软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2024,35(7):3497−3530 [doi: 10.13328/j.cnki.jos.007053]

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

基于手绘草图的视觉内容生成深度学习方法综述*

左 然[1,2], 胡皓翔[1,2], 邓小明[1], 马翠霞[1,2], 王宏安[1,2]

[1](人机交互北京市重点实验室 (中国科学院 软件研究所), 北京 100190)

[2](中国科学院大学 计算机科学与技术学院, 北京 100049)

通信作者: 马翠霞, E-mail: cuixia@iscas.ac.cn

摘 要: 手绘草图通过绘制简单的线条直观呈现用户的创作意图, 支持用户采用手绘的方式快速表达思维过程及设计灵感, 创作目标图像或视频. 随着深度学习的发展, 基于草图的视觉内容生成通过学习草图和视觉对象 (即图像和视频) 的特征分布进行跨领域特征映射, 实现图像自动生成草图以及草图自动生成对应的图像或视频, 与传统的人工创作方式相比有效地提高了生成的效率和多样性, 成为计算机视觉、图形学领域的重要研究方向, 并且在设计、视觉创作等领域具有重要作用. 综述基于草图的视觉内容生成深度学习方法的研究现状和发展趋势, 按照视觉对象的不同将现有工作分为基于草图的图像生成和基于草图的视频生成方法, 并结合草图和视觉内容跨域生成、风格转化、视觉内容编辑等任务对生成模型进行详细分析, 然后比较和总结常用的数据集、针对草图数据不足提出的扩充方法以及生成模型的评估方法, 进一步通过草图在视觉内容生成应用中面临的挑战及生成模型未来发展方向对研究趋势进行展望.

关键词: 人机交互; 手绘草图; 视觉内容生成; 深度学习

中图法分类号: TP391

中文引用格式: 左然, 胡皓翔, 邓小明, 马翠霞, 王宏安. 基于手绘草图的视觉内容生成深度学习方法综述. 软件学报, 2024, 35(7): 3497–3530. http://www.jos.org.cn/1000-9825/7053.htm

英文引用格式: Zuo R, Hu HX, Deng XM, Ma CX, Wang HA. Survey on Deep Learning Methods for Freehand-sketch-based Visual Content Generation. Ruan Jian Xue Bao/Journal of Software, 2024, 35(7): 3497–3530 (in Chinese). http://www.jos.org.cn/1000-9825/7053.htm

Survey on Deep Learning Methods for Freehand-sketch-based Visual Content Generation

ZUO Ran[1,2], HU Hao-Xiang[1,2], DENG Xiao-Ming[1], MA Cui-Xia[1,2], WANG Hong-An[1,2]

[1](Beijing Key Laboratory of Human-computer Interaction (Institute of Software, Chinese Academy of Sciences), Beijing 100190, China)

[2](School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: Freehand sketches can intuitively present users' creative intention by drawing simple lines and enable users to express their thinking process and design inspiration or produce target images or videos. With the development of deep learning methods, sketch-based visual content generation performs cross-domain feature mapping by learning the feature distribution between sketches and visual objects (images and videos), enabling the automated generation of sketches from images and the automated generation of images or videos from sketches. Compared with traditional artificial creation, it effectively improves the efficiency and diversity of generation, which has become one of the most important research directions in computer vision and graphics and plays an important role in design, visual creation, etc. Therefore, this study presents an overview of the research progress and future development of deep learning methods for sketch-based visual content generation. The study classifies the existing work into sketch-based image generation and sketch-based video generation according to different visual objects and analyzes the generation models in detail with a combination of specific tasks including cross-domain generation between sketch and visual content, style transfer, and editing of visual content. Then, it summarizes and compares the

* 基金项目: 国家自然科学基金 (62272447); 北京市自然科学基金 (4212029); 2019 年牛顿奖中国奖 (NP2PB/100047)
收稿时间: 2023-03-02; 修改时间: 2023-06-05, 2023-07-03; 采用时间: 2023-09-18; jos 在线出版时间: 2024-01-31
CNKI 网络首发时间: 2024-02-02

commonly used datasets and points out sketch propagation methods to address in sufficient sketch data and evaluation methods of generated models. Furthermore, the study prospects the research trend based on the challenges faced by the sketch in the application of visual content generation and the future development direction of generated models.

Key words: human-computer interaction; freehand sketch; visual content generation; deep learning

手绘草图是一种直观传达用户意图的表达方式, 它通过简单的线条组合描述用户在持续时间内的思维变化, 在用户创作过程中及时捕捉创作灵感并且激发创新思维. 随着带有触屏功能的智能移动终端的普及, 手绘草图成为一种重要且高效的交互方式支持用户表达抽象的概念及传递信息, 并逐渐应用至识别、检索、生成等多种视觉相关的任务中. 其中基于手绘草图的视觉内容生成旨在通过学习草图和视觉对象的特征分布进行特征映射, 实现草图与视觉对象互相转化生成, 并在生成过程中保持语义内容和外观结构的一致性, 在实际生活中具有重要的应用场景. 该任务支持多样化的草图素材生成, 让用户通过手绘或者直接使用已有草图的方式将大脑构思的视觉效果或设计场景呈现出来, 并且可以按照创作灵感随时对草图进行编辑与修改, 解决传统人工生成方式流程复杂和需要大量重复劳动力的问题, 提高生成效率和生成结果的多样化, 实现智能的设计和视觉创作等. 图像和视频是两种常见的视觉对象, 手绘草图可以直观地描述图像中的物体外观形状、空间结构以及场景布局等并生成对应的图像结果, 图像则通过抽象为线条的形式生成对应的草图. 视频在静态的二维视觉内容基础上增加了时间维度, 草图在视频生成中除了描述视频帧的语义内容, 还可以通过衔接视频帧间的语义关联在一定程度上传递时序信息, 让用户经由草图描述控制视频中物体的运动轨迹并进行视频创作. 因此, 本文对基于草图的图像和视频生成这一计算机视觉、图形学等领域的热点研究问题进行综述.

早期的基于草图进行视觉内容生成的工作多集中在图像领域, 草图在生成图像时通过草图检索与图像合成两个步骤生成图像[1,2], 具体方法为: 先利用草图检索相关的图像元素, 然后将筛选出的图像元素进行合成, 以半自动的方式生成对应的图像. 它们在生成过程中仅利用草图检索现有的图像元素进行合成, 没有充分地考虑草图与生成结果之间的语义一致性, 很大程度限制了生成的真实性和多样性. 图像在生成草图时, 多采用边缘提取的方式获取图像轮廓[3–6], 与真实手绘草图的风格相差较大. 在视频领域, 早期的工作多集中在通过示例视频提供动作驱动生成二维草图动画[7], 或者通过人体骨架[8]提供运动轨迹生成视频, 缺乏草图直接生成或编辑视频的工作. 随着深度学习、人工智能等技术的蓬勃发展, 基于神经网络的生成模型如变分自动编码器 (variational auto-encoder, VAE)[9]和生成对抗网络 (generative adversarial network, GAN)[10]等被提出, 推动了根据输入目标自动生成对应视觉内容任务的发展. 其中 GAN 采取博弈的思想, 利用生成器生成样本以及鉴别器区分生成样本和真实样本, 并通过对抗损失函数让判别器输入真实样本时输出最小化、输入生成样本时输出最大化进行对抗训练提高生成的质量, 它通过训练生成模型实现生成数据的分布与真实数据的分布保持一致. 条件 GAN 网络 (conditional generative adversarial network, cGAN)[11]增加额外的条件输入至生成器和判别器中, 利用条件约束影响生成结果, 被广泛地应用至图像自动生成具有手绘风格的草图以及草图生成对应的图像或视频的工作中. 图像生成草图工作可以实现自动获取丰富的草图素材, 克服手动绘制收集草图的难度, 扩充了草图数据的规模, 并进一步增强草图的应用潜力. 草图生成图像或视频工作通过输入草图直接控制生成的内容, 极大地简化了生成流程, 扩大了基于草图进行视觉内容生成的实际应用范围. 例如在设计领域, 专业设计师和业余的设计人员都可以通过草图描述设计理念并对设计的产品进行智能生成和编辑, 实现基于草图的服装和发型设计、人脸照片美化、工业产品设计和室内设计等. 在视觉创作领域, 用户可以通过草图生成真实的物体实例以及故事场景来进行图像和视频素材的创作, 并且可以进一步应用至电影制作中, 如利用草图描述和修改视频关键帧解决传统计算机图形学使用复杂场景渲染技术时遇到的困难, 简化电影场景的制作流程, 还可以通过草图来智能更换视频中的人脸或者场景, 辅助电影完成后期制作. 在其他专业领域, 如在公安刑事调查时画像师可以通过草图对嫌疑人绘制模拟画像得到自动生成的嫌疑人真实人脸图像, 在目击者提供的细节较少时仍能得到高质量的嫌疑人画像, 提高破案的成功率; 在实际教学过程中, 教师通过草图生成真实的样例来模拟实验内容、演示实验结果等让学生们产生更加直观的教学体验, 利用草图辅助教学并提高教学的效率.

利用深度学习方法实现基于草图的视觉内容生成具有一定的挑战性, 由于草图通过线条组合描述物体的结构形状和纹理细节, 采用深度学习方法提取到的草图特征是稀疏且抽象的, 增大了草图特征与图像特征跨领域映射的难度, 如何通过深度学习模型有效地实现稀疏草图特征与稠密视觉对象特征的映射, 并实现图像生成抽象手绘

草图以及草图生成高质量的图像或视频具有较大挑战. 草图随着用户绘制习惯的变化而具有多样化的风格, 它对物体轮廓边缘和纹理细节的描述与真实的物体存在着一定的偏差, 为图像生成草图并保留草图的绘制风格增加了难度. 此外, 草图在生成图像或视频时如果生成的结果过度依赖输入草图, 输入形变程度较大的草图会严重降低生成结果的真实性, 在生成过程中如何保持与输入草图语义内容和结构形状一致的同时又保证生成结果的真实性同样是十分有挑战的. 基于草图的视觉内容生成采用的深度学习方法需要较大规模的草图-图像数据集和草图-视频数据集, 但草图绘制存在一定的难度, 特别是在绘制人脸草图、场景草图等复杂草图时, 构建足够规模的数据集比较困难. 为解决上述的挑战, 现有的工作提出草图数据扩充方法, 使用图像边缘提取方法得到边缘图或者生成模型生成草图代替手绘草图来扩大数据集规模; 利用生成模型增强草图特征表示, 并通过修改生成器、判别器等方式提高草图特征和视觉特征的映射效果; 利用生成模型纠正形变的草图以及补充草图缺失的细节信息, 提高生成模型输入手绘草图时的泛化性能, 得到较高质量的生成结果.

目前已有工作[12–16]对草图在视觉内容生成领域的应用进行了总结. Xu 等人[12]对草图相关的深度学习任务进行综述, 基于草图的视觉内容生成仅是其中的一个分支, 未被进行具体深入的分析. Elasri 等人[13]和 Zhan 等人[14]分别总结了图像生成相关工作以及基于多模态的图像生成和编辑工作, 它们侧重描述图像领域整体的生成工作研究进展, 仅对草图生成和编辑图像工作进行简单的概述, 没有开展算法层面的比较和分析, 并且该部分包含的相关工作不够完整. Chen 等人[15]面向图像和视频着色问题, 总结了草图在图像和视频着色中的应用, 但是仅包含了基于草图的视觉内容生成一部分内容. Wang 等人[16]总结 GAN 网络方法下草图应用至图像生成领域的相关工作, 他们描述了草图在图像生成中面临的挑战, 从有无精细控制上对任务进行分类, 但是缺乏对其他视觉内容生成任务的详细描述如图像生成草图和草图在视频生成、视频编辑中的应用等, 在相关工作分析时缺乏算法层面的比较, 并且在未来展望部分更侧重于草图生成图像在艺术层面的应用, 缺少分析生成方法的改进方向. 因此本文对草图在图像和视频生成领域的应用进行分析, 并对该领域使用的深度学习方法[17–29]进行详细分类和对比讨论. 如图 1 所示.

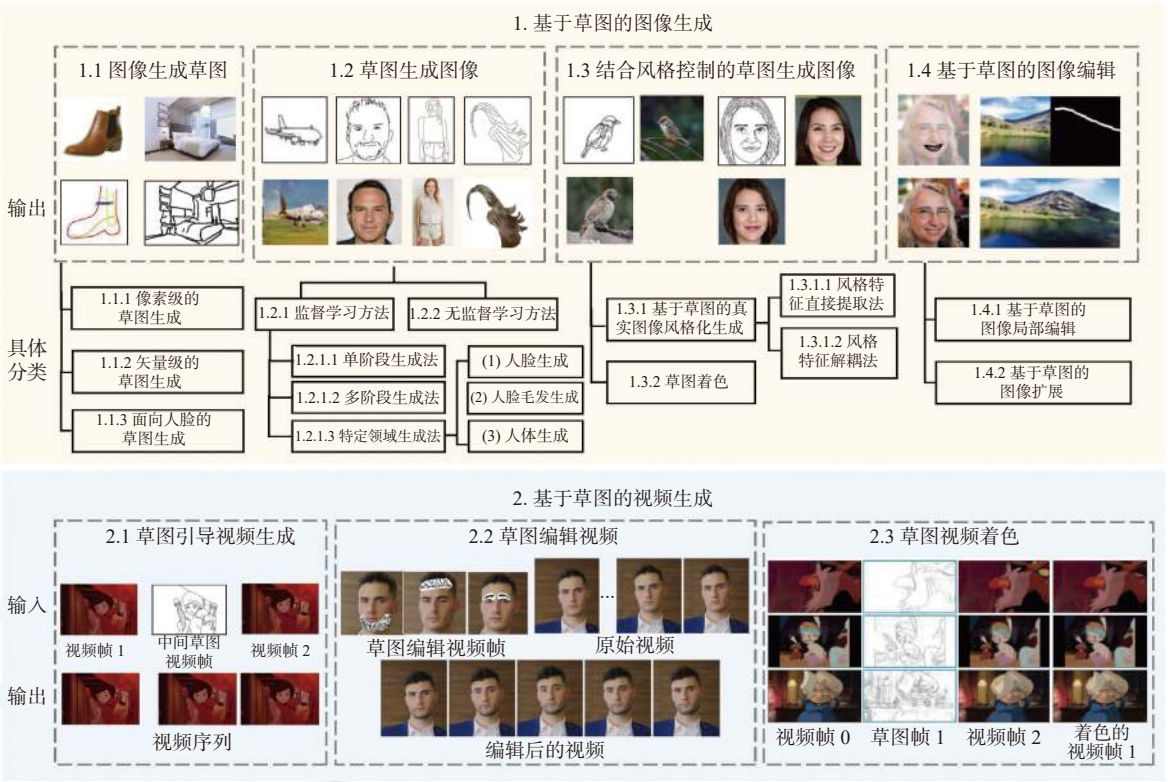


图 1 基于草图的视觉内容生成深度学习方法分类示意图

本文主要面向图像、视频两种视觉对象, 探讨与草图之间的跨域生成技术, 主要从基于草图的图像生成和基于草图的视频生成两类任务介绍草图在视觉内容生成领域的深度学习方法. 进一步地, 我们结合不同生成任务的输入内容和草图的形式对已有的研究工作进行分类, 当草图为完整物体草图时, 通过学习草图和图像的特征分布, 草图域和图像域可以互相转化生成, 实现草图生成图像和图像生成草图; 当输入为完整草图描述物体的结构以及额外的风格控制如彩色线条、纹理块、示例图像等描述外观风格时可以生成风格化图像, 实现风格控制的草图生成图像; 输入为局部草图线条以及待编辑的图像时, 通过补全草图修改的掩模区域并生成新的完整图像, 实现基于草图的图像编辑. 在视频领域, 输入草图描述完整视频帧的内容时, 可以通过草图传递视频在语义层面和时序层面的变化, 由草图引导视频生成; 输入局部草图修改视频帧时, 将修改的结果映射至整个视频序列, 实现草图编辑视频; 输入草图序列描述视频帧内容以及参考视频帧描述视频着色分布时, 通过视频帧和草图帧的色彩映射实现草图视频着色. 之后我们总结相关的数据集, 以及草图数据不足的情况下所提出的一系列草图数据扩充方法, 并列举常用的定量评估和定性评估方法, 最后结合手绘草图在视觉内容生成应用中的主要挑战及模型改进方向总结该任务的未来发展趋势.

1 基于草图的图像生成

1.1 图像生成草图

草图与跨域生成任务相结合时, 既可以通过学习草图的特征分布将其映射至图像域中生成对应的图像, 还可以逆向学习图像的特征分布将其映射至草图域中实现图像生成草图. 利用图像生成草图与传统的图像边缘检测任务不同, 手绘草图在描述物体时是高度抽象的, 它与图像的轮廓细节之间不是严格的对应关系, 并且在实际绘制过程中一张图像经常对应着多种风格的草图, 因此在生成时需要保留草图的绘制风格得到更接近真实手绘的草图. 我们按照草图生成的形式对目前的方法进行分类, 像素级的草图生成直接生成完整的草图图像, 矢量级的草图生成旨在生成由笔画构成的矢量草图, 面向人脸的草图生成根据输入的人脸图像以及人脸特有的属性生成对应的人脸草图.

图像生成草图模型框架如图 2 所示, 图 2(a) 为像素级的草图生成基于 GAN 网络, 通过改进损失函数或添加注意力机制提高生成草图图像的效果[17,30,31]; 图 2(b) 为矢量级的草图生成通过监督和无监督学习结合生成草图, 并改进损失函数提高生成效果, 或通过光栅器生成 Bezier 曲线表示矢量草图[18,32,33]; 图 2(c) 为面向人脸的草图生成基于 GAN 网络, 通过添加人脸语义分区、提取多尺度的图像特征或改进损失函数提高人脸草图生成效果[34–37].

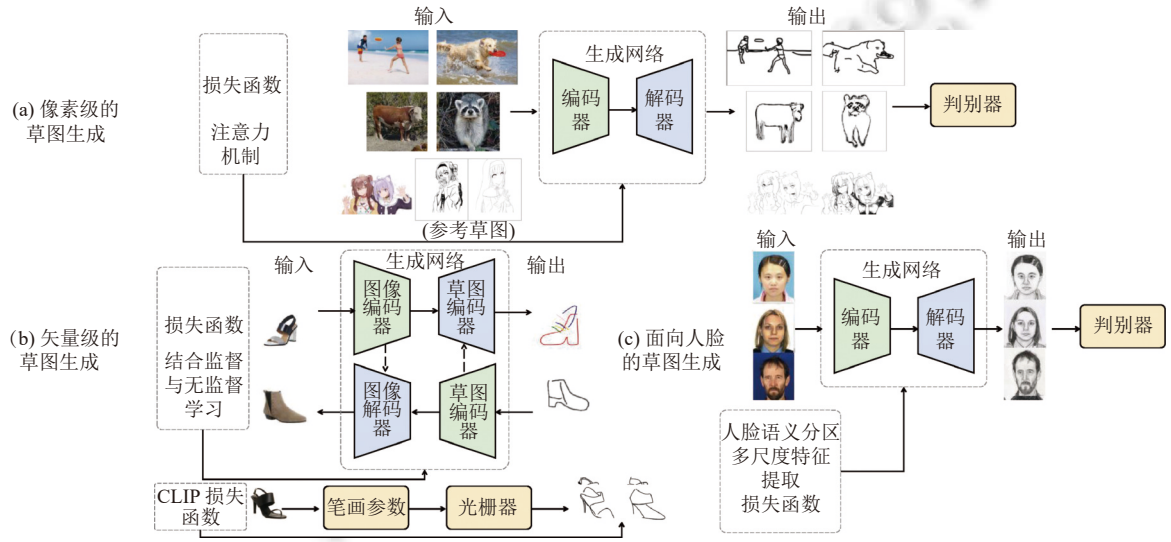


图 2 图像生成草图模型框架结构示意图

1.1.1 像素级的草图生成

像素级的草图生成主要面向包含多个类别的物体实例或者场景, 将图像和草图映射至像素空间并生成对应的草图图像, 在生成过程中需要保持或者增加草图的绘制风格控制 (如图 2(a) 所示). Photo-sketching[17]为了生成更接近真实手绘的轮廓草图图像, 以 cGAN 网络为框架, cGAN 的对抗损失函数为:

$$\mathcal{L}_{\text{cGAN}}(G,D)=\min_{D}\max_{G}\mathbb{E}_{x,y}\left[\log D(x,y)\right]+\mathbb{E}_{x,z}\left[1-\log D(x,G(x,z))\right] \tag{1}$$

其中, G 表示生成器, D 表示判别器, x,y 分别代表输入的条件控制和真实图像, z 表示随机噪声. Photo-sketching 利用图像和草图之间一对多的关系提出 MM (min-mean) 损失, 将图像与对应的多个草图的对抗损失取平均以及 L_1 损失取最小值, 实现轮廓草图的重建, L_1 损失函数为:

$$\mathcal{L}_{L_1}(G)=\mathbb{E}_{x,y,z}\left[\|y-G(x,z)\|_1\right] \tag{2}$$

MM 损失函数为:

$$\mathcal{L}(\text{min-mean})=\frac{\lambda}{N^i}\sum_{k=1}^{N^i}\mathcal{L}_{\text{cGAN}}(x_i,y_i^k)+\min_{k\in\{1,\ldots,N^i\}}\mathcal{L}_{L_1}(x_i,y_i^k) \tag{3}$$

其中, x_i 和 $y_1^1,\ldots,y_i^k,\ldots,y_i^{N^i}$ 代表输入的图像和它对应的多张草图, λ 是超参数控制每项损失函数的重要性.

Kampelmuhler 等人[30]采用全卷积的编码器-解码器结构实现图像生成草图, 并添加 AdaIN 将类别特征逐层插入解码器中. 他们提出了感知相似度量损失函数, 先利用 Sketchy 数据集预训练 CNN 分类模型去隐式地学习草图的形状表示, 再根据模型中间层的权重将通过生成模型得到的生成草图和真实草图之间特征距离最小化, 生成与手绘草图相似的草图图像. Ashtari 等人[31]通过输入图像和参考草图, 生成与图像视觉内容一致和与参考草图风格一致的草图. 他们采用自参考的方式对草图进行空间旋转和变化作为参考草图进行训练, 并添加 CBAM 注意力机制至图像和参考草图特征编码器中, 之后将参考草图和图像进行特征拼接, 并利用 CBAM 注意力机制学习图像和参考草图的对应关系, 经过解码器得到内容和风格与输入图像和参考草图一致的草图.

1.1.2 矢量级的草图生成

为了实现图像生成笔画层面的手绘草图, 现有的方法采用基于 RNN 的解码器或利用 Bezier 曲线生成矢量化的草图, 并改进生成网络结构和损失函数提高笔画级草图的生成效果 (如图 2(b) 所示). Song 等人[18]将 CNN 与 RNN 相结合, 分别构建图像与草图的编码器和解码器, 并且提出一种结合监督学习和无监督学习的混合学习框架, 监督训练图像生成草图和草图生成图像, 无监督训练图像生成图像和草图生成草图, 采用快捷循环一致性损失函数在单领域内进行重建. 该方法有效地利用了图像和多风格草图的弱监督关系, 按照笔画逐步生成手绘草图. Zhang 等人[32]将基于 RNN 的 VAE 模型和 GAN 网络相结合, 提出无监督的矢量化草图生成方式, 他们在已有的循环一致性损失函数中加入边缘损失函数, 通过最小化图像对应的边缘图和生成草图之间的特征距离实现草图和图像在隐空间的正确匹配, 确保生成的草图与输入图像的一致性. 上述方法模拟人类绘制草图的过程得到逐笔生成的草图, 但生成的类别有限. CLIPasso[33]采用一系列 Bezier 曲线表示草图并利用图像直接生成矢量化的草图, 它通过提取输入图像的显著性区域初始化笔画位置, 并通过最小化 CLIP 提取的草图和图像特征距离优化笔画属性和笔画位置. CLIPasso 实现任意类别物体生成对应的草图, 并通过笔画数量有效地控制草图的抽象程度.

1.1.3 面向人脸的草图生成

如图 2(c) 所示, 为了提高生成人脸草图的真实性, Sketch-Transformer[34]使用多尺度的图像特征编码器获取不同维度的图像特征和位置编码, 添加自注意力模块建立不同位置特征编码的依赖关系并进行特征更新, 将多尺度特征输入基于 SPADE[38]的解码器中重建对应的人脸草图. SCA-GAN[35]、SDGAN[36]和 EADT[37]利用人脸的语义分区指导人脸图像生成对应的草图, 并将语义分区插入到生成网络的解码器中, 提高生成的人脸草图效果. 为进一步增强对生成草图的语义控制, SCA-GAN 通过两阶段 GAN 网络细化生成结果, SDGAN 基于语义分区各类别的像素平均值和方差构建自适应特征加权图, 将生成草图与真实草图距离最小化, EADT 对语义分区进行边缘一致性过滤, 实现语义分区边缘的正确分类.

图像生成像素级的草图将草图按照图像的方式进行处理, 丢失了草图的笔画属性并且生成的轮廓比较模糊.

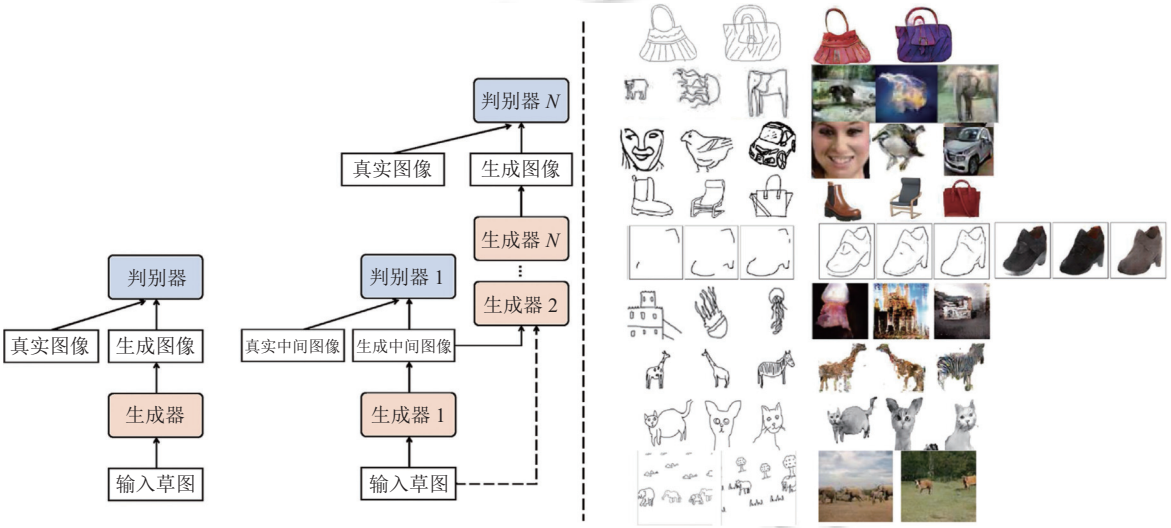
图像生成矢量级的草图可以得到清晰的草图轮廓, 符合实际手绘过程中使用的草图形式, 但是在生成草图的笔画时容易产生误差, 例如生成与实际不符的形状和线条或者错误的线条细节等, 并且当笔画数量过多时会给生成模型带来巨大的负担, 特别是逐笔生成将变得十分困难. 图像生成人脸草图主要生成的是素描人脸, 与真实的手绘人脸相差较大. 总之, 图像生成草图的方法可以应用至草图数据扩充, 扩大目前草图相关的数据集规模, 提高草图生成图像模型的性能. 但是当物体结构复杂、姿态形变度较高或者背景复杂时, 现有的模型较难学习到草图的抽象语义表示, 生成效果将会显著下降. 如何实现复杂图像生成高质量的草图并且获得草图的笔画属性以及保持特有的手绘风格是之后的研究重点.

1.2 草图生成图像

根据使用的数据集是否提供成对的草图-图像示例对, 草图生成图像方法分为监督学习和无监督学习方法. 监督学习方法以真实的草图-图像对约束生成模型, 保持草图与生成图像语义内容的一致性. 由于草图需要大量的人工绘制导致收集成对的草图-图像数据难度较大, 因此无监督的生成方法被提出, 该方法输入一系列未建立匹配关系的源域草图和目标域图像进行训练, 学习草图域和图像域的特征分布并将草图转化生成目标域图像, 实现无监督的草图生成对应的图像.

1.2.1 基于监督学习的草图生成图像方法

基于监督学习的草图生成图像方法示意图及生成结果示意图如图 3 所示[19,39–46].



(a) 单阶段生成法 (b) 多阶段生成法基本框架示意图 (c) 相关工作生成结果示意图

图 3 基于监督学习的草图生成图像

1.2.1.1 单阶段生成法

单阶段生成法面向常见类别的单物体实例草图, 如图 3(a) 所示, 模型中只有单个生成器和鉴别器, 直接通过生成器学习与草图相近的分布并从分布中采样生成图像, 再由判别器判别生成图像和真实图像. Pix2Pix[39]是监督学习下跨域图像生成的典型网络, 它基于 cGAN, 支持输入边缘图并生成对应的图像. 生成器采用基于 U-Net[47]的编码器-解码器结构, 并加入跳跃连接实现底层信息的共享, 然后将草图-真实图像对和草图-生成图像对输入至判别器中进一步增强生成结果. 判别器 PatchGAN 将图像分割为多个局部图像块并判断真假, 捕捉图像的局部细节或高频信息从而生成高质量的图像. 虽然 Pix2Pix 在输入实际手绘草图时生成质量下降, 它仍是草图生成图像模型中的重要基线网络.

SketchyGAN[19]提出输入包含多种物体类别的手绘草图训练 GAN 网络得到草图生成图像模型, 并利用该模型生成对应的图像. 该工作在生成器和判别器中引入掩模残差单元 (MRU), 并将不同尺寸的草图输入至生成器的多

个 MRU 模块中, 通过学习内部掩模动态地从输入草图中提取新特征并与特征图相结合, 有效地传递草图特征信息, 判别器与生成器操作相同. SketchyGAN 除了采用 GAN 网络的对抗损失函数外, 还引入分类损失函数判定生成图像的类别确保生成图像具有正确的语义标签, 并采用 L_1 损失函数与感知损失函数分别最小化生成图像和真实图像像素层面与 Inception-v4 网络[48]提取的特征层面的距离, 感知损失的公式为:

$$\mathcal{L}_{\text{perceptual}}(G)=\sum_i \lambda_p \|\varphi_i(G(x,z))-\varphi_i(y)\|_1 \tag{4}$$

其中, φ_i 代表的是 Inception 模型第 i 层的特征输出, λ_p 是超参数决定每一层计算的损失的重要性. 此外, SketchyGAN 添加多样性损失函数将不同噪声引入同一输入来最大化不同噪声生成图像的距离, 提高图像的生成质量和多样性.

由于 Pix2Pix[39]和 SketchyGAN[19]的输入草图与生成图像之间存在着较强的约束关系, 生成结果过度依赖于输入草图的好坏, 当输入绘制较差的手绘草图时会丢失图像的真实性. 因此 Contextual GAN[40]提出采用图像补全的方法, 由草图提供弱监督的上下文约束补全生成对应的图像. 在训练阶段将草图和图像作为联合图像输入至 GAN 网络中, 生成器将联合图像映射到联合空间中, 通过学习级联分布获取草图和图像的对应关系. 在补全阶段, 利用草图部分作为上下文信息预测联合图像被遮挡的部分, 通过反向迭代传播将遮挡的联合图像映射到生成器的特征空间中, 生成对应的联合图像. 该方法不需要草图和图像进行跨域的特征匹配, 削弱草图和生成图像之间的严格对应关系, 生成更加真实且多样化的结果.

Koley 等人[41]采用解耦编码器和解码器的训练策略实现输入抽象程度较高的草图时生成高质量的真实图像. 他们将 StyleGAN[49]作为解码器并提前在图像上进行预训练, 之后通过训练自回归的草图匹配网络将草图映射至 StyleGAN 的隐空间 W^+, 并在映射过程中依据隐向量的时序关系进行特征映射, 建立草图抽象程度与 W^+ 空间隐向量的关联关系, 利用抽象草图控制高层隐向量特征更新从而影响图像的主要语义结构, 随机生成低层隐向量反映抽象草图的不确定性, 生成与用户意图一致的多样化真实图像.

单阶段生成法通过简单的网络结构与训练过程学习单物体实例草图生成图像, 但是由于草图绘制风格多变、语义信息表述抽象, 草图和图像在结构上不是完全对齐的关系, 存在较大的模态差异, 因此直接生成的图像分辨率较低, 物体边缘生成模糊, 与草图描述的语义内容在精细程度和准确性上有较大的差距. Koley 等人[41]极大地提高了生成图像的质量, 但他们仅在简单的类别上进行验证, 无法体现模型输入抽象草图时生成多种复杂类别图像的性能.

1.2.1.2 多阶段生成法

如图 3(b) 所示, 为了提高草图生成图像的质量及多样性, 让生成模型学习到更多的草图信息, 多阶段生成法将生成任务进行分解, 用多个 GAN 网络去分别学习草图包含的语义内容和类别属性, 或者通过叠加多个 GAN 网络提高生成图像的粒度, 以及使用多个 GAN 网络建立草图和图像之间前景物体和背景信息的对应关系.

对于单物体实例草图, iSketchNFill[42]为了减轻用户输入草图的负担提出交互式的草图生成图像系统并构建基于 GAN 的两阶段生成模型, 它先对输入的稀疏局部草图进行推荐补全, 快速形成符合用户需求的完整草图, 再通过草图生成完整的图像. 草图生成图像阶段的生成器采用 MUNIT[50]的编码器-解码器结构, 并且加入物体类别为条件向量的门控制机制, 利用输出的参数去调整生成网络的特征让模型适应于多类别的物体实例生成. Li 等人[43]基于 cGAN, 先输入类别标签生成图像以学习各类别的公共表示, 再将学习到的类别先验与草图结合输入至第 2 阶段的草图生成图像网络中生成更高质量的图像. 文献 [44] 基于 Pix2Pix 逐层学习草图和图像在结构、纹理上的映射关系[44], 模型先输入草图和类别信息生成边缘图, 进一步拼接草图、边缘图以及类别信息生成对应的图像, 并在生成过程中加入互信息指导模型学习特征之间的相关性, 实现没有类别标签指导的图像生成. 文献 [45] 构建双层级联 GAN 网络[45], 第 1 层网络学习手绘草图简单的纹理和结构生成粗粒度图像, 第 2 层补充细节特征将粗粒度图像生成高清图像.

SketchyCOCO[46]提出由多个物体实例组成场景草图生成对应的图像. 考虑到直接将场景草图与图像进行特征映射有较大难度, 该方法将前景物体和背景分别进行生成, 即先采用草图分割的方法得到场景草图的物体实例, 并通过前景生成模块生成对应的物体, 再输入背景草图和前景物体生成最后的场景图像. 在前景生成阶段提出 EdgeGAN

网络 (如图 4 所示), 该网络在训练时通过输入随机向量分别生成边缘图及对应的图像, 学习两个模态共有的属性向量和类别向量, 并将属性向量与类别相结合构成完整的特征向量表示, 在测试时将手绘草图生成图像转化为手绘草图编码为属性向量并进一步生成对应的图像, 生成的图像与输入草图弱相关并且结果更加真实. 但是生成效果受实例分割影响, 当输入过于抽象的草图实例时分割效果和生成质量会显著下降.

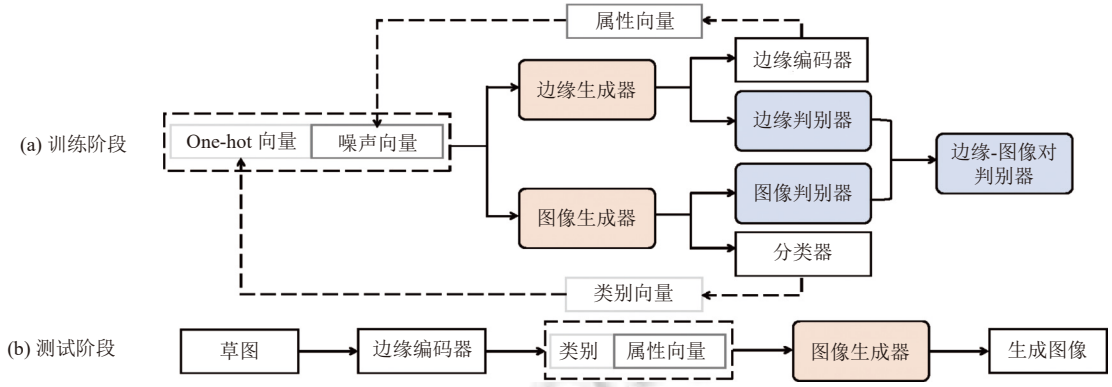


图 4 EdgeGAN 模型示意图

与单阶段生成法相比, 多阶段生成法在生成过程中可以从稀疏的草图输入中捕捉更多的细节特征, 明显提高了生成图像的质量和草图与图像之间的语义一致性. 但是多阶段生成法的生成网络更加复杂, 大多数现有的方法在第 2 阶段生成模型中会利用第 1 阶段的生成结果, 导致生成模型的训练难度增大以及生成偏差累积, 造成最终图像生成质量的下降.

1.2.1.3 特定领域生成法

不同于常见的多种物体类别生成, 特定领域的草图如人脸草图、人脸毛发区域草图 (包括发型或胡须) 和人体草图等通常具有明确的语义结构, 特定领域生成法除了基于通用的草图特性改进生成网络外, 还会结合类别的特有属性以及实际应用提出对应的方法来提高生成图像的质量.

(1) 人脸生成

人脸生成模型框架结构示意图如图 5 所示. 图 5(a) 为在生成网络中添加边缘图生成图像, 实现对手绘草图的形变纠正提高其生成人脸图像的效果[51]; 图 5(b) 为利用两阶段生成网络细化人脸图像生成[52]; 图 5(c) 为利用人脸语义分区编码人脸草图特征生成人脸图像[20].

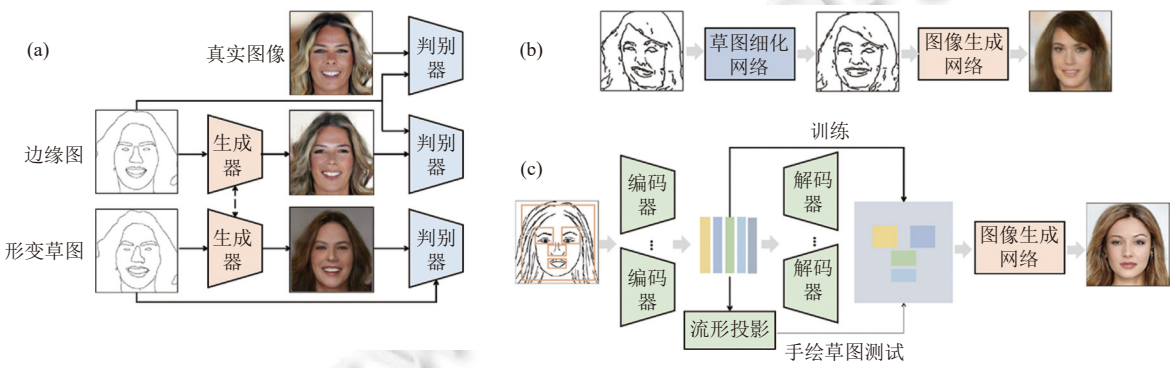


图 5 人脸生成模型框架结构示意图

与常见的草图类别如动物、椅子、鞋等相比, 人脸草图的描述更加复杂, 构建对应的手绘草图-图像数据集成本较高, 所以目前多采用边缘图-图像对的方式进行学习, 但是边缘图与原始图像的结构基本对齐, 而实际的手绘

草图是稀疏、形变的, 缺乏对图像的细节描述甚至是轮廓相关的必要线条, 因此现有的方法建立草图与边缘图之间的映射关系, 提高边缘图训练的生成模型在输入手绘草图时的生成效果, 生成既与手绘草图的形状具有一定相似度又能保持真实性的人脸图像. 如图 5(a) 所示, DeepFacePencil[51]提出空间注意力池化层模块, 通过笔画的宽度反映草图与边缘图的限制关系, 自适应地采用不同内核进行池化操作以设置不同的笔画宽度, 并将笔画松弛至公差区域内, 然后构建空间注意力图根据草图的形变程度对草图特征图进行松弛度分配, 越多失真的笔画松弛度越大, 让图像在生成时与绘制较好的草图区域保持语义一致而在形变度高的草图区域则降低关联来保持生成的真实度. 该工作还提出构建双生成器, 将边缘对齐的草图和形变草图分别输入至生成器中, 并采用生成器特征匹配损失函数将双生成器中形变草图和边缘图的特征距离最小化, 让生成器感知草图线条形变并在特征空间进行校正. 如图 5(b) 所示, Cali-sketch[52]提出两阶段草图生成人脸图像模型, 首先对输入的草图进行笔画校准并补全生成细粒度的人脸草图, 再进一步生成逼真的人脸图像. 但是两种方法的草图形变程度都较低, 与实际的手绘草图差距较大. 所以 Yang 等人[53]提出可控的人脸草图生成图像任务, 允许用户通过调整草图和边缘图的匹配程度决定输入草图和输出图像之间的相似程度. 他们首先采用草图细化网络, 基于实际绘制中由粗到细的过程实现手绘草图生成边缘图, 通过扩张的方法将粗糙草图转化为覆盖边缘图的绘制区域, 训练 Pix2Pix 网络匹配粗糙草图与边缘图, 并将控制参数编码为风格向量, 风格化生成器中的草图特征, 实现可控的草图细化, 然后将细化过的草图输入至预训练的边缘图生成图像网络中去生成真实图像. 但是当草图的细化程度降低到一定程度后无法保证得到较高的生成图像质量.

除此之外, 人脸草图还具有特定的语义结构, 各个语义分区的特征存在一定的关联性, 为了捕捉人脸的长距离依赖并建模整体的结构, LinesToFacePhoto[54]提出添加 MRU 模块的 CSAGAN 模型, 生成网络采用编码器-解码器结构, 并在解码器最后一层添加自监督模块学习特征图之间的关系, 采用多尺度的判别器提高生成图像结构和纹理的质量. DeepFaceDrawing[20]则利用人脸特定的语义结构对人脸进行固定分区, 并采用局部到整体的方法与真实图像建立局部到全局的语义关联, 生成高质量人脸图像. 模型构架图如图 5(c) 所示, 将人脸草图分为左眼、右眼、鼻子、嘴巴和其余部分, 通过自动编码器学习人脸草图各区域的特征向量, 再将各部分特征向量通过多通道特征映射并按照空间位置组合为特征图, 最后采用 cGAN 网络生成高质量的真实人脸图像. 当输入手绘草图时, 流形投影将局部区域样本的特征向量作为当前区域流形的点样本, 通过局部线性嵌入 (locally linear embedding, LLE) 算法[55]对人脸草图局部特征进行线性重构, 帮助有效地学习人脸草图特征分区, 提高真实手绘的人脸草图生成的图像质量.

为生成多样化的图像结果, S2FGAN[56]在草图生成人脸图像中加入人脸属性编辑任务, 控制特定人脸属性的生成强度. S2FGAN 基于所有属性正交、当前属性编辑时其他属性保持不变的定理构建两种属性映射网络, 修改潜在空间中的语义, 并通过控制特定属性的强度生成对应的结果. pSp 模型[57]支持草图作为一种输入对图像进行风格化生成, 它提取输入模态的多层特征并编码为对应的风格向量, 然后采用预训练的 StyleGAN 网络实现图像生成. pSp 的生成器编码的是图像风格特征而不是空间输入, 它支持多样化的生成, 但是丢失掉了草图本身的结构特征导致生成的结果与草图不够相似.

目前的草图生成人脸图像虽然通过形变边缘图的方式实现手绘草图生成高质量图像, 但质量较高的生成结果仍是草图与边缘图比较接近的情况, 当草图绘制抽象或者绘制面部特征不完整时, 生成结果将明显下降. 而且数据集中的图像多为没有旋转和平移的正面图像, 缺乏不同角度和表情的图像生成或者带有配饰产生局部遮挡的人脸生成. 如何进一步提高边缘图生成图像在输入真实手绘草图时的泛化能力并实现复杂人脸图像的多样化生成是后续的研究方向.

(2) 人脸毛发区域生成

草图生成人脸毛发区域和人体的模型框架结构如图 6 所示. 图 6(a) 草图生成人脸毛发区域在毛发区域生成网络的基础上, 相关工作采用多个 GAN 网络补充毛发生成粒度[21,58]; 直接输入草图, 采用多个 GAN 网络先生成掩模再实现毛发区域更换[59]. 图 6(b) 草图生成人体在人体生成网络的基础上, 通过叠加多个 GAN 网络提高生成粒度[22,60].

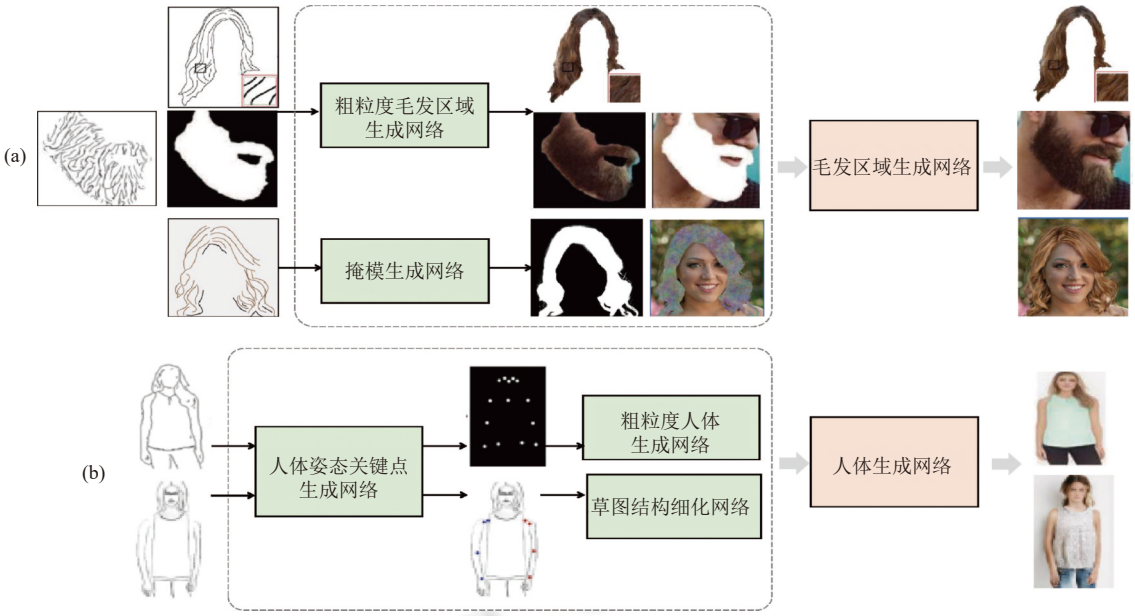


图 6　草图生成人脸毛发区域和人体的模型框架结构示意图

发型和胡须是人脸的重要组成部分, 它们影响着人脸的整体风格并且在实际的造型设计、人物识别任务中具有重要应用场景. 发型图像可以被当作二维分段平滑矢量场[61]并以方向图描述发型走向, 而草图通过笔画线条组合描述发型的结构和风格, 可以形成发型的方向图, 并与掩模组合输入生成对应的发型. 胡须与发型具有相同的属性, 利用草图可以有效地实现人脸局部发型、胡须的更换与设计. 现有的工作通过多阶段生成方法实现细粒度的人脸毛发区域生成, 支持在生成过程中将草图与掩模相结合进行毛发区域更换 (如图 6(a) 所示).

Qiu 等人[21]输入由掩模和描述头发方向的笔画构成的发型草图, 先生成结构与目标图像一致但缺乏纹理信息并且模糊的粗粒度发型图像, 再通过 Gabor 滤波器从生成的图像中提取方向图和纹理图输入到生成网络中生成高分辨率的发型图像. Olszewski 等人[58]为实现直接在人脸图像上进行胡须或发型的更换, 将掩模和一系列引导草图输入 Pix2Pix 为框架的生成网络, 生成胡须或发型图像, 进一步对图像进行细化和背景合并生成胡须或发型更换后的完整人脸图像, 该方法还支持修改草图笔画的结构及颜色对胡须或发型进行编辑. 上述方式利用输入草图构建发型的方向图, 过多地关注局部纹理的生成效果而忽略了发束整体之间的连接关系以及发束之间的遮挡问题, 当处理复杂发型时生成结果较差. 它们还需要额外的掩模输入描述发型的形状, 掩模影响着发型边缘的生成效果并且增大了用户输入负担.

因此, SketchHairSalon[59]仅输入带有颜色的草图描述复杂发型的结构、外观和形状, 生成逼真的发型图像. SketchHairSalon 采用两阶段生成模型, 先根据掩模生成网络由草图生成发型掩模, 将少量的非发型区域的笔画与发型草图组合输入至生成器中, 添加自注意力模块学习长距离草图包含的全局结构特征, 再将彩色草图、发型掩模和背景图像输入至发型生成网络得到更换发型的完整图像. 该方法分别训练生成编织发型和非编织发型, 除了对抗损失函数和感知损失函数外, 非编织发型由 L1 损失确保生成发型的像素质量, 编织发型通过形状损失函数即先通过高斯滤波器进行平滑再计算 L1 损失, 保证编织形状的一致性. SketchHairSalon 简化了生成模型的输入并生成高质量的发型结果. 但是当发型的边缘存在较多细碎毛发时, 细碎毛发无法正常生成并且与背景图像结合处存在失真感. 目前的二维草图无法描述深度信息, 生成具有层次的卷发时效果较差. 后续的研究将侧重细粒度的发型生成, 处理细碎毛发在生成时的边缘过渡问题, 以及定义草图深度或者输入额外的深度信息生成复杂的卷发发型, 增强草图生成发型的通用性.

(3) 人体生成

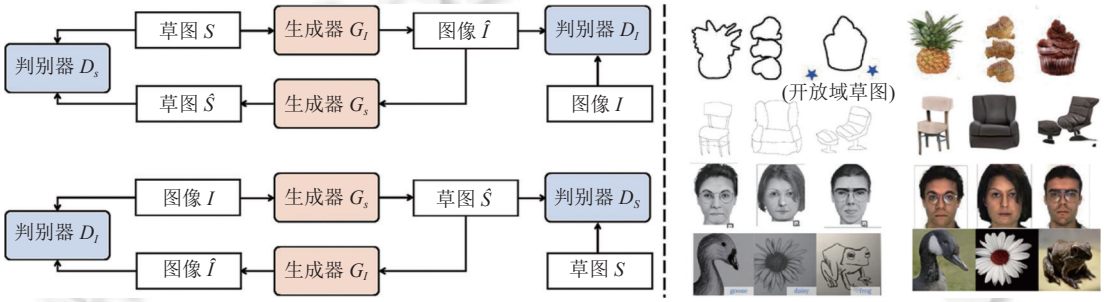
和人脸图像相比人体图像包含更加复杂的结构, 增加了人体草图的绘制难度, 并且在生成中需要考虑人体的

姿态、形状和衣着等, 给草图生成人体图像带来极大的挑战. 现有的工作利用人体具有固定姿态关键点的特征与草图结合生成人体图像, 并叠加多个 GAN 网络提高草图生成人体图像的效果, 如图 6(b) 所示. Ho 等人[60]首次在生成过程中加入语义关键点, 实现与草图姿态一致的人体图像生成. 它先输入人体草图, 利用 GAN 网络生成初始的图像并从中提取对应的姿态关键点, 然后将人体的姿态关键点和草图相结合通过多层 GAN 网络生成对应的人体图像. DeepPortraitDrawing[22]输入语义分割的草图并分别对形状和姿态进行优化, 它基于 LLE 算法[55]建立每个部位的流形并生成形状接近真实图像的人体草图与掩模, 进一步提取姿态关键点并通过级联优化的策略对人体草图和掩模的形状进行迭代优化, 减少草图和掩模与目标图像在形状和姿态上的误差, 最后将人体草图和掩模输入 GauGAN[38]为框架的生成网络中生成对应的人体图像. 但是人体草图本身具有一定的绘制难度, 当输入的草图与边缘图差距较大时, 调整结构和形状后生成的结果仍远偏离于真实图像.

1.2.2 基于无监督学习的草图生成图像方法

基于无监督学习的草图生成图像如图 7 所示[62–66]. CycleGAN[62]是无监督生成方法的基线网络, 目前的草图无监督跨域生成图像的工作多数都以 CycleGAN 为框架, 并结合草图的特性对网络进行改进. 应用在无监督学习的草图生成图像中的 CycleGAN 通过构建两个生成器和两个判别器, 实现草图领域和图像领域的循环生成, 基本的模型框架如图 7(a) 所示. CycleGAN 在对抗损失函数的基础上加入循环一致性损失函数, 具体指草图域内的草图 S 生成图像 $\hat{I}=G_I(S)$ 再反向生成草图 $\hat{S}=G_S(G_I(S))$, 图像域内的图像 I 生成草图 $\hat{S}=G_S(I)$ 再反向生成图像 $\hat{I}=G_I(G_S(S))$, 让生成草图与真实草图以及生成图像与真实图像差异最小, 损失函数的公式为:

$$\mathcal{L}_{\text{cyc}}(G_S,G_I)=\mathbb{E}_{S\sim p_{\text{data}}(S)}[\|G_S(G_I(S))-S\|]+\mathbb{E}_{I\sim p_{\text{data}}(I)}[\|G_I(G_S(I))-I\|] \tag{5}$$



(a) 使用的 CycleGAN 框架示意图 (b) 相关工作结果示意图

图 7 基于无监督学习的草图生成图像

AODA[63]为实现类别没有出现在训练数据中的开放域草图生成对应的图像, 基于 CycleGAN 构建级联框架, 先利用草图生成器实现图像生成草图, 进一步将生成草图和真实草图通过图像生成器生成图像. 在图像生成草图网络中, 采取随机混合取样的策略, 每个类别的真实草图中加入一定比例的生成草图, 根据域内的草图-图像对应关系将开放域的图像映射为草图, 减少图像生成器在输入生成草图时的生成误差, 提高开放域图像的生成质量.

Liu 等人[64]也利用 CycleGAN 网络先将草图生成灰度图, 为了从风格多样的抽象草图中提取样式不变的信息, 他们在手绘草图的基础上合成噪声草图并通过自监督的方式利用噪声草图重建原始草图, 让模型忽略与风格相关的无关笔画, 添加注意力机制降低笔画密集区域的权重, 有效地去除笔画干扰区域. 之后模型由灰度图生成真实图像, 通过自监督的强度损失函数让生成图像与灰度图像在 Lab 彩色空间距离最小化训练生成网络, 该阶段支持加入参考图像并通过 AdaIN 将参考图像的风格信息与灰度图相结合以生成多样化的结果.

Kazemi 等人[65]提出改进 CycleGAN 中的损失函数与鉴别器实现无监督的草图生成人脸图像. 他们将多层感知损失的均值定义为新的循环一致性损失函数, 在重建图像中保留高层次的面部特征, 并且提出了几何形状判别器, 将代表人脸形状的高层语义特征即感知损失层特征输入至判别器中判别生成图像和真实图像, 提高生成图像的纹理及形状对应.

不同于上述方法基于 CycleGAN 实现无监督的草图生成图像, MaskSketch[66]输入类别标签和草图, 采用预训

练的图像生成模型 MaskGIT[67]生成对应的真实图像. 该模型利用 VQ-GAN[68]编码图像为掩码序列, 在采样过程中并行预测掩码, 按照模型置信度对掩码进行过滤以及迭代预测. MaskSketch 在此基础上利用自注意力图指导模型采样过程, 通过计算草图和图像自注意力图之间的距离指导掩码特征更新, 让模型按照草图结构控制生成对应的图像. 该方法不需要对模型微调以及使用草图-图像数据集进行模型训练, 可以实现不同抽象程度的草图无监督生成对应的高质量图像.

1.2.3 讨 论

表 1 整理了目前草图生成图像的监督学习方法和无监督学习方法的对比分析, 对现有的研究方法进行细分类并总结各自的优缺点.

表 1 草图生成图像方法对比分析

方法分类		草图类型	方法描述	优点	缺点	相关工作
监督学习	单阶段生成法	多类别单物体实例	通过修改生成器和判别器优化模型生成效果	生成结果与草图保持一致性	生成结果对输入草图依赖性过强, 真实性降低	[19,39]
			将草图和图像作为联合图像, 草图提供弱监督信息用于联合图像补全生成对应的图像	减少草图对生成图像的约束, 提高生成图像的真实性	图像生成质量整体较低	[40]
			解耦编码器和解码器, 单独训练编码器将草图映射至StyleGAN隐空间	支持输入抽象草图并生成高质量的真实图像	缺少在复杂多类别草图生成图像的验证	[41]
	多阶段生成法	多类别单物体实例	基于多个GAN网络实现草图补全和草图生成图像相结合	通过草图推荐生成符合用户需求的多类别图像	生成图像的质量一般, 并且缺乏在结构复杂的类别上的训练	[42]
			基于多个GAN网络将类别先验叠加至草图生成图像	利用类别标签增强草图生成图像的语义控制		[43]
			基于多个GAN网络从粗粒度到细粒度生成图像	叠加GAN网络补充草图特征, 提高生成图像的粒度		[44,45]
		多类别多物体实例	基于多个GAN网络分别学习前景物体和背景信息	分阶段生成前景和背景物体, 提高生成的效率和图像的真实性	生成结果受实例分割的影响, 前景与背景融合部分生成结果较差	[46]
	特定领域生成法	人脸草图	通过建立人脸手绘草图与边缘图的匹配关系, 让边缘图训练的人脸生成模型适应手绘草图输入	提高输入手绘人脸草图时的生成效果, 满足生成图像的真实性和与输入草图的语义一致性	训练时大部分使用的草图形变程度较低, 在实际输入形变程度高的手绘草图时, 生成质量较差	[51–53]
			通过自监督模块捕捉人脸分区特征的长距离依赖	提高人脸整体结构的生成效果	仅考虑人脸的全局结构, 忽略人脸局部细节的学习	[54]
			利用人脸固定分区从局部到整体生成人脸图像	有效地捕捉人脸草图的局部细节特征, 生成高质量的人脸图像	未对草图存在形变和局部遮挡的情况提出解决方法	[20]
			增加属性编辑[56]或者采用风格编码[57]生成多样化的人脸图像结果	提高草图生成人脸图像的多样性	需要通过额外输入或者丢失草图的结构信息得到多样化的结果	[56,57]
		人脸毛发区域草图	通过将掩模和描述头发或胡须方向的草图输入至多个GAN网络生成发型或胡须	生成与草图描述方向一致的发型或胡须	忽略全局的发束连接和遮挡关系, 并且需要额外的掩模输入	[21,58]
			通过发型草图先生成对应的发型掩模, 再进一步生成发型图像	发型草图直接描述发型的结构和形状, 生成包括非编织发型和编织发型的多种复杂发型	无法生成包含深度信息的发型	[59]
		人体草图	将人体草图与描述人体姿态的关键点结合生成人体图像	利用草图生成结构和外观更加复杂的人体图像	输入的手绘草图与实际人体结构有差异时, 生成结果偏离真实图像	[22,60]

表 1 草图生成图像方法对比分析 (续)

方法分类	草图类型	方法描述	优点	缺点	相关工作
无监督学习	多类别单物体实例	基于CycleGAN[62]构建级联框架	实现开放域草图生成对应的图像	缺乏在多样化的复杂类别上进行训练和验证	[63]
		基于CycleGAN由草图生成灰度图, 再将灰度图映射为真实图像	实现无监督的物体实例草图生成多样化的结果		[64]
		利用草图和图像的自注意力图距离控制MaskGIT采样过程中掩码的更新		增加了采样迭代过程, 生成结果依赖于MaskGIT预训练模型的训练样本	[66]
	人脸草图	改进CycleGAN中的损失函数与鉴别器	实现无监督的人脸草图生成对应图像	训练使用的草图基本与图像对齐, 模型缺少对手绘草图的适应能力	[65]

图 3(c) 和图 7(b) 展示了监督学习方法和无监督学习方法的生成结果示例. 总之, 监督学习方法通过匹配的草图-图像对将生成模型的输入与输出联系起来, 形成草图对生成图像的约束关系, 生成的图像与草图相似性高, 但是草图的抽象性会导致图像生成的真实性下降, 难以得到较高分辨率的图像. Koley 等人[41]虽然提出了抽象草图生成高质量图像的解决方法, 但缺乏输入多种复杂类别草图的验证. 如何从包含多种复杂类别的抽象手绘草图里提取通用的语义内容, 并且在生成过程中既保持草图内容与图像的一致性又能提高生成图像的真实性是监督学习草图生成图像的重要挑战. 无监督草图生成图像工作虽然解决了草图-图像对缺失的问题, 但是仅支持输入简单的类别, 当草图结构复杂时生成效果显著下降, 大部分工作都基于 CycleGAN 框架, 除了文献 [64] 引入 AdaIN 结构以外其他方法给定源域的草图仅支持生成一种样式的图像, 根据输入的草图无监督生成高质量、多样化的图像是之后的研究重点.

除了通过 GAN 网络学习草图和图像领域的特征分布实现草图生成图像外, 现有的工作通过输入少量的草图样本调整预训练生成网络的参数权重, 实现生成的图像与草图形状和姿态保持一致. 其中 Wang 等人[69]将预训练的 GAN 网络生成的图像经过 Photo-sketching[17]生成对应的草图, 并利用判别器判别生成的草图和输入草图, 保证生成网络得到的图像与输入草图形状的一致性, 同时新增判别器判别原始生成模型的生成图像和训练图像以同时保证图像的生成质量和多样性. Israr 等人[70]在此基础上, 将原始生成模型的生成图像和目标生成模型的生成图像输入至新增的判别器中进行判别, 同时增加域间距离一致性损失让目标模型和源模型拥有相似的特征分布防止模型过拟合. 这些方法在草图控制生成特定风格和复杂动作的图像时结果较差, 但它们提出利用草图控制生成模型实现内容可控的多样化图像生成, 是提高草图生成图像多样性的创新研究方向.

1.3 结合风格控制的草图生成图像

图像会随着颜色与纹理的变化呈现不同的风格, 但手绘草图仅包含稀疏的视觉内容, 它可以大致描述物体的结构和布局, 但缺乏细粒度的纹理和外观描述, 在生成时会忽略图像风格的多样化. 因此, 目前的工作通过增加额外的风格控制输入让草图生成具有不同纹理和色彩的图像. 此外, 现有的工作利用额外的色彩控制输入对草图进行着色, 生成着色后的彩色绘画, 实现动漫、漫画和游戏的艺术创作. 本文将相关工作分为基于草图的真实图像风格化生成和草图着色. 结合风格控制的草图生成图像模型结构如后文图 8 所示. 图 8(a) 直接提取风格特征生成风格化图像[23,71]. 图 8(b) 对示例图像进行特征解耦, 提高风格化生成效果[24,72]. 图 8(c) 输入颜色控制实现对草图着色[73,74].

1.3.1 基于草图的真实图像风格化生成

风格输入从纹理和色彩两方面风格化草图对应的生成图像, 其中 Scribbler[75]和 TextureGAN[76]分别控制生成图像的色彩和纹理. Scribbler 通过输入草图和稀疏的彩色笔画在不同的草图区域进行对应的色彩控制, 生成用户指定颜色的逼真图像. 但是它在不同着色的区域边缘生成结果模糊, 并且使用的对抗损失函数严格限制了生成物体颜色和形状的真实性, 无法灵活生成不常见的结果. TextureGAN 输入草图和任意位置、尺寸的纹理块

控制生成图像的纹理, 但是纹理渲染至整张图像时效果不够自然, 并且当草图中包含多个纹理块时生成的图像在纹理衔接区域变化突兀, 与真实的图像差距较大. 上述方法无法同时决定生成图像的纹理和色彩, 所以现有的工作多采用输入示例图像的方式控制草图生成真实图像的整体风格. 本文将相关工作分为风格特征直接提取法和风格特征解耦法, 其中风格特征直接提取法通常直接提取示例图像的整体特征作为风格特征, 风格特征解耦法是指将图像解耦为语义内容和风格表示, 并实现在生成过程中由草图提供语义信息、示例图像提供与语义内容无关的风格特征.

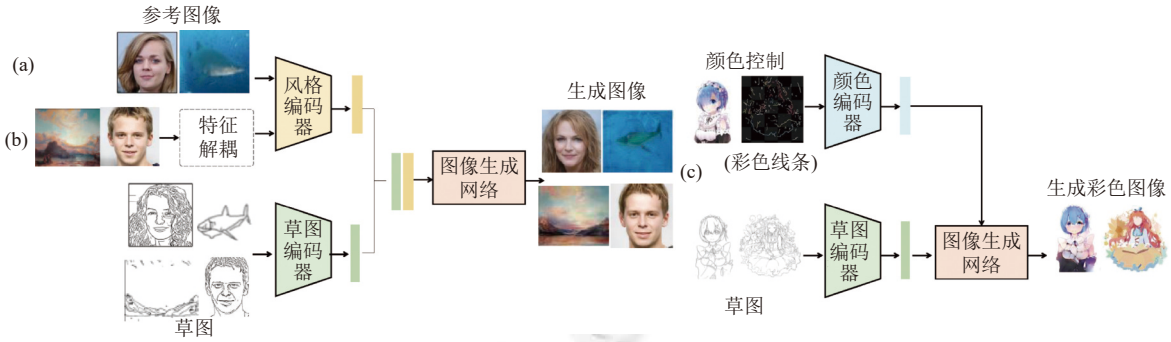


图 8 结合风格控制的草图生成图像模型结构示意图

1.3.1.1 风格特征直接提取法

风格特征直接提取法将示例图像输入现有的编码器中并将提取到的特征直接作为风格特征, 通过改进生成网络让生成结果与示例图像保持风格一致, 如图 8(a) 所示. Li 等人[77]输入示例图像控制草图生成艺术绘画风格的图像, 他们先输入草图生成图像然后对生成图像进行风格转化, 利用 VGG19 网络[78]分别提取生成图像和一系列示例图像的特征, 通过噪声图像与生成图像的内容特征和与示例图像的风格特征最小化重建风格化图像, 添加基于 PageRank 算法[79]的自适应权重方法将一系列示例艺术图像的风格特征更好的结合. CoGS[23]利用 VQ-GAN 中的编码器编码草图和示例图像, 之后通过 Transformer 学习在当前草图、示例图像和类别向量条件下真实图像的编码表示, 并增加风格损失函数让解码器重建的图像与示例图像的风格一致. 风格损失公式为:

$$\mathcal{L}_{\text{style}} = MSE(A(s), A(\hat{y})) \tag{6}$$

其中, s 是输入的风格图像, $\hat{y}$ 是生成的图像, $A(\cdot)$ 为计算的 ALADIN 风格特征向量. CoGS 还增加了 VAE 结构通过检索图像并在特征空间进行插值来微调生成结果, 生成更符合用户需求的风格化图像. MDSIT[80]直接通过编码器将获取到的示例图像特征作为风格特征, 并将草图的内容特征与示例图像的风格特征多次组合输入至解码器中, 提高图像的风格化效果, 该方法还支持输入不同绘制程度的草图生成风格化的图像. MichiGAN[81]通过输入示例图像、掩模和草图笔画, 实现在当前人脸背景不变的前提下控制局部发型的形状、结构、外观等多个属性的生成. 它包含的外观控制模块将示例图像的外观特征直接映射至输入的发型掩模中, 形状及结构控制模块输入掩模和方向图控制头发的形状和结构, 背景控制模块基于掩模的控制生成背景特征. MichiGAN 将各个条件控制模块逐步加入以 SPADE[38]为基础框架的生成模型中, 生成更换发型后的完整人脸图像. Yang 等人[71]基于 StyleGAN 网络由 W 和 W^+ 空间编码器分别提取草图的语义特征和结构特征, W^+外观编码器编码示例图像特征, 实现草图控制生成图像的语义和结构, 示例图像控制生成图像的外观. 他们还通过建立草图笔画与语义之间的对应关系提高生成模型对不同风格草图的适应能力.

上述工作仅提供全局的风格特征, 忽略示例图像和草图的局部特征对应. 因此, Lee 等人[82]对草图和示例图像进行编码, 添加注意力机制学习草图与示例图像之间的像素对应关系, 并通过基于相似度的三元组损失函数实现示例图像对草图进行正确着色. CoCosNet[83,84]等工作支持输入草图、语义分割图和姿态关键点等多种模态, 通过弱监督的方式建立输入和示例图像语义级别的对应关系, 生成风格与示例图像一致的目标图像.

1.3.1.2 风格特征解耦法

风格特征解耦法对图像特征进行分解并且通过一系列方式提取示例图像中与内容无关的风格特征, 确保生成图像与输入草图的语义一致性, 如图 8(b) 所示. Zuiderveld 等人[85]认为内容特征和风格特征相加可以得到完整的图像特征, 他们将草图特征减去草图风格特征得到草图内容特征, 之后与目标领域的风格特征结合生成对应的图像. 该方法使用文本-图像预训练模型 CLIP[86]作为草图和图像的编码器, 草图的风格表示由 Sketchy 数据集中与输入草图类别不重合的草图特征取均值得到, 目标领域的风格表示通过 ImageNet[87]中的图像或者文本编码得到, 利用文本-图像预训练模型实现零样本草图风格化生成图像.

Sketch2Art[72]通过参考图像、艺术家名字或者类别标签指定输入的风格, 生成风格化图像. 该方法在生成器前引入特征图转换模块, 通过特征池化及填充操作提取图像的显著特征作为图像的风格特征, 并且在判别器中采用 AdaIN 的逆过程解耦生成图像的风格特征和内容特征, 预测图像的风格及对应的草图内容, 提高生成的风格图像质量. 除此之外, Sketch2Art 在生成器中将双掩模注入层 DMI 加入前向传播过程中, DMI 基于草图掩模对草图边缘和空白区域的风格特征赋予不同的权重并重新输出特征图, 确保生成的图像轮廓清晰且构图正确.

Liu 等人[88]采用自监督的方法生成风格化图像. 他们为了解耦示例图像中的风格特征与内容特征, 对图像进行裁剪、水平翻转、旋转、缩放变化后将得到的多张图像一起输入风格编码器中得到风格向量, 并由三元组损失函数增强图像的风格一致性, 进一步提出动量互信息损失函数, 即利用交叉熵损失函数进行风格分类, 并确保输入内容特征时得到的类别概率一样从而无法进行正确的风格分类, 实现内容特征与风格特征的分离. 内容编码器与风格编码器类似, 对草图进行变化得到内容向量后由三元组损失函数自监督地训练编码器, 最后基于 GAN 网络生成风格化图像.

DeepFaceEditing[24]将人脸特征解耦为外观特征 (颜色和纹理) 和几何特征 (脸部形状及褶皱等其他脸部结构细节), 并利用草图控制几何特征、参考图像控制外观特征. 它先通过局部解耦模块分别得到人脸各区域的几何特征和外观特征, 外观编码器采用全局平均池化的方式编码示例图像中与几何结构无关的外观特征, 几何编码器得到输入草图的几何特征表示, 再将局部人脸特征进行全局融合生成完整的人脸.

风格特征解耦法通过对示例图像的语义内容和风格特征进行分离, 在生成时减少示例图像对生成结果的干扰, 有效地提高了生成的风格化图像的质量. 现有的工作[24,72,88]由于数据集规模的限制包含的风格类别是有限的, 而文献 [85] 采用零样本的方式生成风格特征, 打破当前风格化类别的局限, 是未来基于草图的风格化图像生成重要研究方向.

1.3.2 草图着色

如图 8(c) 所示, 当输入额外的色彩信息如彩色图像、颜色线条、色彩标签时, 现有的工作将色彩控制输入至生成网络中, 控制草图不同位置的着色并生成对应的彩色图像. Style2Paints V1[73]输入示例图像对草图进行着色, 模型将增强的 U-Net 和 AC-GAN[89]相结合, 它将 VGG19 网络提取到的示例图像风格特征添加至 U-Net 的中间层, 通过两个损失函数让中间层生成图像与真实图像最小化, 并且添加至 U-Net 中间层的入口和出口处, 解决 U-Net 在中间层梯度消失的问题, 提升草图着色的风格化效果. Style2Paints V3[90]将多个颜色块作为颜色控制指导草图着色, 该方法构建了基于 U-Net 的两阶段生成模型, 模型在第 1 阶段将输入的色彩组合并预测生成彩色草图, 第 2 阶段支持用户输入新的颜色指示细化生成结果, 检测和修复错误着色. 它有效地提高了生成结果并且支持用户控制颜色, 但增加了用户输入负担. Style2Paints V5[74]对草图进行平面色彩填充, 在不同的区域填涂固定颜色. 它采用分割填充机制将用户输入的彩色线条进行分组, 再将每一组的颜色填充至对应的区域内, 生成符合用户感知的绘画. Tag2Pix[91]输入多个颜色标签指导草图着色, 它通过结构编码器编码草图的结构特征, 颜色编码器编码颜色标签的颜色特征, 生成器基于 U-Net 结构将中间层特征与结构特征拼接输入解码器, 颜色特征基于 SENet[92]与中间层特征拼接, 生成高质量的着色结果. 它在解码器的第 1 层添加引导解码器, 通过新的损失函数解决梯度消失问题, 采用两阶段训练方法并在第 2 阶段添加分类损失函数对颜色标签进行多分类预测, 提高着色效果. 现有的草图着色方法在处理包含复

杂背景元素的草图时, 模型在前景物体和背景的边缘过渡处将产生着色误差, 因此未来的草图着色改进方向可以建立草图和颜色控制局部结构和语义的对应, 并在生成过程中考虑色彩的渐变以及色彩的亮度变化.

1.3.3 讨 论

如表 2 所示, 现有的方法一般输入彩色笔画、局部纹理块和示例图像控制草图生成真实图像的纹理和色彩. 在输入示例图像时, 风格特征直接提取法未对示例图像的特征做出其他处理, 当输入多种领域的风格图像时示例图像会弱化草图提供的语义内容, 导致生成的图像在形状与结构上与输入的草图不一致. 采用解耦的方法从示例图像中提取与内容无关的风格特征, 从草图中提取内容特征, 有效地提高了风格化图像的生成质量. 在草图着色任务中, 现有的方法输入示例彩色图像、颜色块、彩色线条、色彩标签等多种方式控制草图着色. 在未来的研究中, 结合风格控制的草图生成图像可以侧重于增加风格控制的输入元素如文本, 让用户可以自定义生成图像的纹理和色彩并且通过修改草图及风格控制元素对图像的语义和风格进行编辑, 还可以继续改进生成模型以精准地解耦风格控制输入的风格特征以及草图的语义内容, 并且从局部到全局细粒度地建立草图和风格控制的对应关系, 灵活生成逼真的多样化风格图像.

表 2 结合风格控制的草图生成图像方法对比分析

任务	风格控制输入	控制内容	风格编码方式	具体风格化方式	数据集的草图类型	相关工作
基于草图的真实图像风格化生成	彩色笔画	颜色	—	输入彩色笔画对草图区域正确着色	室内场景、人脸、车	[75]
	纹理块	纹理	—	输入多个局部纹理块控制生图像的纹理	包、鞋子、衣服	[76]
	示例图像	颜色和纹理	直接提取法	基于PageRank算法将一系列示例艺术图像的特征与草图生成图像后的特征相结合生成风格化图像	猫、建筑	[77]
				基于VQ-GAN提取示例图像的特征, 基于风格损失函数实现解码器重建的图像与示例图像的风格一致	125个类别物体实例	[23]
				编码器编码示例图像特征, 并与草图的内容特征多次组合输入至解码器中	人脸	[80]
				示例发型图像的外观特征映射至输入的发型掩模中	发型	[81]
				基于StyleGAN网络的 W^+ 外观编码器编码示例图像特征	人脸	[71]
				基于注意力机制建立草图和示例图像素级对应关系, 并对草图进行着色	动画人物、猫、狗、车、人脸、鞋子	[82]
				建立草图和示例图像语义级别的对应关系	语义图、人脸、人体骨架	[83,84]
			解耦法	草图特征减去草图风格特征, 得到草图内容特征再与目标领域风格特征相加, 得到完整目标领域特征	多类别物体实例	[85]
				特征图转换模块提取图像的风格特征, 在判别器中利用AdaIN的逆过程解耦生成图像的风格特征与内容特征并分别预测	风景	[72]
				利用动量互信息损失函数训练风格编码器, 解耦示例图像的风格特征和内容特征	风景、人脸	[88]
				人脸特征解耦为外观特征和几何特征, 将示例图像全局平均池化得到外观特征	人脸	[24]
草图着色	示例图像	颜色	—	将VGG19网络提取到的示例图像颜色特征添加至生成模型U-Net的中间层	动漫角色	[73]
	颜色块			将多个颜色块和草图输入两阶段生成模型进行草图着色	动漫角色	[90]
	彩色线条			对彩色线条进行分组和草图输入生成模型进行草图着色	艺术线条画	[74]
	颜色标签			颜色编码器编码颜色特征并与草图编码器的中间层特征拼接输入解码器	动漫角色	[91]

1.4 基于草图的图像编辑

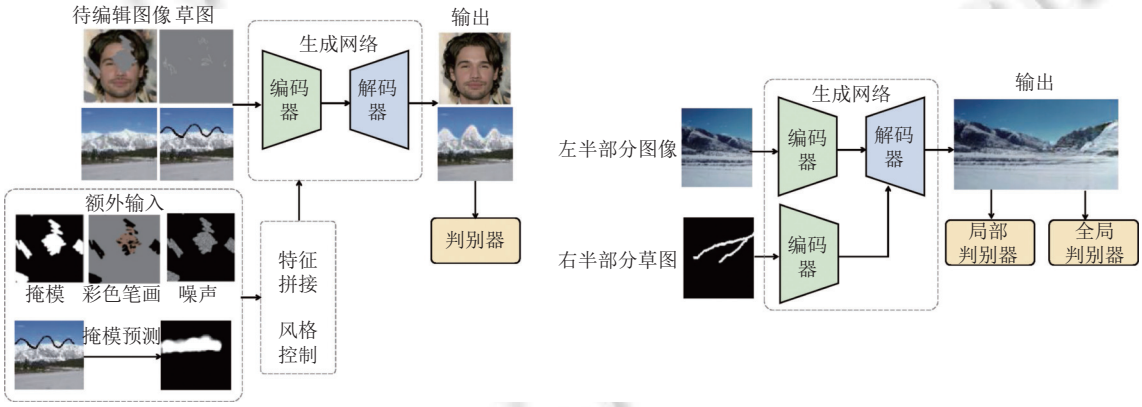
基于草图的图像编辑模型如图 9 所示[25,26,93], 包括基于草图的图像局部编辑和基于草图的图像扩展.

1.4.1 基于草图的图像局部编辑

草图在局部编辑图像时, 通过草图笔画修改图像的局部结构, 生成新的图像, 在编辑过程中要保持编辑区域和其他区域的语义关联及风格一致性, 如图 9(a) 所示. FaceShop[94]构建草图局部编辑人脸图像的系统, 模型的输入除了局部草图外, 还有彩色笔画、旋转的矩形掩模和被遮挡的图像, 图像补全网络由多层 CNN 构成, 其中生成器为编码器-解码器结构, 判别器包括局部判别器和全局判别器, 基于 WGAN-GP 损失函数[95]补全生成完整的人脸图像, WGAN-GP 损失函数为:

$$\mathcal{L}_{\text{WGAN-GP}} = \mathbb{E}\left[D\left(\hat{I}\right)\right] - \mathbb{E}\left[D(I)\right] + \lambda\mathbb{E}\left[\left(\left\|\nabla_{I_u} D(I_u)\right\|_2 - 1\right)^2\right] \tag{7}$$

其中, $\hat{I}$ 是生成样本, 包括被编辑生成的图像、草图、颜色约束以及编辑区域的掩模, I 是真实样本, 包括原始图像和随机掩模, I_u 是 I 和 $\hat{I}$ 之间沿直线均匀采样的数据点, λ 是超参数. FaceShop 将掩模输入至局部判别器和全局判别器时, 局部判别器会对局部区域大小进行放缩造成一定的信息丢失, 并且模型在擦除区域过大时生成的图像结果不合理. 所以 SC-FEGAN[93]通过优化网络结构, 实现输入草图、颜色和任意形状的掩模进行人脸图像编辑. 图像补全网络的生成器与 U-Net 结构类似由门控卷积层构成[96], 鉴别器为 SN-PatchGAN[96]实现对任意形状掩模的生成, 除了 WGAN-GP 损失函数以外还增加风格损失函数最小化生成图像和目标图像特征图的距离, 保证较大的人脸区域进行编辑时的生成效果. Deep plastic surgery[97]利用不同绘制程度的手绘草图进行人脸图像编辑, 它提出草图细化方法将粗粒度的手绘草图进行边缘扩张形成绘制区域, 并在绘制区域内由粗粒度草图生成对应的细粒度边缘图, 草图的细化程度与扩张半径有关, 基于 AdaIN 将控制草图细化程度的扩张半径参数编码为风格向量控制草图的生成粒度. 在编辑阶段, 以 Pix2Pix 为框架, 输入待编辑的图像、掩模以及草图生成完整的图像. 当图像被掩模全部掩盖并用草图进行描述时, 就转换成了草图生成图像的问题. 这些工作将草图和待编辑图像直接映射到特征空间, 草图的稀疏特征无法得到有效提取导致编辑区域补全效果下降. 因此 DeFLOCNet[98]在编码器和解码器的多层跳跃连接中添加草图和颜色控制, 增强图像的结构特征编码, 并添加额外的解码器生成纹理特征, 对编辑区域进行补全.



(a) 基于草图的图像局部编辑 (b) 基于草图的图像扩展

图 9 基于草图的图像编辑模型示意图

上述工作除了 DeFLOCNet 支持风景图像编辑外, 主要面向人脸图像的编辑, 并且需要输入草图以及修改区域的掩模, 掩模会造成较大的输入负担并且遮挡上下文内容的信息. SketchEdit[25]面向人脸、其他类别的物体实例和风景图像的编辑, 它直接输入局部草图并预测对应的掩模来对图像进行编辑. 在训练过程中, 如图 9(a) 所示, SketchEdit 将草图和形变的图像输入至掩模预测网络中, 并通过双向掩模归一化损失函数分别利用形变图像和草

图重建原始图像以及原始图像和形变草图重建形变图像, 实现对掩模区域进行正确的预测. 然后它将草图、预测的掩模区域和被编辑的图像以及掩模区域的风格特征一并输入至生成网络中, 利用两个生成器从粗到细的生成图像并让图像的编辑区域与图像的整体风格保持一致.

1.4.2 基于草图的图像扩展

基于草图的图像扩展通过输入图像和左半部分的草图生成图像的右半边部分内容, 得到完整的图像, Wang 等人[26]提出草图对风景图像进行扩展. 如图 9(b) 所示, 在训练阶段, 生成器为基于 LSTM[99]的编码器-解码器结构, 编码器编码左半部分图像和对应的草图以及右半部分草图, 他们在生成模型中加入位置通道建立风景图像空间位置与语义信息像素级的对应联系, 并在解码器中加入跳跃连接共享右半部分的草图特征, 除了局部判别器和全局判别器判别生成图像的真实性, 还引入额外的草图对齐损失函数实现生成图像的边缘图和输入的草图距离最小, 确保生成的右半部分图像与草图的结构一致性. 该方法允许用户根据个人偏好使用手绘草图控制图像扩展的内容, 但是图像扩展生成的部分分辨率低, 生成质量一般.

目前基于草图的图像编辑工作中草图编辑的图像结果在区域边缘仍存在不真实性, 在生成时要进一步考虑编辑区域与其他区域在语义及风格上的统一. 除此之外, 掩模作为编辑任务的一个重要输入, 当面向复杂的编辑情况时如同时编辑多处区域并且区域有重叠, 或者草图线条的编辑范围覆盖整张图像时, 掩模将会造成图像编辑产生较大的误差, 因此利用草图线条本身提供的结构信息直接建立草图和待编辑图像之间的语义关联进行图像编辑是后续研究的重点.

2 基于草图的视频生成

基于草图生成视频模型如图 10 所示, 其中图 10(a) 为草图引导视频生成[27,100,101], 图 10(b) 为草图编辑视频[28], 图 10(c) 为草图视频着色[29,102].

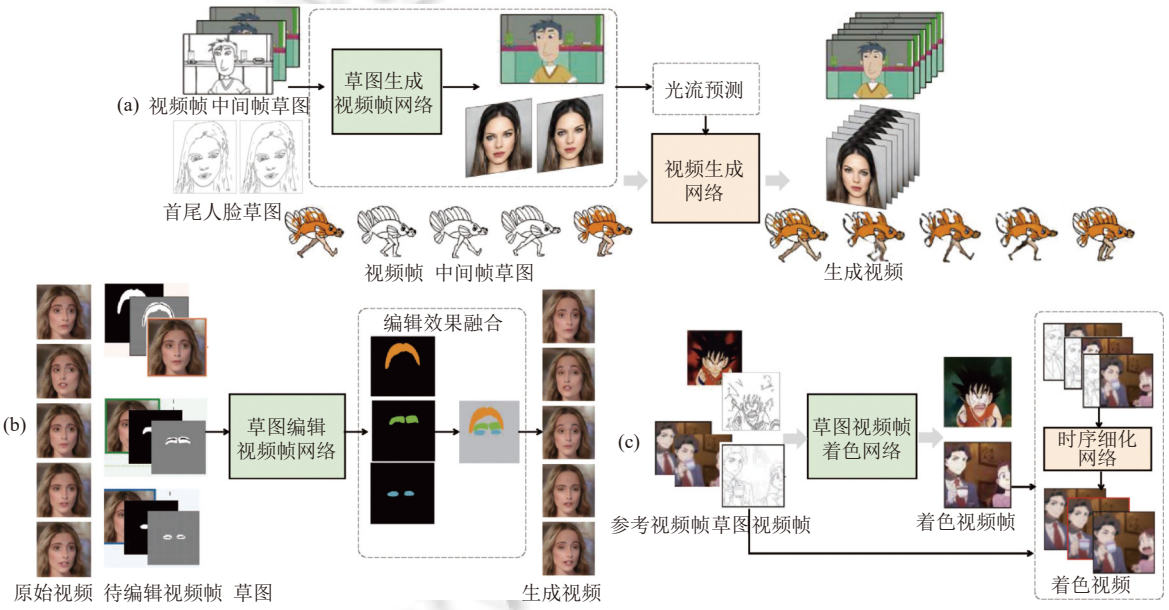


图 10 基于草图生成视频模型示意图

2.1 草图引导视频生成

如图 10(a) 所示, 草图可以描述完整的视频帧内容, 并通过与前后视频帧相衔接描述物体姿态的变化过程, 传递对应的时序信息, 引导生成新的视频. Li 等人[27]提出由用户输入中间帧草图, 引导动画视频的中间帧插值生成.

他们通过中间帧草图和首尾视频帧进行光流预测生成中间视频帧, 然后采用任意时刻视频插值方法生成更多的中间视频帧, 最后通过时序处理方法生成时序一致性的视频. Zhang 等人[100]输入首尾两张人脸草图生成对应的视频帧序列. 他们采用两阶段生成方法, 先通过局部编码器编码数据集中人脸草图局部区域的特征, 再采用 LLE 算法[55]检索得到各个人脸草图区域在语义空间的特征映射, 然后基于 Pix2PixHD[103]的框架构建局部生成器生成真实的起始帧和结束帧, 并将草图的语义特征序列送入第 2 阶段. 第 2 阶段基于 cVAE[104]结构将草图和训练视频 (测试阶段视频服从标准正态分布) 作为条件输入映射至运动空间得到对应的运动变量, 之后将运动变量与第 1 阶段得到的草图的语义特征相结合并通过视频解码器, 输出带有掩模的光流序列并生成最终的视频. 由于视频中的运动可能要比草图表示的更复杂, 为了让草图和视频的运动分布尽可能接近, 模型加入平滑运动损失函数最小化光流序列和首尾两帧线性插值的光流距离. SketchBetween[101]是视频转换的子任务, 它输入描述物体外观的关键帧以及描述运动的中间草图组成的视频序列, 通过 VQ-VAE[105]模型生成完全渲染的动画, 实现二维动画的生成. 但是动作生成依赖于训练集, 在测试阶段出现不常见的动作时生成效果将显著下降.

2.2 草图编辑视频

草图在编辑视频时, 先输入局部的线条编辑视频帧, 然后将编辑结果映射至整个视频序列, 得到完整的新视频 (如图 10(b) 所示). DeepFaceVideoEditing[28]基于 StyleGAN 框架利用草图局部编辑人脸视频, 包括时序一致性编辑和动态编辑两种编辑方式. 该方法先将输入图像映射为隐向量 w, 再输入草图和掩模得到编辑后的图像并映射为代表草图结构编辑的新的隐向量 w_{edit}, 然后将所有视频帧映射至隐空间得到隐向量序列并生成编辑后的视频. 时序一致性编辑指草图仅编辑人脸形状, 将编辑效果直接映射至所有的视频帧. 动态编辑包括面部运动编辑和特定表情编辑, 面部运动编辑通过指定时间窗口, 采用分段线性函数得到每张视频帧的编辑向量, 生成平滑的面部运动; 特定表情编辑通过三维重建的方法提取每张视频帧的表情向量生成对应的编辑向量, 实现表情编辑. 当对多张图像进行编辑时, 它按照多个编辑的区域进行特征映射生成编辑后的视频.

2.3 草图视频着色

随着动漫产业的发展, 为了降低中间草图帧上色的成本, 现有的工作采用深度学习方法自动着色一系列视频帧, 在着色时保持相邻视频帧着色的一致性. 如图 10(c) 所示, TCVC[102]提出保持时序一致性的草图视频着色方法, 它将 Pix2Pix 扩展至视频领域, 将当前草图视频帧和前一帧的视频着色作为条件输入, 实现生成过程中视频帧着色随着时序变化保持一致. 它通过 VGG19 提取生成彩色视频帧和真实彩色视频帧的特征, 新增内容损失函数最小化它们之间的特征距离, 以及风格损失函数最小化特征的格拉姆矩阵距离. 但是逐帧生成的方式会造成着色的误差积累. 因此, Shi 等人[29]提出多张彩色帧作为参考指导草图视频着色, 并采用从局部到全局的方式建立待着色草图视频帧和参考视频帧之间的对应关系. 模型包括颜色转化网络和时序细化网络, 颜色转化网络通过距离注意力层, 将目标草图帧和参考视频帧映射为距离图, 通过计算草图帧和参考视频帧的特征距离建立目标视频帧和参考视频帧相同位置特征的对应关系, 获取目标草图帧的局部颜色特征, 再通过基于 AdaIN 的颜色编码器编码参考视频帧的全局风格保证全局的色彩一致性, 时序细化网络将参考和目标的草图-视频帧对输入基于 3D U-Net 的三维卷积网络, 生成按时间顺序的彩色图像序列, 细化着色结果.

视频生成是视觉内容生成的另一项重要内容, 视频基于空间和时序层面的视觉信息形成的复杂数据结构增大了视频生成的难度, 因此基于草图的视频生成工作目前仍处于初步探索中. 草图引导视频生成仅利用草图生成接近线性的简单动作, 但草图具有描述复杂姿态变化的能力, 如何利用草图引导更加复杂的动作变化生成对未来研究有重要挑战. 草图编辑视频目前局限于编辑变化幅度较小的正面人脸视频, 当人脸视频为其他朝向、面部变化复杂或者局部区域存在覆盖情况时将严重影响生成效果, 未来应扩大草图编辑视频的范围并利用草图灵活地编辑视频中的语义内容和时序信息. 草图视频着色在中间视频帧新增物体时无法得到较好的结果, 尽管 TCVC 可以通过重新训练的方式解决该问题, 但训练难度增大, 模型缺乏足够的泛化性能, 并且在细节处存在着色错误或者色彩溢出, 未来可以通过分割的方式进行实例对齐后对视频帧进行着色, 并且添加风格预测解决新物体出现的视频帧着色.

3 数据集介绍

3.1 常见数据集

由于手绘草图存在一定的绘制难度, 目前具有成对示例的数据集主要为手绘草图-图像数据集, 数据集的类别有限并且规模较小. 许多方法对草图进行扩充, 利用现有的图像数据集和视频数据集生成伪草图[23]构建匹配对. 表 3 列举了基于草图的视觉内容生成任务中用到的草图-图像数据集以及经常使用的图像数据集和视频数据集, 图 11 展示了各个数据集的示例.

表 3 基于草图的视觉内容生成中常见的数据集

类别	名称	内容	相关工作
草图-图像数据集	SketchyCOCO	17种背景物体类别, 14 081张场景草图-图像对, 47 881个物体草图-图像-边缘图匹配对	[46]
	ShoeV2[106] ChairV2[106]	6 648张草图, 2 000张鞋子图像, 笔画级、多对一的草图-图像对 1 297张草图, 400张椅子图像, 笔画级、多对一的草图-图像对	[41,64]
	Photo-sketching	1 000张户外图, 5 000张轮廓草图, 1张图像对应5张草图	[17]
	Sketchy[107]	125个类别, 12 500张图像, 75 471张草图, 一张图像对应5–6张草图	[19]
	Sketch2Hair	640个发型草图-图像对, 草图描述发束方向	[21]
	SketchHairSalon	4 500个发型草图-发型掩模-图像匹配对, 草图笔画包含颜色信息	[59]
图像数据集	CUFS[108]	606张人脸图像-草图对	[34–37]
	CUFSF[109]	1 194张人脸图像-草图对	[34–37]
	CelebA[110]	20万张人脸图像	[40]
	Caltech-UCSD Birds-200-2011[111]	11 700张鸟类图像	[40]
	Stanford's Cars[112]	16 000张车辆图像	[40]
	WikiArt[113]	1万张包含55中艺术风格的图像	[72]
	CelebA-HQ[114]	3万张人物肖像图像	[25,42,51,53,54,80,88,93,94,97,98]
	CelebAMaskHQ[115]	3万张人物肖像图像及人脸各属性对应的掩模	[20,24,56,80,100]
	FFHQ	70万张人脸图像	[70,80]
	Flickr-Faces-HQ	7万张肖像图像	[81]
	Danbooru2017[116]	2.94万张绘画图像	[74,90,91]
	NS6K[117]	6 040张自然风景图像	[26]
	DeepFashion[118]	52 712张身着不同服装的人体图像	[22,60]
	Places2[119]	180万张图像和365个场景类别	[98]
视频数据集	VoxCeleb2[120]	150 480个说话人脸视频	[100]
	MGIF[121]	1 000个卡通动物走路、跑和跳的视频	[101]

常见的草图-图像数据集采用人工绘制的方式绘制图像对应的草图, 而 SketchyCOCO[46]和 SketchHairSalon[59]采用以下方式构建数据集.

● SketchyCOCO[46]包含 14 个前景物体类别和 3 个背景元素类别, 草图和图像分别从已有数据集 Sketchy[107], Tuber-lin[122], QuickDraw[123]和 COCO-stuff[124]中筛选. 为构建场景草图, 将背景元素 (云、草和树) 的草图随机填充在对应图像掩模的位置, 前景物体通过类别检索与图像最相似的草图形成图像-草图-边缘图匹配对.

● SketchHairSalon[59]收集不同视角的肖像画, 然后分割发型区域并自动获取对应的发型掩模, 让绘制人员通过草图笔画绘制发型结构.

CelebA-HQ[114]数据集是草图在人脸生成任务中最常使用的一个数据集, 现有的工作在基本的人脸跨域生成[42,51,53]、人脸局部编辑[25,93,94,97,98]和人脸风格转化[88]等任务的基础上, 结合草图本身的特点提出新的生成任务, 如利用草图

具有形变的抽象线条在实际生成中纠正形变线条提高生成图像的真实性[51], 或者控制线条形变程度实现可控的图像生成[53]或图像编辑[97], 还可以利用草图的绘制过程先补全推荐草图再生成图像[42]. 图 12 列出不同生成任务在 CelebA-HQ 数据集下的生成结果对比图.

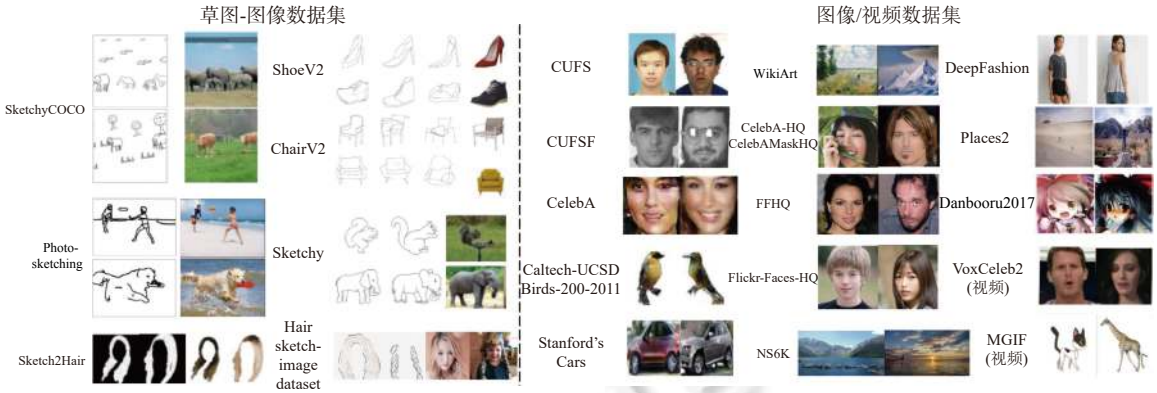


图 11 基于草图的视觉内容生成相关数据集示例图

图 12 基于 CelebA-HQ 数据集的生成结果对比图

3.2 草图数据扩充方法

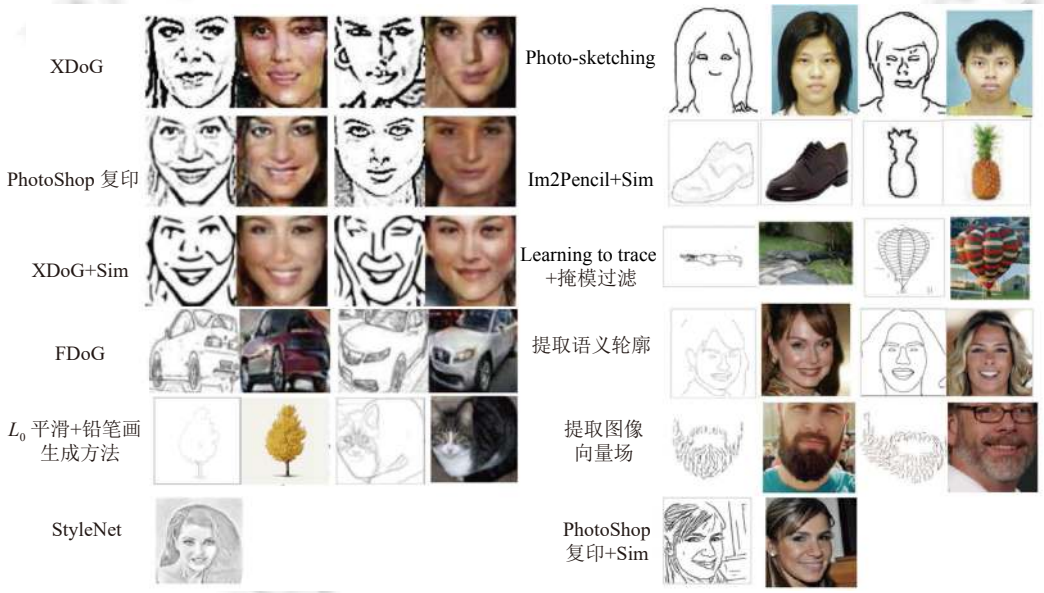
目前草图数据扩充方法包括提取图像边缘得到草图和利用生成模型生成草图两大类. 常见的提取图像边缘生成草图的方法主要为 HED 边缘检测器[3]、PhotoShop 复印[40]、XDoG 边缘检测器[4]、FDoG 过滤器[5]和铅笔画生成方法[6], 常见的利用生成模型生成草图的方法有 Photo-sketching[17]、Im2Pencil[125]、StyleNet[126]、Learning to trace[127]等, 这些方法提取到的边缘突出主体的语义与结构并具有清晰的轮廓, 但是线 条数量较多, 存在很多语义无关的多余线条及模糊阴影, 并且在轮廓上和图像基本对齐, 与实际的手绘草图相距较远. 在处理人脸草图时, 文献 [25,51]提取语义图的轮廓得到轮廓清晰简洁的草图, 但同时也丢失了草图描述的细节特征. 在生成发型草图时, 文献 [58]提取图像向量场生成, 但它无法描述复杂的发型结构. 为了让获得的草图更接近手绘草图, 现有的方法会采取不同的预处理和后处理方式 (如表 4 所示) 处理生成结果, 得到线条平滑并且与手绘草图风格接近的伪草图.

表 4 草图数据扩充方法中额外的预处理和后处理方式

	方法	特点	具体方法	边缘提取方法	代表工作
边缘预处理	边缘加强	增加边缘部分的显著性	加权最小二乘滤波器	HED	[45]
	图像平滑	提取图像形状	L_0平滑算法	铅笔画生成方法	[77]
	掩模过滤	减少背景对物体边缘的干扰	利用高斯核区分实例掩模和背景区域	Learning to trace	[23]
边缘后处理	线条形变	形变边缘图以接近真实的手绘草图	采样层[128]形变	HED	[53]
			AutoTrace[129]矢量化线条图并形变线条	HED	[94]
				提取语义图的轮廓	[51]
			移动最小二乘策略	HED	[45]
	线条细化	将图像变为线条图边缘	线条二值化及细化[130]	HED	[19,60]
			直接线条细化		[39,56]
	线条过滤	删减多余的线条	随机擦除	HED	[53]
			0–3个矩形区域		
			阈值过滤线条		[19,60,94]
			去除独立连接的区域		[19,60]
	线条简化	得到结构清晰且平滑的草图	平滑线条的控制点	HED	[94]
			Mastering sketching[131]	HED	[100]
			Sim[132]	HED	[60]
				PhotoShop复印	[20,24]
				XDoG	[40,52]
				Im2Pencil	[22,42]

为了减少输入边缘图和输入草图时模型生成的效果差异并且提高模型对不同风格的手绘草图的学习能力, 目前的方法通过多种风格的草图输入来增强模型的鲁棒性, 如利用几类草图数据扩充方法处理同一数据集, 每种扩充方法作为一种风格[40,52,75,80], 而 TOM[88]构建了无监督的图像生成草图模型, 在训练时输入提供内容的图像和提供目标风格的草图并在一个 batch 里进行随机匹配, 实现一张图像生成对应的多种风格的草图.

图 13 展示了草图数据扩充方法的结果示例, 现有的草图数据扩充方法增加了伪草图的抽象性和多样性, 极大程度地缩小生成的伪草图与真实手绘草图之间的差距, 是草图数据的重要补充.



(a) 不同边缘提取方法及对应预处理和后处理方式组合

图 13 草图数据扩充方法及不同预处理和后处理方式组合结果示意图

(b) HED 边缘提取方法及不同预处理和后处理方式组合

图 13 草图数据扩充方法及不同预处理和后处理方式组合结果示意图 (续)

4 评估方法

视觉内容生成的评估方法主要从定量评估和定性评估两个方面对生成结果进行合理的评估. 定量评估通过一系列评估指标量化生成结果的质量和多样性, 但它有时与实际人类感知的情况存在差异导致对生成结果的评估存在一定误差. 定性评估让用户对生成结果进行打分, 采用人工的方式判断生成结果的质量, 人为的判断在判断标准的设定上存在较强的主观性, 当图像逼真度很高时定性评估将不再准确. 所以现有的方法将定量评估与定性评估结合起来使用, 综合地衡量生成结果.

4.1 定量评估

定量评估使用的评估指标利用生成图像的像素、深度特征以及图像处理相关的任务, 通过多种算法或者模型对生成结果量化评估. 目前基于草图的视频生成任务是对生成的视频帧进行定量评估, 因此与图像生成使用的评估指标一致. 以下列举了常见的评估指标.

● Inception score (IS)[133]使用在 ImageNet 上预训练的 Inception-v3[134]网络来计算条件分布和边缘分布之间的 Kullback-Leibler (KL) 散度, IS 值越高代表图像的质量和多样性越高, 但它对变化感知敏感导致分数不够准确.

● Frechet inception distance (FID)[135]利用 Inception-v3 提取特征, 计算生成图像的特征分布与真实图像的特征分布距离, FID 可以衡量生成图像的质量, 分数越低代表生成图像的分布越接近真实图像.

● Kernel inception distance (KID)[136]通过计算最大均值差异的平方评估 Inception-v3 网络提取的生成图像和真实图像之间的差异, KID 值越低代表生成结果越好.

● 学习感知图像块相似性 (LPIPS)[137]使用深度特征度量生成图像与真实图像之间的相似度, LPIPS 数值越低表示两张图像越相似, 数值越大代表生成图像的多样性越高.

● 结构相似性差异 (structural similarity metric, SSIM)[138]用来评估生成图像和真实图像之间的结构相似性, 数值越大代表生成图像与真实图像越相似.

● 特征相似性 (feature similarity index metric, FSIM)[139]用来评估生成图像和真实图像的相位一致性特征和梯度特征的相似性, 数值越大代表生成图像越相似.

● 峰值信噪比 (PSNR) 用来评估图像失真程度, 值越大代表失真程度越低, 生成图像的质量越高.

● FCN 分数[140]将生成图像进行语义分割, 通过像素精度 (per-pixel accuracy)、类别精度 (per-class accuracy) 和类别的 IoU 值 (intersection over union) 比较生成图像和真实图像的语义标签.

● Style relevance (SR) 计算生成图像和输入的风格图像低层感知特征的距离, 衡量生成图像的风格一致性.

● 形状形似性 (accuracy SS)[46]计算输入草图和生成图像形成的边缘图之间的特征距离, 是结合草图输入提出的新评价指标, 分数越低代表图像对输入草图的忠实度越高.

PSNR、SSIM、FSIM 基于像素层面进行计算, 类似的指标还有 L_1 损失[25,27]和 L_2 损失[93]. Xiao 等人[59]使用绝对差值总和 (SAD) 来衡量发型掩模生成的准确率以及 IoU 衡量发型掩模边界区域的准确性, 它们主要用来评

估图像生成的质量, 在数据集包含的图像底层特征较为相似时, 无法进行精确的度量. IS、FID、KID、LPIPS、FCN 分数等指标基于深度学习模型评估生成图像的质量以及多样性, 更能捕捉图像之间的特征差异性以及代表用户实际的视觉感知. 如表 5 所示, FID 是应用最广泛的指标, 在像素层面的指标如 PSNR、SSIM 等相差范围较小时, FID 仍可以明显地区分各种方法生成效果的差异, 与人类视觉感知相匹配. 除了从特征层面和像素层面评估生成图像, 现有的方法还会结合具体的图像处理任务评估生成结果. 分类准确率是指将生成的图像用于分类任务中, 准确率越高代表生成的图像质量越好. 在图像生成草图任务[18,32]中, 将生成草图用于细粒度图像检索任务中, 由检索精度验证生成草图的质量.

表 5 不同数据集中生成方法的定量评价

数据集	模型	FID↓	LPIPS↓	FSIM↑	IS↑	PSNR↑	SSIM↑
CUFS[108]	Sketch-Transformer[34]	20.92	0.3019	0.7350	—	—	—
	SCA-GAN[35]	34.2	—	0.716	—	—	—
	SDGAN[36]	33.408	0.2767	0.7446	—	—	—
	EADT[37]	20.97	0.178	0.742	—	—	—
CUFSF[109]	Sketch-Transformer[34]	9.39	0.3400	0.7259	—	—	—
	SCA-GAN[35]	18.2	—	0.729	—	—	—
	SDGAN[36]	30.594	0.3358	0.7328	—	—	—
	EADT[37]	27.12	0.144	0.775	—	—	—
CelebA-HQ[114]	Controllable sketch-to-image translation[53]	113.669	—	—	—	—	—
	DeepFacePencil[51]	242.1	—	—	2.411	—	—
	SC-FEGAN[93]	—	—	—	—	31.1687	0.9671
	DeFLOCNet[98]	9.92	—	—	—	25.42	0.90
	SketchEdit[25]	0.844	—	—	—	34.15	0.9604
	MDSIT[80]	56.264	—	—	—	—	—

4.2 定性评估

定性评估通过众包的方式让用户对生成图像或者视频的质量进行比较, 主要包括生成结果的真实度、生成结果的忠实度 (与输入草图的一致性) 以及生成模型的可用性评估. 目前的方法在评估生成结果的真实度时, 让用户在生成图像与真实图像之间判定真假[42,62,65], 按真实程度进行打分[26], 或者给定草图及一系列对比方法的生成结果从中选择更加真实的结果[19,20,23,46,53,64], 在结合风格控制的草图生成图像任务中是选择风格着色至草图效果最好的结果[72]; 评估生成结果的忠实度时, 给定生成图像和多个草图, 选择与图像一致的草图[19,20,46,69]. 在评估生成模型的可用性时, 主要衡量输入手绘草图时是否可以得到质量较高的生成结果, 让没有绘制经验的用户输入手绘草图至构建的生成系统中例如让用户通过手绘草图更换发型[59]、利用手绘草图编辑图像[94]等并且评估生成结果, 或者挑选绘制水平较差的草图生成的视频结果[100]进行评估.

5 总结和展望

草图具有直观表达、灵活输入的特性, 它可以描述常见的多种类别的物体以及结合实际应用描述人脸、发型和人体等特定类别, 或者以笔画线条的方式描述局部结构内容, 所以目前的工作利用草图的多种形式和具体的视觉内容生成任务如跨域生成、风格转化、视觉内容编辑等对草图在视觉内容生成中的应用进行了深入的研究, 解决草图由于稀疏性、抽象性和风格多样性增加了视觉内容生成难度的问题, 但是在实际应用中仍然面临手绘草图收集难度高, 以及输入多样化的手绘草图时难以生成高质量的图像或视频的问题. 因此本文对草图在视觉内容生成中面临的挑战进行分析并提出改进策略, 然后对未来研究方向进行展望.

5.1 基于草图的视觉内容生成挑战

(1) 手绘草图的获取难度

为提高图像或视频的生成效果, 基于草图的视觉内容生成模型需要大规模的草图-图像数据集或者草图-视频数据集作为支撑. 采用人工绘制的方式可以收集真实手绘草图并且保留草图多样化的风格, 但花费的成本高且数据规模有限. 目前多数工作采用草图数据扩充方法得到结构简单、存在形变且线条平滑的伪草图代替真实手绘草图进行训练. 但现有的方法利用边缘提取方法或者图像生成草图模型输入图像得到伪草图时都进行的是整体处理, 忽略了局部的语义信息, 例如场景草图中前景的多个物体实例以及背景之间包含着不同的语义信息, 如果采用现有的方法对场景图像直接进行处理会得到大量的冗余线条并且丢失场景语义信息, 如果采用实例分割的方法得到每个物体对应的草图会缺失场景整体的风格一致性. 而且目前的伪草图与真实手绘风格仍有较大差距, 手绘草图在绘制物体或场景的局部区域时会根据用户绘制习惯以及当前区域视觉特征的稠密程度产生不同密度的线条分布, 例如人脸的眼睛与眉毛、嘴巴等区域相比线条分布数量更多, 这是目前伪草图生成时未考虑的内容. 未来草图数据扩充的研究重点为通过改进图像生成草图模型让生成的草图具有语义信息并且更接近实际手绘草图的线条分布, 例如模型可以采用从局部到全局的方式生成既包含语义信息又具有绘制风格的草图.

(2) 跨域差异

与图像稠密的视觉表示不同, 草图具有抽象且稀疏的表示, 直接提取草图特征并与图像特征对齐难度较高, 降低了跨域转化的生成效果. 因此很多方法选择增加边缘图[52,53]作为中间生成结果, 将稀疏的手绘草图先映射为特征稠密的边缘图再由边缘图生成图像, 或者增加掩模控制生成区域[59]将草图和掩模结合起来生成图像. 在未来工作中, 通过增加中间模态建立草图和图像之间的特征关联, 对草图特征进行补充以减小草图与图像之间的特征差异, 是提高图像生成质量的研究方向. 为了从语义内容和时序信息两方面缩小草图与视频的跨域差异, 未来的研究中可以采用局部到全局的方式建立草图和视频帧之间细粒度的语义对应, 并通过额外的草图笔画控制运动方向或增加草图序列的数量描述复杂运动的过程, 建立草图与视频时序信息的对应关系, 实现草图细粒度地控制视频生成的内容及运动信息, 进行更高质量的视频创作.

(3) 生成模型的泛化性能

目前大多数方法在训练生成模型时使用的伪草图与真实图像仍十分接近, 当输入真实的手绘草图时生成效果显著下降, 生成模型的泛化能力较差. Yang 等人[53]获取不同绘制粒度的草图共同训练生成模型, 提出可控的人脸草图生成图像任务. SketchyGAN[19]通过调整训练策略, 在训练开始时输入伪草图-图像匹配对, 并随着训练进行逐渐增加真实草图-图像匹配对的比重实现草图和伪草图的统一训练. 在未来工作中, 提高模型的泛化能力是生成模型的改进方向之一, 其中增加输入草图的多样性让输入草图尽可能接近真实手绘草图的多种风格, 在生成模型中学习纠正形变的草图并补充草图缺失的细节信息, 或者是修改训练模型策略让模型更适应于手绘草图的输入都是可行的改进方向. 此外, 生成模型在提高输入手绘草图时生成结果的质量和真实性的同时, 还需要提高从抽象草图中提取关键语义内容的能力, 保持生成结果与输入草图的语义一致性, 得到符合用户预期的生成结果.

5.2 研究趋势

(1) 场景草图在视觉内容生成领域的应用

与含有单个物体实例的草图相比, 场景草图[46]通常包含多个前景物体实例以及背景信息, 更接近实际生活中出现的图像或视频. 目前的工作仅局限于场景草图生成对应的图像, 在未来的研究可以将场景草图与风格转化、视觉内容编辑任务相结合, 进行风格控制的场景草图生成图像, 并且在生成时考虑每个实例和场景整体的风格化效果, 或者利用场景草图对多种场景类别的图像或者视频进行编辑. 除此之外, 未来工作还应考虑到场景草图中多个实例间存在的互相覆盖和遮挡的问题, 提高草图实例分割的结果辅助场景草图进行视觉内容生成, 以及结合场景和不同实例的语义内容有效地提取场景级和实例级的特征, 对场景草图和图像/视频之间进行特征对齐, 实现包含复杂语义的场景草图生成高质量的图像或视频.

(2) 结合多模态的细粒度视觉内容生成

由于草图的稀疏表示无法完全包含图像和视频中稠密的视觉信息, 未来工作可以将草图与其他模态相结合进行信息互补, 全面地刻画用户意图. 目前已有方法输入彩色笔画、纹理块和示例图像控制草图生成图像的风格, 后续的研究还可以将草图与多种模态如文本、语音、图像、视频等相结合, 在图像生成中提供外观风格的控制, 在视频生成中提供外观风格及时序信息的控制, 按照用户期望进行视觉内容生成. 进一步地, 未来研究可以利用多种模态输入的特征与视觉对象从语义、实例、局部区域等多个层面建立对应关系, 实现用户从局部细节到整体特征的控制生成, 进行可控的细粒度图像或者视频生成.

(3) 小样本或零样本草图下的视觉内容生成

草图的获取难度使得目前数据集的规模较难满足模型训练的需求, 因此小样本或零样本的草图-图像匹配对和草图-视频匹配对下的视觉内容生成具有重要的研究意义. 文献 [69,70] 利用微调模型参数实现少量的草图样本生成图像, Zuiderveld[85]提出特征分离的假设将图像特征分解为风格特征和内容特征, 实现零样本的草图生成图像. 但是生成模型的泛化能力和生成结果的质量仍然需进一步提高. 后续的工作可以通过迁移学习或者特征增强的方法, 让模型从少量的草图样本或者已有的草图样本中提取到有效的特征来提高模型的鲁棒性, 这是小样本或零样本草图下视觉内容生成可行的改进方向.

(4) 扩散模型下基于草图的视觉内容生成

不同于 GAN 网络通过生成器和判别器进行对抗学习, 扩散模型 (diffusion models) 基于马尔可夫链将噪声逐步添加至数据中, 并通过学习逆向过程从噪声中修复数据实现视觉内容生成. 目前扩散模型在图像生成的应用效果已经逐渐超过 GAN 网络, 在未来视觉内容生成中具有重要应用场景. 草图可以作为条件输入控制扩散模型的生成过程, 决定生成物体的形状结构和空间布局. DiffSketching[141]将扩散模型应用至多物体实例草图生成图像中, DiSS[142]由草图和彩色笔画控制扩散模型生成图像的结构、颜色和真实度, 它们生成的图像质量得到大幅度提高. ControlNet[143]和 T2I-Adapter[144]通过将多种结构信息如语义图、草图、空间布局等输入至扩散模型中, 控制文本生成图像的结构和布局. Voynov 等人[145]和 MaungMaung 等人[146]结合草图固有的属性, 利用草图控制文本生成图像中的物体结构, 并生成高质量的图像. Kim 等人[147]基于草图和示例图像编辑图像的局部结构和外观风格. DiffFaceSketch[148]基于扩散模型输入草图生成高质量的人脸图像. 在未来工作中, 扩散模型可以将草图作为条件输入, 提高多种视觉内容生成任务的生成质量, 有效提升输入手绘草图时的模型泛化能力, 让用户在实际输入手绘草图时仍实现高质量的图像和视频生成.

(5) 复杂三维场景下基于草图的视觉内容生成

本文主要描写草图在二维视觉内容 (图像和视频) 生成中的应用, 但是随着智能触屏设备的普及和 VR/AR 技术的发展, 草图可以从二维层面描述物体的几何拓扑结构并通过深度学习方法或者图形学方法建模生成对应的实体, 被逐渐应用至三维实体的生成中. 目前的主要研究方向包括输入绘制抽象的手绘草图[149−151]或者专业绘制的草图[152,153]通过对草图直接编码生成多个类别的三维实体. 此外, SketchSampler[154]将草图与三维点云相结合生成对应的三维物体, SKED[155]则利用草图编辑文本生成的三维图像, 通过输入两个视角的草图经由 NeRF[156]渲染和编辑物体的形状.

此外草图可以与专业领域的知识相结合进行三维设计, 如发型设计、人脸设计、人体姿态设计、服装设计、建筑设计和计算机辅助设计 (CAD). CaricatureShop[157]利用草图修改人脸面部曲线, 对人脸进行夸张表示, 生成个性化的真实三维人脸. SEMM[158]通过草图笔画线条控制人脸的皱纹改变三维人脸的表情和年龄. SketchFace-NeRF[159]基于 NeRF 实现二维草图生成和编辑三维人脸, 它利用二维草图和参考图像作为输入, 预测包含色彩和立体信息的草图三维平面特征, 投影至三维隐空间并利用 NeRF 实现高质量三维人脸生成. SMPL[160]利用人体草图生成三维人体骨架, 并进一步结合生成对应姿势的三维人体. Sketch2Pose[161]从单张的位图草图分别推断二维和三维的人体骨架并结合生成符合用户期望的三维人体. DeepSketchHair[162]利用草图生成三维发型进行发型设计. 专

家和业余时装设计师通过绘制二维草图生成对应的三维服装进行服装设计, Foldsketch[163]根据平面草图绘制衣服的折叠和褶皱图案生成对应三维服装的褶皱变化, 降低现有的繁琐、耗时的迭代流程, Wang 等人[164]利用二维草图设计服装并显示生成的三维服装在不同身形上的效果, Knit sketching[165]基于针织服装制作中的裁剪和缝制过程, 通过草图表示服装的二维平面和缝制顺序, 实现在二维平面上对三维针织服装的设计. Sketch2PQ[166]提出基于草图的三维建模系统让用户通过草图自由设计以平面四边形 (PQ) 网格表示的建筑表面屋顶形状. 工业草图[167]可以描述多个几何图元和它们之间的拓扑关系, 被用于工业产品、家居等的 CAD 设计中. 草图简化了三维建模中复杂的参数输入, 支持没有专业领域知识的用户建模生成逼真的三维实体, 在三维视觉内容生成中具有很好的应用前景, 但目前局限于单个或者特定样式的三维实体生成. 未来工作可以由草图描述复杂三维场景, 生成多个三维实体并构建完整的场景, 进一步应用至建筑设计、机械制造、室内设计以及三维动画或电影渲染等更广泛的领域.

总之, 未来基于草图的视觉内容生成工作应该重点面向解决草图本身因为抽象性和风格多样性带来的挑战, 提高输入手绘草图时模型整体的生成能力, 以及结合草图具有语义表征的特性和生成领域的新方向, 扩大草图在视觉内容生成中的应用范围, 推动视觉内容生成的发展.

References:

[1] Chen T, Cheng MM, Tan P, Shamir A, Hu SM. Sketch2Photo: Internet image montage. ACM Trans. on Graphics, 2009, 28(5): 1–10. [doi: 10.1145/1618452.1618470]

[2] Eitz M, Richter R, Hildebrand K, Boubekeur T, Alexa M. PhotoSketcher: Interactive sketch-based image synthesis. IEEE Computer Graphics and Applications, 2011, 31(6): 56–66. [doi: 10.1109/MCG.2011.67]

[3] Xie SN, Tu ZW. Holistically-nested edge detection. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision. Santiago: IEEE, 2015. 1395–1403. [doi: 10.1109/ICCV.2015.164]

[4] Winnemöller H, Kyprianidis JE, Olsen SC. XDoG: An extended difference-of-Gaussians compendium including advanced image stylization. Computers & Graphics, 2012, 36(6): 740–753. [doi: 10.1016/j.cag.2012.03.004]

[5] Kang H, Lee S, Chui CK. Coherent line drawing. In: Proc. of the 5th Int'l Symp. on Non-photorealistic Animation and Rendering. San Diego: ACM, 2007. 43–50. [doi: 10.1145/1274871.1274878]

[6] Lu CW, Xu L, Jia JY. Combining sketch and tone for pencil drawing production. In: Proc. of the 2012 Symp. on Non-photorealistic Animation and Rendering. Annecy: Eurographics Association, 2012. 65–73.

[7] Su QK, Bai X, Fu HB, Tai CL, Wang J. Live sketch: Video-driven dynamic deformation of static drawings. In: Proc. of the 2018 CHI Conf. on Human Factors in Computing Systems. Montreal: ACM, 2018. 662. [doi: 10.1145/3173574.3174236]

[8] Dvorožnák M, Li W, Kim VG, Sýkora D. Toonsynth: Example-based synthesis of hand-colored cartoon animations. ACM Trans. on Graphics, 2018, 37(4): 167. [doi: 10.1145/3197517.3201326]

[9] Kingma DP, Welling M. Auto-encoding variational Bayes. In: Proc. of the 2nd Int'l Conf. on Learning Representations. Banff: ICLR, 2013.

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

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

[12] Xu P, Hospedales TM, Yin QY, Song YZ, Xiang T, Wang L. Deep learning for free-hand sketch: A survey. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2023, 45(1): 285–312. [doi: 10.1109/TPAMI.2022.3148853]

[13] Elasri M, Elharrouss O, Al-Maadeed S, Tairi H. Image generation: A review. Neural Processing Letters, 2022, 54(5): 4609–4646. [doi: 10.1007/s11063-022-10777-x]

[14] Zhan FN, Yu YC, Wu RL, Zhang JH, Lu SJ, Liu LJ, Kortylewski A, Theobalt C, Xing E. Multimodal image synthesis and editing: A survey. arXiv:2112.13592v3, 2021.

[15] Chen SY, Zhang JQ, Zhao YY, Rosin PL, Lai YK, Gao L. A review of image and video colorization: From analogies to deep learning. Visual Informatics, 2022, 6(3): 51–68. [doi: 10.1016/j.visinf.2022.05.003]

[16] Wang JX, Shi YJ, Liu H, Huang HQ, Du F. Research on freehand sketch to image translation based on generative adversarial networks. Application Research of Computers, 2022, 39(8): 2249–2256 (in Chinese with English abstract). [doi: 10.19734/j.issn.1001-3695.2022.01.0027]

[17] Li MT, Lin Z, Mech R, Yumer E, Ramanan D. Photo-sketching: Inferring contour drawings from images. In: Proc. of the 2019 IEEE

Winter Conf. on Applications of Computer Vision. Waikoloa: IEEE, 2019. 1403–1412. [doi: 10.1109/WACV.2019.00154]

[18] Song JF, Pang KY, Song YZ, Xiang T, Hospedales TM. Learning to sketch with shortcut cycle consistency. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 801–810. [doi: 10.1109/CVPR.2018.00090]

[19] Chen W, Hays J. SketchyGAN: Towards diverse and realistic sketch to image synthesis. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 9416–9425. [doi: 10.1109/CVPR.2018.00981]

[20] Chen SY, Su WC, Gao L, Xia SH, Fu HB. DeepFaceDrawing: Deep generation of face images from sketches. ACM Trans. on Graphics, 2020, 39(4): 72. [doi: 10.1145/3386569.3392386]

[21] Qiu HN, Wang C, Zhu H, Zhu XY, Gu JJ, Han XG. Two-phase hair image synthesis by self-enhancing generative model. Computer Graphics Forum, 2019, 38(7): 403–412. [doi: 10.1111/cgf.13847]

[22] Wu X, Wang C, Fu HB, Shamir A, Zhang SH. DeepPortraitDrawing: Generating human body images from freehand sketches. Computers & Graphics, 2023, 116: 73–81. [doi: 10.1016/j.cag.2023.08.005]

[23] Ham C, Tarrés GC, Bui T, Hays J, Lin Z, Collomosse J. CoGS: Controllable generation and search from sketch and style. In: Proc. of the 2022 European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 632–650. [doi: 10.1007/978-3-031-19787-1_36]

[24] Chen SY, Liu FL, Lai YK, Rosin PL, Li CP, Gao L. DeepFaceEditing: Deep face generation and editing with disentangled geometry and appearance control. ACM Trans. on Graphics, 2021, 40(4): 90. [doi: 10.1145/3450626.3459760]

[25] Zeng Y, Lin Z, Patel VM. SketchEdit: Mask-free local image manipulation with partial sketches. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 5941–5951. [doi: 10.1109/CVPR52688.2022.00586]

[26] Wang YX, Wei YC, Qian XM, Zhu L, Yang Y. Sketch-guided scenery image outpainting. IEEE Trans. on Image Processing, 2021, 30: 2643–2655. [doi: 10.1109/TIP.2021.3054477]

[27] Li XY, Zhang B, Liao J, Sander PV. Deep sketch-guided cartoon video inbetweening. IEEE Trans. on Visualization and Computer Graphics, 2022, 28(8): 2938–2952. [doi: 10.1109/TVCG.2021.3049419]

[28] Liu FL, Chen SY, Lai YK, Li CP, Jiang YR, Fu HB, Gao L. DeepFaceVideoEditing: Sketch-based deep editing of face videos. ACM Trans. on Graphics, 2022, 41(4): 167. [doi: 10.1145/3528223.3530056]

[29] Shi M, Zhang JQ, Chen SY, Gao L, Lai YK, Zhang FL. Reference-based deep line art video colorization. IEEE Trans. on Visualization and Computer Graphics, 2023, 29(6): 2965–2979. [doi: 10.1109/TVCG.2022.3146000]

[30] Kampelmuhler M, Pinz A. Synthesizing human-like sketches from natural images using a conditional convolutional decoder. In: Proc. of the 2020 IEEE Winter Conf. on Applications of Computer Vision. Snowmass: IEEE, 2020. 3192–3200. [doi: 10.1109/WACV45572.2020.9093440]

[31] Ashtari A, Seo CW, Kang C, Cha SH, Noh J. Reference based sketch extraction via attention mechanism. ACM Trans. on Graphics, 2022, 41(6): 207. [doi: 10.1145/3550454.3555504]

[32] Zhang Y, Su GY, Qi YG, Yang J. Unpaired image-to-sketch translation network for sketch synthesis. In: Proc. of the 2019 IEEE Visual Communications and Image Processing. Sydney: IEEE, 2019. 1–4. [doi: 10.1109/VCIP47243.2019.8965725]

[33] Vinker Y, Pajouheshgar E, Bo JY, Bachmann RC, Bermano AH, Cohen-Or D, Zamir AR, Shamir A. CLIPasso: Semantically-aware object sketching. ACM Trans. on Graphics, 2022, 41(4): 86. [doi: 10.1145/3528223.3530068]

[34] Zhu MR, Liang CC, Wang NN, Wang XY, Li ZF, Gao XB. A Sketch-Transformer network for face photo-sketch synthesis. In: Proc. of the 30th Int'l Joint Conf. on Artificial Intelligence. Montreal: IJCAI.org, 2021. 1352–1358. [doi: 10.24963/ijcai.2021/187]

[35] Yu J, Xu XX, Gao F, Shi SJ, Wang M, Tao DC, Huang QM. Toward realistic face photo-sketch synthesis via composition-aided GANs. IEEE Trans. on Cybernetics, 2021, 51(9): 4350–4362. [doi: 10.1109/TCYB.2020.2972944]

[36] Qi XQ, Sun MY, Wang WN, Dong XX, Li Q, Shan CF. Face sketch synthesis via semantic-driven generative adversarial network. In: Proc. of the 2021 IEEE Int'l Joint Conf. on Biometrics (IJCB). Shenzhen: IEEE, 2021. 1–8. [doi: 10.1109/IJCB52358.2021.9484393]

[37] Zhang CY, Liu DC, Peng CL, Wang NN, Gao XB. Edge aware domain transformation for face sketch synthesis. IEEE Trans. on Information Forensics and Security, 2022, 17: 2761–2770. [doi: 10.1109/TIFS.2022.3195383]

[38] Park T, Liu MY, Wang TC, Zhu JY. Semantic image synthesis with spatially-adaptive normalization. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 2337–2346. [doi: 10.1109/CVPR.2019.00244]

[39] Isola P, Zhu JY, Zhou TH, Efros AA. Image-to-image translation with conditional adversarial networks. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 5967–5976. [doi: 10.1109/CVPR.2017.632]

[40] Lu YY, Wu SZ, Tai YW, Tang CK. Image generation from sketch constraint using contextual GAN. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 205–220. [doi: 10.1007/978-3-030-01270-0_13]

[41] Koley S, Bhunia AK, Sain A, Chowdhury PN, Xiang T, Song YZ. Picture that sketch: Photorealistic image generation from abstract aketches. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 6850–6861.

[doi: 10.1109/CVPR52729.2023.00662]

[42] Ghosh A, Zhang R, Dokania PK, Wang O, Efros A, Torr P, Shechtman E. Interactive sketch & fill: Multiclass sketch-to-image translation. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 1171–1180. [doi: 10.1109/ICCV.2019.00126]

[43] Li ZY, Deng C, Yang EK, Tao DC. Staged sketch-to-image synthesis via semi-supervised generative adversarial networks. IEEE Trans. on Multimedia, 2021, 23: 2694–2705. [doi: 10.1109/TMM.2020.3015015]

[44] Zong YJ. A two-stage method and application implementation for image generation from sketch [MS. Thesis]. Dalian: Dalian University of Technology, 2021 (in Chinese with English abstract). [doi: 10.26991/d.cnki.gdllu.2021.003556]

[45] Cai YT, Chen ZJ, Ye DY. Bi-level cascading GAN-based heterogeneous conversion of sketch-to-realistic images. Pattern Recognition and Artificial Intelligence, 2018, 31(10): 877–886 (in Chinese with English abstract). [doi: 10.16451/j.cnki.issn1003-6059.201810002]

[46] Gao CY, Liu Q, Xu Q, Wang LM, Liu JZ, Zou CQ. SketchyCOCO: Image generation from freehand scene sketches. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 5174–5183. [doi: 10.1109/CVPR42600.2020.00522]

[47] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Proc. of the 18th Int'l Conf. on Medical Image Computing and Computer-assisted Intervention. Munich: Springer, 2015. 234–241. [doi: 10.1007/978-3-319-24574-4_28]

[48] Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-ResNet and the impact of residual connections on learning. In: Proc. of the 31st AAAI Conf. on Artificial Intelligence. San Francisco: AAAI Press, 2017. 4278–4284.

[49] 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]

[50] Huang X, Liu MY, Belongie S, Kautz J. Multimodal unsupervised image-to-image translation. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 172–189. [doi: 10.1007/978-3-030-01219-9_11]

[51] Li YH, Chen XJ, Yang BX, Chen ZH, Cheng ZH, Zha ZJ. DeepFacePencil: Creating face images from freehand sketches. In: Proc. of the 28th ACM Int'l Conf. on Multimedia. Seattle: ACM, 2020. 991–999. [doi: 10.1145/3394171.3413684]

[52] Xia WH, Yang YJ, Xue JH. Cali-sketch: Stroke calibration and completion for high-quality face image generation from human-like sketches. Neurocomputing, 2021, 460: 256–265. [doi: 10.1016/j.neucom.2021.07.029]

[53] Yang S, Wang ZY, Liu JY, Guo ZM. Controllable sketch-to-image translation for robust face synthesis. IEEE Trans. on Image Processing, 2021, 30: 8797–8810. [doi: 10.1109/TIP.2021.3120669]

[54] Li YH, Chen XJ, Wu F, Zha ZJ. LinesToFacePhoto: Face photo generation from lines with conditional self-attention generative adversarial networks. In: Proc. of the 27th ACM Int'l Conf. on Multimedia. Nice: ACM, 2019. 2323–2331. [doi: 10.1145/3343031.3350854]

[55] Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science, 2000, 290(5500): 2323–2326. [doi: 10.1126/science.290.5500.2323]

[56] Yang Y, Hossain Z, Gedeon T, Rahman S. S2FGAN: Semantically aware interactive sketch-to-face translation. In: Proc. of the 2022 IEEE/CVF Winter Conf. on Applications of Computer Vision. Waikoloa: IEEE, 2022.3162–3171. [doi: 10.1109/WACV51458.2022.00322]

[57] Richardson E, Alaluf Y, Patashnik O, Nitzan Y, Azar Y, Shapiro S, Cohen-Or D. Encoding in style: A StyleGAN encoder for image-to-image translation. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 2287–2296. [doi: 10.1109/CVPR46437.2021.00232]

[58] Olszewski K, Ceylan D, Xing J, Echevarria J, Chen ZL, Chen WK, Li H. Intuitive, interactive beard and hair synthesis with generative models. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 7444–7454. [doi: 10.1109/CVPR42600.2020.00747]

[59] Xiao CF, Yu D, Han XG, Zheng YY, Fu HB. SketchHairSalon: Deep sketch-based hair image synthesis. ACM Trans. on Graphics, 2021, 40(6): 216. [doi: 10.1145/3478513.3480502]

[60] Ho TT, Virtusio JJ, Chen YY, Hsu CM, Hua KL. Sketch-guided deep portrait generation. ACM Trans. on Multimedia Computing, Communications, and Applications, 2020, 16(3): 88. [doi: 10.1145/3396237]

[61] Chen H, Zhu SC. A generative sketch model for human hair analysis and synthesis. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2006, 28(7): 1025–1040. [doi: 10.1109/TPAMI.2006.131]

[62] Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 2242–2251. [doi: 10.1109/ICCV.2017.244]

[63] Xiang XY, Liu D, Yang X, Zhu YH, Shen XH, Allebach JP. Adversarial open domain adaptation for sketch-to-photo synthesis. In: Proc. of the 2022 IEEE/CVF Winter Conf. on Applications of Computer Vision. Waikoloa: IEEE, 2022. 944–954. [doi: 10.1109/WACV51458.2022.00102]

[64] Liu RT, Yu Q, Yu SX. Unsupervised sketch to photo synthesis. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 36–52. [doi: 10.1007/978-3-030-58580-8_3]

[65] Kazemi H, Taherkhani F, Nasrabadi NM. Unsupervised facial geometry learning for sketch to photo synthesis. In: Proc. of the 2018 Int'l Conf. of the Biometrics Special Interest Group. Darmstadt: IEEE, 2018. 1–5. [doi: 10.23919/BIOSIG.2018.8552937]

[66] Bashkirova D, Lezama J, Sohn K, Saenko K, Essa I. MaskSketch: Unpaired structure-guided masked image generation. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 1879–1889.

[67] Chang HW, Zhang H, Jiang L, Liu C, Freeman WT. MaskGIT: Masked generative image transformer. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 11305–11315. [doi: 10.1109/CVPR52688.2022.01103]

[68] Esser P, Rombach R, Ommer B. Taming transformers for high-resolution image synthesis. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 12868–12878. [doi: 10.1109/CVPR46437.2021.01268]

[69] Wang SY, Bau D, Zhu JY. Sketch your own GAN. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 14030–14040. [doi: 10.1109/ICCV48922.2021.01379]

[70] Israr SM, Zhao F. Customizing GAN using few-shot sketches. In: Proc. of the 30th ACM Int'l Conf. on Multimedia. Lisboa: ACM, 2022. 2229–2238. [doi: 10.1145/3503161.3548415]

[71] Yang BX, Chen XJ, Wang CQ, Zhang C, Chen ZH, Sun XY. Semantics-preserving sketch embedding for face generation. IEEE Trans. on Multimedia, 2022. 1–15. [doi: 10.1109/TMM.2023.3239182]

[72] Liu BC, Song KP, Zhu YZ, Elgammal A. Sketch-to-art: Synthesizing stylized art images from sketches. In: Proc. of the 15th Asian Conf. on Computer Vision. Kyoto: Springer, 2020. 207–222. [doi: 10.1007/978-3-030-69544-6_13]

[73] Zhang LM, Ji Y, Lin X, Liu CP. Style transfer for anime sketches with enhanced residual U-Net and auxiliary classifier GAN. In: Proc. of the 4th IAPR Asian Conf. on Pattern Recognition (ACPR). Nanjing: IEEE, 2017. 506–511. [doi: 10.1109/ACPR.2017.61]

[74] Zhang LM, Li CZ, Simo-Serra E, Ji Y, Wong TT, Liu CP. User-guided line art flat filling with split filling mechanism. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 9889–9898. [doi: 10.1109/CVPR46437.2021.00976]

[75] Sangkloy P, Lu JW, Fang C, Yu F, Hays J. Scribbler: Controlling deep image synthesis with sketch and color. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 6836–6845. [doi: 10.1109/CVPR.2017.723]

[76] Xian W, Sangkloy P, Agrawal V, Raj A, Lu JW, Fang C, Yu F, Hays J. TextureGAN: Controlling deep image synthesis with texture patches. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 8456–8465. [doi: 10.1109/CVPR.2018.00882]

[77] Li JN, Liu SQ, Cao MY. Line artist: A multiple style sketch to painting synthesis scheme. arXiv:1803.06647, 2018.

[78] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego: ICLR, 2015.

[79] Page L, Brin S, Motwani R, Winograd T. The PageRank citation ranking: Bringing order to the Web. Technical Report, Stanford InfoLab, 1999.

[80] Huang JL, Jing L, Tan ZF, Kwong S. Multi-density sketch-to-image translation network. IEEE Trans. on Multimedia, 2021, 24: 4002–4015. [doi: 10.1109/TMM.2021.3111501]

[81] Tan ZT, Chai ML, Chen DD, Liao J, Chu Q, Yuan L, Tulyakov S, Yu N. MichiGAN: Multi-input-conditioned hair image generation for portrait editing. ACM Trans. on Graphics, 2020, 39(4): 95. [doi: 10.1145/3386569.3392488]

[82] Lee J, Kim E, Lee Y, Kim D, Chang J, Choo J. Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 5800–5809. [doi: 10.1109/CVPR42600.2020.00584]

[83] Zhang P, Zhang B, Chen D, Yuan L, Wen F. Cross-domain correspondence learning for exemplar-based image translation. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 5143–5153. [doi: 10.1109/CVPR42600.2020.00519]

[84] Zhou XR, Zhang B, Zhang T, Zhang P, Bao JM, Chen D, Zhang ZF, Wen F. CoCosNet v2: Full-resolution correspondence learning for image translation. In: Proc. of the 2021 IEEE Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 11465–11475.

[85] Zuiderveld J. Style-content disentanglement in language-image pretraining representations for zero-shot sketch-to-image synthesis. arXiv:2206.01661v1, 2022.

[86] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I. Learning transferable visual models from natural language supervision. In: Proc. of the 38th Int'l Conf. on Machine Learning. PMLR, 2021. 8748–8763.

[87] Deng J, Dong W, Socher R, Li LJ, Li K, Li FF. ImageNet: A large-scale hierarchical image database. In: Proc. of the 2009 IEEE Conf. on Computer Vision and Pattern Recognition. Miami: IEEE, 2009. 248–255. [doi: 10.1109/CVPR.2009.5206848]

[88] Liu BC, Zhu YZ, Song KP, Elgammal A. Self-supervised sketch-to-image synthesis. In: Proc. of the 37th AAAI Conf. on Artificial Intelligence. Washington: AAAI Press, 2021. 2073–2081. [doi: 10.1609/aaai.v35i3.16304]

[89] Odena A, Olah C, Shlens J. Conditional image synthesis with auxiliary classifier GANs. In: Proc. of the 34th Int'l Conf. on Machine Learning. Sydney: PMLR, 2017. 2642–2651.

[90] Zhang LM, Li CZ, Wong TT, Li Y, Liu CP. Two-stage sketch colorization. ACM Trans. on Graphics, 2018, 37(6): 261. [doi: 10.1145/3272127.3275090]

[91] Kim H, Jhoo HY, Park E, Yoo S. Tag2Pix: Line art colorization using text tag with SECat and changing loss. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 9055–9064. [doi: 10.1109/ICCV.2019.00915]

[92] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141. [doi: 10.1109/CVPR.2018.00745]

[93] Jo Y, Park J. SC-FEGAN: Face editing generative adversarial network with user's sketch and color. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 1745–1753. [doi: 10.1109/ICCV.2019.00183]

[94] Portenier T, Hu QY, Szabo A, Bigdeli SA, Favaro P, Zwicker M. FaceShop: Deep sketch-based face image editing. ACM Trans. on Graphics, 2018, 37(4): 99. [doi: 10.1145/3197517.3201393]

[95] Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC. Improved training of Wasserstein GANs. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 5769–5779.

[96] Yu JH, Lin Z, Yang JM, Shen XH, Lu X, Huang T. Free-form image inpainting with gated convolution. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 4470–4479. [doi: 10.1109/ICCV.2019.00457]

[97] Yang S, Wang ZY, Liu JY, Guo ZM. Deep plastic surgery: Robust and controllable image editing with human-drawn sketches. In: Proc. of the16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 601–617. [doi: 10.1007/978-3-030-58555-6_36]

[98] Liu HY, Wan ZY, Huang W, Song YB, Han XT, Liao J, Jiang B, Liu W. DeFLOCNet: Deep image editing via flexible low-level controls. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 10760–10769. [doi: 10.1109/CVPR46437.2021.01062]

[99] Graves A. Long short-term memory. In: Graves A, ed. Supervised Sequence Labelling with Recurrent Neural Networks. Berlin: Springer, 2012. 37–45. [doi: 10.1007/978-3-642-24797-2]

[100] Zhang HC, Yu G, Chen T, Luo GZ. Sketch me a video. arXiv:2110.04710v1, 2021.

[101] Loftsdottir D, Guzdial M. SketchBetween: Video-to-video synthesis for sprite animation via sketches. In: Proc. of the 17th Int'l Conf. on the Foundations of Digital Games. Athens: ACM, 2022. 32. [doi: 10.1145/3555858.3555928]

[102] Thasarathan H, Nazeri K, Ebrahimi M. Automatic temporally coherent video colorization. In: Proc. of the 16th Conf. on Computer and Robot Vision (CRV). Kingston: IEEE, 2019. 189–194. [doi: 10.1109/CRV.2019.00033]

[103] Wang TC, Liu MY, Zhu JY, Tao A, Kauta J, Catanzaro B. High-resolution image synthesis and semantic manipulation with conditional GANs. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 8798–8807. [doi: 10.1109/CVPR.2018.00917]

[104] Xue TF, Wu JJ, Bouman KL, Freeman WT. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In: Proc. of the 30th Int'l Conf. on Neural Information Processing Systems. Barcelona: Curran Associates Inc., 2016. 91–99.

[105] van den Oord A, Vinyals O, Kavukcuoglu K. Neural discrete representation learning. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6309–6318.

[106] Yu Q, Liu F, Song YZ, Xiang T, Hospedales TM, Loy CC. Sketch me that shoe. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 799–807. [doi: 10.1109/CVPR.2016.93]

[107] Sangkloy P, Burnell N, Ham C, Hays J. The sketchy database: Learning to retrieve badly drawn bunnies. ACM Trans. on Graphics, 2016, 35(4): 119. [doi: 10.1145/2897824.2925954]

[108] Tang XG, Wang XO. Face photo-sketch synthesis and recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2009, 31(11): 1955–1967. [doi: 10.1109/TPAMI.2008.222]

[109] Zhang W, Wang WG, Tang XO. Coupled information-theoretic encoding for face photo-sketch recognition. In: Proc. of the 2011 Conf. on Computer Vision and Pattern Recognition (CVPR). Colorado Springs: IEEE, 2011. 513–520. [doi: 10.1109/CVPR.2011.5995324]

[110] Liu ZW, Luo P, Wang XG, Tang XO. Deep learning face attributes in the wild. In: Proc of the 2015 IEEE Int'l Conf. on Computer Vision. Santiago: IEEE, 2015. 3730–3738. [doi: 10.1109/ICCV.2015.425]

[111] Wah C, Branson S, Welinder P, Perona P, Belongie S. The Caltech-UCSD Birds-200-2011 dataset. California: California Institute of Technology, 2011. https://authors.library.caltech.edu/records/cvm3y-5hh21

[112] Krause J, Stark M, Deng J, Li FF. 3D object representations for fine-grained categorization. In: Proc. of the 2013 IEEE Int'l Conf. on Computer Vision Workshops. Sydney: IEEE, 2013. 554–561. [doi: 10.1109/ICCVW.2013.77]

[113] Pirrone R, Cannella V, Gambino O, Pipitone A, Russo G. WikiArt: An ontology-based information retrieval system for arts. In: Proc. of the 9th Int'l Conf. on Intelligent Systems Design and Applications. Pisa: IEEE, 2009. 913–918. [doi: 10.1109/ISDA.2009.219]

[114] Karras T, Aila T, Laine S, Lehtinen J. Progressive growing of GANs for improved quality, stability, and variation. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: OpenReview.net, 2018.

[115] Lee CH, Liu ZW, Wu LY, Luo P. Maskgan: Towards diverse and interactive facial image manipulation. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 5548–5557. [doi: 10.1109/CVPR42600.2020.00559]

[116] Branwen G, Gokaslan A. Danbooru2017: A large-scale crowdsourced and tagged anime illustration dataset. 2017. https://www.gwern.net/Danbooru2017

[117] Yang ZX, Dong J, Liu P, Yang Y, Yan SC. Very long natural scenery image prediction by outpainting. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 10560–10569. [doi: 10.1109/ICCV.2019.01066]

[118] Liu ZW, Luo P, Qiu S, Wang XG, Tang XO. DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 1096–1104. [doi: 10.1109/CVPR.2016.124]

[119] Zhou BL, Lapedriza A, Khosla A, Oliva A, Torralba A. Places: A 10 million image database for scene recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2018, 40(6): 1452–1464. [doi: 10.1109/TPAMI.2017.2723009]

[120] Chung JS, Nagrani A, Zisserman A. VoxCeleb2: Deep speaker recognition. In: Proc. of the 19th Annual Conf. of the Int'l Speech Communication Association (Interspeech 2018). Hyderabad: ISCA, 2018. 1086–1090. [doi: 10.21437/Interspeech.2018-1929]

[121] Siarohin A, Lathuilière S, Tulyakov S, Ricci E, Sebe N. Animating arbitrary objects via deep motion transfer. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 2372–2381. [doi: 10.1109/CVPR.2019.00248]

[122] Eitz M, Hays J, Alexa M. How do humans sketch objects? ACM Trans. on Graphics, 2012, 31(4): 1–10. [doi: 10.1145/2185520.2185540]

[123] Ha D, Eck D. A neural representation of sketch drawings. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: OpenReview.net, 2018.

[124] Caesar H, Uijlings J, Ferrari V. COCO-stuff: Thing and stuff classes in context. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 1209–1218. [doi: 10.1109/CVPR.2018.00132]

[125] Li YJ, Fang C, Hertzmann A, Shechtman E, Yang MH. Im2Pencil: Controllable pencil illustration from photographs. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 1525–1534. [doi: 10.1109/CVPR.2019.00162]

[126] Gatys LA, Ecker AS, Bethge M. Image style transfer using convolutional neural networks. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 2414–2423. [doi: 10.1109/CVPR.2016.265]

[127] Inoue N, Ito D, Xu N, Yang J, Price B, Yamasaki T. Learning to trace: Expressive line drawing generation from photographs. Computer Graphics Forum, 2019, 38(7): 69–80. [doi: 10.1111/cgf.13817]

[128] Recasens A, Kellnhofer P, Stent S, Matusik W, Torralba A. Learning to zoom: A saliency-based sampling layer for neural networks. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 52–67. [doi: 10.1007/978-3-030-01240-3_4]

[129] Weber M. AutoTrace. 2018. http://autotrace.sourceforge.net/

[130] Zhang TY, Suen CY. A fast parallel algorithm for thinning digital patterns. Communications of the ACM, 1984, 27(3): 236–239. [doi: 10.1145/357994.358023]

[131] Simo-Serra E, Iizuka S, Ishikawa H. Mastering sketching: Adversarial augmentation for structured prediction. ACM Trans. on Graphics, 2018, 37(1): 11. [doi: 10.1145/3132703]

[132] Simo-Serra E, Iizuka S, Sasaki K, Ishikawa H. Learning to simplify: Fully convolutional networks for rough sketch cleanup. ACM Trans. on Graphics, 2016, 35(4): 121. [doi: 10.1145/2897824.2925972]

[133] 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.

[134] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proc. of the 2016

IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 2818–2826. [doi: 10.1109/CVPR.2016.308]

[135] Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6629–6640.

[136] Bińkowski M, Sutherland DJ, Arbel M, Gretton A. Demystifying MMD GANs. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: OpenReview.net, 2018.

[137] Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 586–595. [doi: 10.1109/CVPR.2018.00068]

[138] Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: From error visibility to structural similarity. IEEE Trans. on Image Processing, 2004, 13(4): 600–612. [doi: 10.1109/TIP.2003.819861]

[139] Zhang L, Zhang L, Mou XQ, Zhang D. FSIM: A feature similarity index for image quality assessment. IEEE Trans. on Image Processing, 2011, 20(8): 2378–2386. [doi: 10.1109/TIP.2011.2109730]

[140] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 3431–3440. [doi: 10.1109/CVPR.2015.7298965]

[141] Wang Q, Kong D, Lin FY, Qi YG. DiffSketching: Sketch control image synthesis with diffusion models. In: Proc. of the 33rd British Machine Vision Conf. London: BMVA Press, 2022. 67

[142] Cheng SI, Chen YJ, Chiu WC, Tseng HY, Lee HY. Adaptively-realistic image generation from stroke and sketch with diffusion model. In: Proc. of the 2023 IEEE/CVF Winter Conf. on Applications of Computer Vision. 2023. 4043–4051. [doi: 10.1109/WACV56688.2023.00404]

[143] Zhang LM, Agrawala M. Adding conditional control to text-to-image diffusion models. arXiv:2302.05543, 2023.

[144] Mou C, Wang XT, Xie LB, Wu YZ, Zhang J, Qi ZG, Shan Y, Qie XH. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv:2302.08453, 2023.

[145] Voynov A, Aberman K, Cohen-Or D. Sketch-guided text-to-image diffusion models. In: Proc. of the 2023 ACM Special Interest Group on Computer Graphics and Interactive Techniques Conf. (SIGGRAPH 2023) Conf. Los Angeles: ACM, 2023. 1–11. [doi: 10.1145/3588432.3591560]

[146] MaungMaung A, Shing M, Mitsui K, Sawada K, Okura F. Text-guided scene sketch-to-photo synthesis. arXiv:2302.06883, 2023.

[147] Kim K, Park S, Lee J, Choo J. Reference-based image composition with sketch via structure-aware diffusion model. arXiv:2304.09748, 2023.

[148] Peng YC, Zhao CQ, Xie HR, Fukusato T, Miyata K. DiffFaceSketch: High-fidelity face image synthesis with sketch-guided latent diffusion model. arXiv:2302.06908, 2023.

[149] Huang HB, Kalogerakis E, Yumer E, Mech R. Shape synthesis from sketches via procedural models and convolutional networks. IEEE Trans. on Visualization and Computer Graphics, 2017, 23(8): 2003–2013. [doi: 10.1109/TVCG.2016.2597830]

[150] Wang LJ, Qian C, Wang JF, Fang Y. Unsupervised learning of 3D model reconstruction from hand-drawn sketches. In: Proc. of the 26th ACM Int'l Conf. on Multimedia. Seoul: ACM, 2018. 1820–1828. [doi: 10.1145/3240508.3240699]

[151] Guillard B, Remelli E, Yvernay P, Fua P. Sketch2Mesh: Reconstructing and editing 3D shapes from sketches. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 13003–13012. [doi: 10.1109/ICCV48922.2021.01278]

[152] Zhang SH, Guo YC, Gu QW. Sketch2Model: View-aware 3D modeling from single free-hand sketches. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 6000–6017. [doi: 10.1109/CVPR46437.2021.00595]

[153] Zhong Y, Qi YG, Gryaditskaya Y, Zhang HG, Song YZ. Towards practical sketch-based 3D shape generation: The role of professional sketches. IEEE Trans. on Circuits and Systems for Video Technology, 2021, 31(9): 3518–3528. [doi: 10.1109/TCSVT.2020.3040900]

[154] Gao CJ, Yu Q, Sheng L, Song YZ, XU D. SketchSampler: Sketch-based 3D reconstruction via view-dependent depth sampling. In: Proc. of the 17th European Conf. on Computer Vision. Cham: Springer, 2022. 464–479. [doi: 10.1007/978-3-031-19769-7_27]

[155] Mikaeili A, Perel O, Safaee M, Cohen-Or D, Mahdavi-Amiri A. SKED: Sketch-guided text-based 3D editing. arXiv:2303.10735, 2023.

[156] Mildenhall B, Srinivasan PP, Tancik M, Barron JT, Ramamoorthi R, Ng R. NeRF: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM, 2021, 65(1): 99–106. [doi: 10.1145/3503250]

[157] Han XG, Hou KC, Du D, Qiu YD, Cui SG, Zhou K, Yu YZ. CaricatureShop: Personalized and photorealistic caricature sketching. IEEE Trans. on Visualization and Computer Graphics, 2020, 26(7): 2349–2361. [doi: 10.1109/TVCG.2018.2886007]

[158] Ling JW, Wang ZB, Lu M, Wang Q, Qian C, Xu F. Structure-aware editable morphable model for 3D facial detail animation and manipulation. In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 249–267. [doi: 10.1007/978-3-031-

20062-5_15]

[159] Gao L, Liu FL, Chen SY, Jiang KW, Li CP, Lai YK, Fu HB. SketchFaceNeRF: Sketch-based facial generation and editing in neural radiance fields. ACM Trans. on Graphics, 2023, 42(4): 159. [doi: 10.1145/3592100]

[160] Yang KZ, Lu JT, Hu SY, Chen XJ. Deep 3D modeling of human bodies from freehand sketching. In: Proc. of the 27th Int'l Conf. on Multimedia Modeling. Prague: Springer, 2021. 36–48. [doi: 10.1007/978-3-030-67835-7_4]

[161] Brodt K, Bessmeltsev M. Sketch2Pose: Estimating a 3D character pose from a bitmap sketch. ACM Trans. on Graphics, 2022, 41(4): 85. [doi: 10.1145/3528223.3530106]

[162] Shen YF, Zhang CG, Fu HB, Zhou K, Zheng YY. DeepSketchHair: Deep sketch-based 3D hair modeling. IEEE Trans. on Visualization and Computer Graphics, 2021, 27(7): 3250–3263. [doi: 10.1109/TVCG.2020.2968433]

[163] Li MC, Sheffer A, Grinspun E, Vining N. Foldsketch: Enriching garments with physically reproducible folds. ACM Trans. on Graphics, 2018, 37(4): 133. [doi: 10.1145/3197517.3201310]

[164] Wang TY, Ceylan D, Popović J, Mitra NJ. Learning a shared shape space for multimodal garment design. ACM Trans. on Graphics, 2018, 37(6): 203. [doi: 10.1145/3272127.3275074]

[165] Kaspar A, Wu K, Luo YY, Makatura L, Matusik W. Knit sketching: From cut & sew patterns to machine-knit garments. ACM Trans. on Graphics, 2021, 40(4): 63. [doi: 10.1145/3450626.3459752]

[166] Deng Z, Liu Y, Pan H, Jabi W, Zhang JY, Deng BL. Sketch2PQ: Freeform planar quadrilateral mesh design via a single sketch. IEEE Trans. on Visualization and Computer Graphics, 2023, 29(9): 3826–3839. [doi: 10.1109/TVCG.2022.3170853]

[167] Willis KDD, Jayaraman PK, Lambourne JG, Chu H, Pu YW. Engineering sketch generation for computer-aided design. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 2105–2114. [doi: 10.1109/CVPRW53098.2021.00239]

附中文参考文献:

[16] 王建欣, 史英杰, 刘昊, 黄海峤, 杜方. 基于 GAN 的手绘草图图像翻译研究综述. 计算机应用研究, 2022, 39(8): 2249–2256. [doi: 10.19734/j.issn.1001-3695.2022.01.0027]

[44] 宗雨佳. 两阶段草图至图像生成模型与应用实现 [硕士学位论文]. 大连: 大连理工大学, 2021. [doi: 10.26991/d.cnki.gdllu.2021.003556]

[45] 蔡雨婷, 陈昭炯, 叶东毅. 基于双层级联 GAN 的草图到真实感图像的异质转换. 模式识别与人工智能, 2018, 31(10): 877–886. [doi: 10.16451/j.cnki.issn1003-6059.201810002]

左然(1995－), 女, 博士生, 主要研究领域为计算机视觉, 人机交互.

马翠霞(1975－), 女, 博士, 研究员, 博士生导师, CCF 杰出会员, 主要研究领域为人机交互, 媒体大数据可视分析.

胡皓翔(2000－), 男, 博士生, 主要研究领域为计算机视觉.

王宏安(1963－), 男, 博士, 研究员, 博士生导师, CCF 高级会员, 主要研究领域为自然人机交互, 实时智能计算.

邓小明(1980－), 男, 博士, 研究员, 博士生导师, CCF 高级会员, 主要研究领域为计算机视觉, 人机交互.