPR.2019.00160]

[25] Collins E, Bala R, Price B, Süsstrunk S. Editing in style: Uncovering the local semantics of GANs. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 5770–5779. [doi: 10.1109/CVPR42600.2020.00581]

[26] Dorta G, Vicente S, Campbell NDF, Simpson IJA. The GAN that warped: Semantic attribute editing with unpaired data. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 5355–5364. [doi: 10.1109/CVPR42600.2020.00540]

[27] Lang YN, He Y, Dong JF, Yang F, Xue H. Design-gan: Cross-category fashion translation driven by landmark attention. In: Proc. of the ICASSP 2020–2020 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. Barcelona: IEEE, 2020. 1968–1972. [doi: 10.1109/ICASSP40776.2020.9053880]

[28] Li BW, Qi XJ, Torr PHS, Lukasiewicz T. Image-to-image translation with text guidance. arXiv: 2002.05235, 2020.

[29] Liang XD, Zhang H, Lin L, Xing E. Generative semantic manipulation with mask-contrasting GAN. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 574–590. [doi: 10.1007/978-3-030-01261-8_34]

[30] Chen YC, Shen XH, Lin Z, Lu X, Pao IM, Jia JY. Semantic component decomposition for face attribute manipulation. In: Proc. of 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 9851–9859. [doi: 10.1109/CVPR.2019.01009]

[31] Tang H, Liu H, Sebe N. Unified generative adversarial networks for controllable image-to-image translation. IEEE Trans. on Image Processing, 2020, 29: 8916–8929. [doi: 10.1109/TIP.2020.3021789]

[32] Chen JB, Shen YL, Gao JF, Liu JJ, Liu XD. Language-based image editing with recurrent attentive models. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 8721–8729. [doi: 10.1109/CVPR.2018.00909]

[33] Zhang LS, Chen QC, Hu BT, Jiang SR. Text-guided neural image inpainting. In: Proc. of the 28th ACM Int'l Conf. on Multimedia. Seattle: ACM, 2020. 1302–1310. [doi: 10.1145/3394171.3414017]

[34] Li BW, Qi XJ, Lukasiewicz T, Torr PHS. Controllable text-to-image generation. In: Proc. of the 33rd Int'l Conf. Neural Information Processing Systems. Vancouver, 2019. 2063–2073.

[35] Zhou XR, Huang SY, Li B, Li YM, Li JC, Zhang ZF. Text guided person image synthesis. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3658–3667. [doi: 10.1109/CVPR.2019.00378]

[36] Cheng Y, Gan Z, Li YT, Liu JJ, Gao JF. Sequential attention GAN for interactive image editing. In: Proc. of the 28th ACM Int'l Conf. on Multimedia. Seattle: ACM, 2020. 4383–4391. [doi: 10.1145/3394171.3413551]

[37] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 4396–4405. [doi: 10.1109/CVPR.2019.00453]

[38] Chen YP, Kalantidis Y, Li JS, Yan SC, Feng JS. Multi-fiber networks for video recognition. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 364–380. [doi: 10.1007/978-3-030-01246-5_22]

[39] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft COCO: Common objects in context. In: Proc. of the 13th European Conf. on Computer Vision. Zurich: Springer, 2014. 740–755. [doi: 10.1007/978-3-319-10602-1_48]

[40] Wah C, Branson S, Welinder P, Perona P, Belongie S. The caltech-UCSD birds-200–2011 dataset. Pasadena: California Institute of Technology, 2011.

[41] Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X. Improved techniques for training GANs. In: Proc. of the 30th Int'l Conf. on Neural Information Processing Systems. Barcelona: Curran Associates Inc., 2016. 2234–2242.

[42] Barratt S, Sharma R. A note on the inception score. arXiv: 1801.01973, 2018.

[43] Dong H, Yu SM, Wu C, Guo YK. Semantic image synthesis via adversarial learning. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 5707–5715. [doi: 10.1109/ICCV.2017.608]

[44] Zhu XJ, Li TF, de Melo G. Exploring semantic properties of sentence embeddings. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: ACL, 2018. 632–637. [doi: 10.18653/v1/P18-2100]

[45] Brock A, Donahue J, Simonyan K. Large scale GAN training for high fidelity natural image synthesis. In: Proc. of the 7th Int'l Conf. on

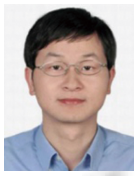
Learning Representations. New Orleans: OpenReview.net, 2019.

附中文参考文献:

[10] 陈佛计, 朱枫, 吴清潇, 郝颖明, 王恩德, 崔芸阁. 生成对抗网络及其在图像生成中的应用研究综述. 计算机学报, 2021, 44(2): 347–369. [doi: 10.11897/SP.J.1016.2021.00347]

[14] 杨婉香, 严严, 陈思, 张小康, 王菡子. 基于多尺度生成对抗网络的遮挡行人重识别方法. 软件学报, 2020, 31(7): 1943–1958. http://www.jos.org.cn/1000-9825/5932.htm [doi: 10.13328/j.cnki.jos.005932]

吴福祥(1984－), 男, 博士, 助理研究员, CCF 专业会员, 主要研究领域为多模态深度学习, 文本图像合成, 自然语言处理.



程俊(1977－), 男, 博士, 研究员, 博士生导师, 主要研究领域为机器视觉, 机器人, 机器智能和控制.