

基于残差空间网络的建筑图像编辑

朱成, 赵光哲, 林本旺, 王雪平*

(北京建筑大学电气与信息工程学院 北京 102616)
(wangxueping@bucea.edu.cn)

摘要: 建筑图像层次结构的复杂性给图像编辑带来了一定的挑战。虽然部分工作结合重建和编辑的方法进行建筑图像编辑, 但由于图像编码到的空间维度较低, 容易丢失空间信息, 导致编辑后的建筑图像发生扭曲, 难以保持结构一致性。为了解决上述问题, 提出一种基于 StyleGAN 的建筑图像编辑模型 ABEditor。首先提出多层次结构的残差空间网络, 用于学习多层空间特征和风格特征, 并连接到 StyleGAN 对应维度的高维特征空间; 通过逐层校正精确地重建输入图像, 以表达丰富的空间信息; 在合成网络更接近于输出的后期层, 使用编码器将图像映射到低维隐空间, 保证图像的颜色细节。为了更准确地评估编辑效果, 提出一种基于语义分割的平均像素编辑距离(APEDS)评价指标。在 LSUN Church 数据集上与相关方法进行实验的结果表明, ABEditor 在保留建筑图像细节和空间结构方面有所提升, 在 SSIM, FID, L_2 , LPIPS, PSNR 和 APEDS 定量分析指标中, 分别提升 28.64%, 20.74%, 13.67%, 1.67%, 10.01% 和 21.26%。

关键词: 建筑图像编辑; 生成对抗网络反演; 残差空间学习

中图法分类号: TP391.41

DOI: 10.3724/SP.J.1089.2023-00699

Architectural Image Editing Based on Residual Spatial Networks

Zhu Cheng, Zhao Guangzhe, Lin Benwang, and Wang Xueping*

(School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 102616)

Abstract: The complexity of the hierarchical structure of architectural images brings certain challenges to architectural image editing. Some works combine reconstruction and editing for architectural image editing, however, low-dimensional encoding in hidden space often results in loss of spatial information, causing generated images to be distorted and making it challenging to maintain structural consistency. To address these issues, we propose an architectural image editing model based on StyleGAN, named ABEditor. Specifically, we introduce a multi-level residual spatial network to learn multi-layer spatial and style features, which are connected to the corresponding dimensions in StyleGAN's high-dimensional feature space. The input image can be accurately reconstructed through layer-wise correction while preserving rich spatial information. An encoder in the early layers of the synthesis network maps images to low-dimensional latent space, ensuring fidelity in color details. Additionally, to more accurately evaluate editing performance, we propose the average pixel edit distance based on segmentation (APEDS) metric. Comparative evaluations on the LSUN Church dataset show that the ABEditor framework improves in retaining building image details and spatial structure, outperforming related methods by 28.64% in SSIM, 20.74% in FID, 13.67% in L_2 , 1.67% in LPIPS, 10.01% in PSNR, and 21.26% in APEDS.

收稿日期: 2023-10-27; 修回日期: 2024-11-28. 基金项目: 国家自然科学基金(62176018, 62402026); 北京建筑大学青年教师科研能力提升计划(X23026); 北京市教育委员会科研计划(KM202410016010). 朱成(1999—), 男, 硕士研究生, 主要研究方向为计算机视觉、图像编辑; 赵光哲(1979—), 男, 博士, 教授, 博士生导师, 主要研究方向为计算机视觉与图像处理、模式识别、人工智能; 林本旺(1996—), 男, 博士研究生, 主要研究方向为计算机视觉、人脸生成; 王雪平(1990—), 女, 博士, 论文通信作者, CCF 会员, 主要研究方向为模式识别与图像处理、情感计算。

Key words: architectural image editing; generative adversarial network inversion; residual spatial learning

随着深度学习的发展, 图像编辑技术已经成为计算机视觉领域的研究热点, 建筑图像编辑是其中一个重要研究对象, 它允许用户根据需求修改原始图像, 改变图像的特定属性, 增加或减少元素, 以更好地满足不同场景的需求. 建筑图像编辑可以改变建筑外观^[1]、修改室内设计^[2]和景观设计的图像, 为建筑师、设计师和客户提供了强大的工具, 用以可视化建筑项目的外观和布局. 随着生成对抗网络(generative adversarial network, GAN)^[3]的发展, 自动化建筑图像编辑已经成为建筑和城市规划领域的有力工具, 在提高建筑设计效率、改善城市规划等方面具有重要的意义.

当前, 在建筑图像编辑任务中仍面临一些挑战, 包括如何在结构复杂多样的情况下进行精准编辑, 保持建筑结构的完整性和合理性, 解决问题对于进一步拓展建筑图像编辑的应用领域和提高图像编辑的质量至关重要.

近年来, 建筑图像编辑方法经历了显著的演变, 从最初的风格迁移^[4]和图像翻译^[5]技术, 逐渐转向基于 GAN 隐空间的编辑方法. 风格迁移和图像翻译的方法虽然取得了一定的成果, 但是在使用单一模型时只能编辑有限的属性, 并且难以控制精细度, 通常只改变图像的颜色和整体风格. 随着 GAN 生成质量的提高, 使用预训练好的 GAN 模型可有效地解决以上问题. Karras 等^[6]提出的 StyleGAN 提升了生成性能, 可以生成带有隐风格向量的图像, 使得合成网络的不同层可以独立地控制生成图像的不同属性; 其中, 更靠近初始层的早期层主要控制图像的空间布局, 中期层构成较小的结构如窗户大小、数量, 而更靠近输出层的后期层则渲染整个场景的色彩以及更小的纹理细节. StyleGAN 具有一个解纠缠的隐空间 W , 提供了控制和编辑图像属性的能力. 文献[7-8]将输入域映射到 StyleGAN 的隐空间, 通过隐空间间接地控制图像属性. 这些方法遵循“先反转, 后编辑”的思想, 首先将图像反转到 StyleGAN 的隐空间中, 对反转得到的隐码通过语义方向上的编辑得到新的隐码; 然后将该隐码输入 StyleGAN, 输出编辑后的图像. 然而, 将真实的图像精确地映射到 W 空间仍然具有挑战性. 由于图像包含大量的像素和细节, 而需要映射的图像维度远大于 W 空间, 因此将图像编码到 W 空间的过程中会丢失大量空间信息, 在生成过程中的空间约束不能得到保证, 导

致编辑效果不理想. 对于线条细节多的建筑图像, 即便有线条或者窗户发生些许变化也会影响整体的建筑观感, 进而对于建筑图像编辑、重建质量的影响很大. 一般的方法在图像编辑任务中实用性较差, 很难给设计师提供参考, 在实际应用中受到了限制. 此外, StyleGAN 的生成器中, 不同的隐码经过各层卷积运算得到的多尺度特征, 共同构成一个更高维度的特征空间 F 空间, 虽然将图像逐层映射到 F 空间可以完成高精度的重建, 但这会产生大量的伪影附着在输出图像上, 使图像无法编辑.

为了解决以上问题, 本文提出一种自动建筑编辑器模型(auto building editor, ABEditor). 与传统 GAN 反演方法将图像编码到 W 空间不同, ABEditor 利用残差空间网络(residual spatial network, RSN)把图像映射到高维特征空间—— F 空间, 学习更丰富的多层次结构特征. 当输入不同的隐码时, 合成网络可以生成不同风格和内容的图像, 并在中间层捕捉到不同的风格特征, RSN 旨在学习生成的平均图像与输入图像之间风格特征的差异. RSN 通过多个下采样过程, 从图像中提取出一系列不同维度的风格特征, 并将这些特征通过残差连接的方式嵌入预训练的 StyleGAN 合成网络中. 因为高维特征空间的表达性比低维隐码的表达性更强, 这些空间特征的维度大小要远大于原图, 包含更丰富的图像信息; 所以在生成图像时, 在特征空间对齐会让输出的图像结构信息和纹理细节保留更为完整. 为了缓解映射到高维空间产生的伪影带来的编辑效果损失, 在合成网络的后期层, 改为使用 W 编码器将图像映射到低维隐空间 W 空间, 保证图像的颜色细节. 其次, 为了更好地度量图像编辑的效果和质量, 本文提出一种基于语义分割的平均像素编辑距离(average pixel edit distance based on segmentation, APEDS), 将图像分割为建筑物、天空等类别, 分别计算编辑距离, 有效地排除了重建质量的影响.

1 相关工作

1.1 建筑图像编辑

图像编辑指利用计算机算法代替人, 自动完成一些图像修改的方法, 即输入一幅图像, 经过算法计算后输出一幅编辑后的图像. 建筑图像是一

种图像类型,其中包括前景中的建筑物和建筑结构,背景中有天空和植被.建筑图像的特殊构图为图像编辑带来了巨大的挑战.目前,建筑图像编辑的方法分为基于风格迁移的方法、基于图像翻译的方法、基于 GAN 反演的方法和基于多模态扩散模型的方法 4 类.

自 Gatys 等^[9]提出风格迁移以来,出现了很多使用风格迁移编辑建筑图像的方法.甘益波等^[10]通过语义残差连接提取图像域内的语义特征,将场景图变成油画风格;Luan 等^[4]提出一种摄影风格迁移的方法,可以在室外场景和室内场景中产生逼真的风格迁移效果,包括时间、天气、季节和光线的迁移;Chen 等^[11]使用分割网络将建筑图像前景和背景分离开,可以将一幅建筑图像转换为黄昏、清晨、晚上或中午.但是,将另一幅图像的风格迁移到要编辑的图像时难以对单一属性进行精确控制,只能得到大面积的色彩风格转换.使用图像翻译技术的图像编辑方法也可以生成令人满意的结果,该技术旨在学习从源域到目标域的映射关系.梁浩然等^[12]提出的图像文本编辑方法,能够快速地在图像上进行字符编辑,并保持与相邻字符的几何和视觉一致性;对于建筑场景图像,Sangkloy 等^[5]将室内场景草图作为输入,并在用户指定的笔画指导下对其进行着色;Jiang 等^[1]提出一种两步法的图像翻译框架,可以将草图和建筑图像相互转换.但是,使用图像翻译的编辑方法往往只支持一种或几种属性的转换和编辑,如果需要其他属性的编辑,则需要重新训练一个模型.基于 GAN 反演的方法很好地解决了以上问题,其目的是将给定的图像反演回预训练的 GAN 模型生成器的隐空间,然后该生成器可以准确地从反向代码重建图像.GAN 反演使得在现有训练好的 GAN 的隐空间中找到的可控方向适用于编辑真实图像,而不需要专门的成对监督数据来单独训练一个网络.通常,GAN 反演分为 3 种方法:基于优化的方法,基于编码器的方法,以及结合两者的混合方法.基于优化的方法^[13]将每一幅图像单独映射到 GAN 的隐空间,该方法对于每幅图像都需要进行多步的反向传播,可以得到更高质量的重建图像,但需要花费大量时间;基于编码器的方法^[7,14-15]通过额外训练一个编码器,学习真实数据集到 GAN 隐空间的映射,虽然可以取得更快的速度,但重建质量通常比基于优化的方法差;结合两者的混合方法^[8,16]在一定程度上平衡了重建质量和推理时间.此外,Alaluf 等^[17]和 Dinh 等^[18]通过训练一个较小的网络为 StyleGAN 产生权值,

实现更精确的重建;Liu 等^[19]利用初始重建图像与原始图像之间的频域间隙调整生成器权重,重建图像的高频部分.然而,已有的 GAN 反演方法仍然不足以精确地重建建筑图像.

扩散模型展示了基于文本输入的高质量图像生成^[20],这些模型把随机采样的噪声通过多次迭代去噪,生成逼真的图像.基于多模态扩散的图像编辑方法^[21]是一种用文本引导扩散模型编辑图像的方法,拥有更好的生成质量,但是也需要更长的生成时间和更昂贵的设备需求,并且在结构的可控性上仍有不足.

1.2 基于 GAN 隐空间的编辑方法

Xia 等^[22]对 GAN 模型存在的一些隐空间进行总结,除了一般 GAN 的 Z 空间外,还有专门为 StyleGAN 设计的 W 空间、 $W+$ 空间和风格空间 S .由于 StyleGAN 模型实现了先进的 GAN 方法图像合成,因此当前最先进的 GAN 反演技术都围绕着 StyleGAN 的隐空间展开.

GAN 通常是从简单分布(如正态分布或均匀分布)中采样的值映射到生成的图像,这种简单分布被称为 Z 空间,该空间适用于所有的无条件 GAN 模型,如 BigGAN^[3],StyleGANs^[6]等.然而,正态分布的 Z 空间的约束限制了对语义属性的表示能力和解缠性.StyleGAN 由映射网络和合成网络组成,映射网络使用 8 层多层感知机(multilayer perceptron, MLP)实现将 z 向量映射为风格隐码 w , w 所在的空间称为 W 空间.合成网络将风格隐码 w 输入对应的输入层,生成高质量的图像,其合成过程从随机初始化的、空间分辨率为 4×4 的常量特征图开始,随着不断上采样,最终得到 256×256 的结果图像.在逐级生成过程中,每个风格块在权值变换中以 1×512 的隐码 w 作为输入,与后续的卷积运算相关联,控制生成过程.StyleGAN 通过改变隐码 w 控制生成图像的属性.

经过变化后, W 空间不服从于某一已知的特定分布,可以更好地描述所学习数据集的分布,表现出更多的解纠缠性^[13].然而,因为 W 空间包含的信息量较少,所以一些方法将 W 空间拓展到 $W+$ 空间^[14-15]. $W+$ 空间将不同的隐码 w 应用于 StyleGAN 的不同层.将图像反转到 $W+$ 空间虽然可以减少失真,但也会损失编辑性能.风格空间 S ^[23]由风格参数 s 组成,其中, s 表示每个隐码 w 经过仿射变换后得到的风格隐码,其维度和 w 相同.与 $W+$ 空间相比,虽然映射到 S 空间的失真更小,但也损失了更多的编辑性能,所以这种失真仍然无法保证建

筑图像的结构一致性。

以上空间均为较低维的向量空间, 根据数据率失真理论^[9], 将图像反演成低维潜代码不可避免地会导致信息丢失。因此, 一些方法将图像反演到更高维的空间来保留图像的结构信息。 F 空间由合成网络每一层卷积后得到的风格特征 f 组成, 该空间是分层空间, 每层空间的维度是 $c \times w \times h$; 其中, c 表示通道数, w 和 h 表示特征的长和宽。经过计算, S 空间前 9 层维度为 2793472, 全部 14 层维度为 15376384, 远大于 W 空间的 512 和 $W+$ 空间的 7168。Zhu 等^[24]提出 Barbershop 方法, 通过分割掩码将图像嵌入 F 和 S 隐空间, 实现了高质量的发型编辑; Liu 等^[25]将图像转换到 W 空间, 并利用交叉注意力引导 W 空间向 $W+$ 空间和 F 空间的优化; Li 等^[26]引入逐步细节注入方案, 开发了渐进残差模块和全局对齐模块, 以提高编辑质量。然而这些方法都专注于将低维空间引导向高维空间, 或者注入部分高维空间特征以缓解重建失真的问题, 仍然无法在精细尺度下还原建筑结构和纹理。

为了解决图像重建中带来的伪影和扭曲问题, ABEditor 选择将图像映射到高维 F 空间, 并在后期层通过 $W+$ 空间还原图像的高维色彩信息。实验结果表明, 将图像映射到 F 空间可以得到更接近于真实的特征 f , 进而获得更好的重建质量; 在网络的后期层利用 $W+$ 向量进行色彩校正, 以保留出色的编辑效果。

2 ABEditor

ABEditor 是一种基于深度生成模型的建筑图像编辑模型, 其结构如图 1 所示。其中, RSN 模块通过残差连接到 StyleGAN 合成网络, 使整体从无条件生成模型变为有条件生成模型。

给定一幅图像 I_{input} , 本文的目标是学习一个函数 R , 它能够将输入图像编码到 F 空间, 得到特征 $f = R(I_{\text{input}})$, 特征 f 再经过解码器进行解码, 输出

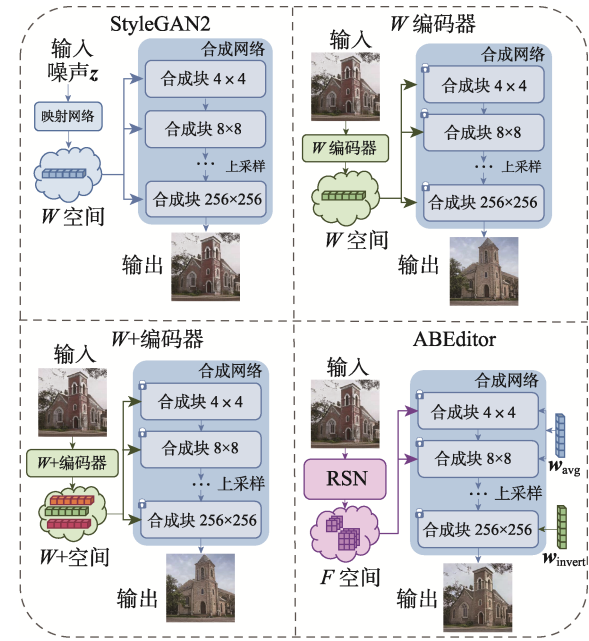


图 1 ABEditor 结构

重建图像 $I_{\text{recon}} = G(f, w_{\text{avg}})$; 其中, w_{avg} 表示解码器中的平均隐码, 为已知的固定值。如图 2 所示, 首先优化函数 R , 再改变解码器中的隐码 w 完成编辑, 得到编辑后的输出图像 $I_{\text{output}} = G(R(I_{\text{input}}), w_{\text{edit}})$ 。

函数 R 为本文设计的编码器 RSN, 解码器是预训练 StyleGAN2 的合成网络, 在训练 ABEditor 时参数固定不动。为了更好地对齐编码器和解码器的特征空间, 采用一种自监督的训练方法, 将合成网络输出的图像 $I_{\text{input}} = G(w_{\text{gt}})$ 作为编码器的输入。其中, $w_{\text{gt}} = M(z)$, $M(\cdot)$ 表示 StyleGAN 的映射网络, z 表示随机采样的高斯噪声, 因此只需要不断地采样噪声 z 即可完成训练。整体优化目标为

$$\min_R \left(G(w_{\text{avg}}, R(G(w_{\text{gt}}))), G(w_{\text{gt}}) \right).$$

2.1 RSN 的结构

ABEditor 中的合成网络使用预训练的 StyleGAN 模型^[4]。RSN 可以将图像映射到 StyleGAN 的 F 空间, 学习风格特征可以保留大量的信息。对于每幅

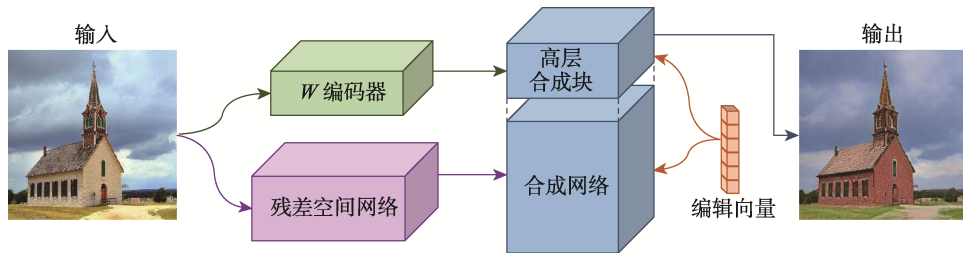


图 2 ABEditor 结构概述

StyleGAN 生成的图像,在噪声每次上采样前得到的特征蕴含大量的属性信息,RSN 旨在找到 StyleGAN 生成图像时每层特征的规律,从而学习到现实图像的属性信息. RSN 首先使用完整的 StyleGAN2 网络得到自监督样本对;然后将合成网络需要的隐码替换为平均隐码 w_{avg} ,将得到的图像样本作为 RSN 的输入. 如图 3 所示,RSN 由 7

个结构相同的 RS 块组成,输入为 StyleGAN 生成的图像样本 $I = G(w_{gt})$,输出为 13 个风格空间特征和 6 个空间图像. StyleGAN 的合成块也有对应的 7 个,包含 13 个空间特征和 7 幅空间图像;然后将 RSN 输出的风格空间特征连接到对应维度的合成块上,最后一幅空间图像(输出图)不进行连接;最终,ABEditor 的输出为合成网络输出的第 7 幅空间图像.

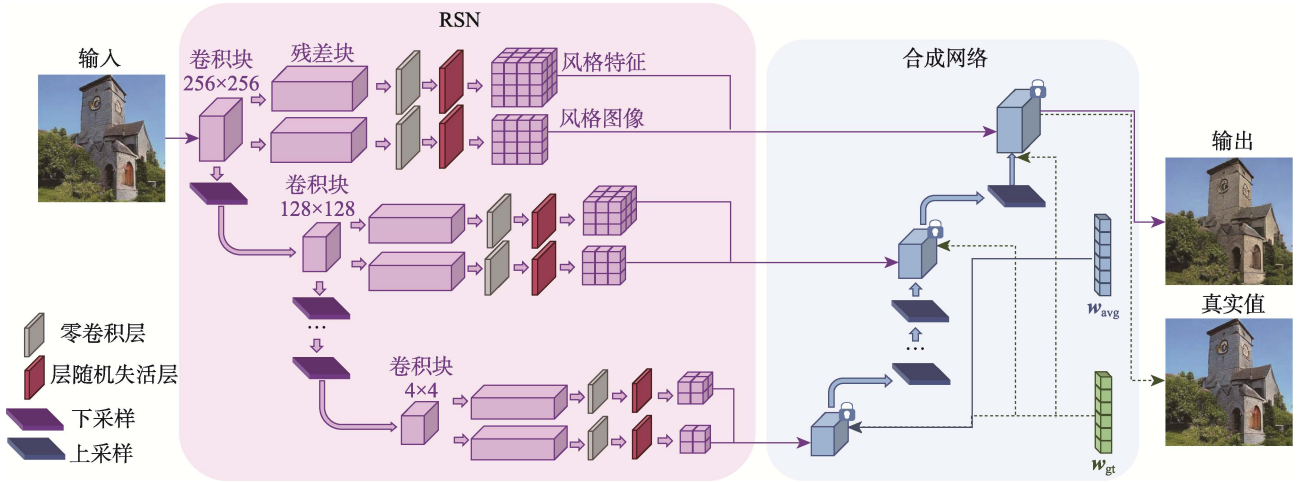


图 3 RSN 结构

每层 RSN 中,如图 4 所示,每个分辨率的 RS 块由卷积层、激活层 ReLU、残差块、零卷积层和层随机失活层 Layer Dropout 组成. 其中,每个 RS 块的卷积层和激活层依次串联,输入/输出特征分别记为 f_i 和 f_{i-1} . 为了让模型更快、更好地收敛,卷积层参数使用预训练的 VGG 来初始化,可以利用学习到的通用特征,如边缘、纹理和形状等,因此可以很好地提取图像特征,加快模型收敛. 残差块包含了 2 个残差连接的卷积层,可以学习更多深层特征,本文使用了 4 个残差块. 零卷积层是权值和偏置初始化为 0 的卷积层,将使训练前的第 1 次前向传播的输出与不连接 RSN 时 StyleGAN 的输出一致,从而使刚开始的每次训练都更靠近优化目标. Layer Dropout 由每 i 层的概率参数 p_i 控制,随机地使 RS 块不产生输出;在训练时使用真实的隐码 w_{gt} 输入到合成块中,则该层的特征空间会更接近于真实值;还可以让模型加速收敛,减少过拟合的可能. 因此,ABEditor 合成网络的每层空间特征 F_i 和空间图像 I_j 分别表示为

$$F_i = G^i(p_i w_{avg} + (1 - p_i) w_{gt}, F_{i-1}) + p_i R^i(x), i \neq 0;$$

$$I_j = G^j(p_j w_{avg} + (1 - p_j) w_{gt}, F_{j-1}) + p_j R^j(x) + U(I_{j-1}), j \neq 0;$$

$$G^i(w_i, F_i) = M(a(w_i), F_{i-1}).$$

其中, $i \in \{1, 2, 3, \dots, 13\}$, $j \in \{1, 2, \dots, 7\}$, $i = 2j - 1$; $G^i(\cdot)$ 表示合成网络的第 i 个特征输出层; $R^i(\cdot)$ 表示第 i 个 RS 块; $U(\cdot)$ 表示 2 倍上采样; $M(\cdot)$ 表示调制卷积; $a(\cdot)$ 表示仿射变换,本文中使用时线性层; p_i 决定第 i 个 Layer Dropout 层是否会激活,其值为 0 或 1; x 表示输入图像; w_i 表示对应的隐码; $j = 0$ 时, $I_0 = 0$; $i = 0$ 时, F_0 为固定好的维度为 $512 \times 4 \times 4$ 的随机初始化特征; $j = 7$ 时, I_7 即为得到的输出图像.

2.2 ABEditor 推理过程

将全部的 RSN 层映射到合成网络的隐空间时,模型容易产生过拟合现象. 这是因为后期层特征具有更高的空间维度,并且包含了大量的细节信息,在图像重建的过程中,这些细节信息可能会导致图像上出现不必要的纹理和伪影,影响编辑效果. 为了解决这个问题,ABEditor 在推理过程中选择性地使用部分 RSN 层进行残差连接,以获得更佳的编辑效果. 考虑早期层隐空间更多体现了图像的结构和线条信息,所以本文选用中早期层进行高维空间映射,但这种方法也意味着失去了部分后期层特征的映射,可能会导致色彩的丢失. 为

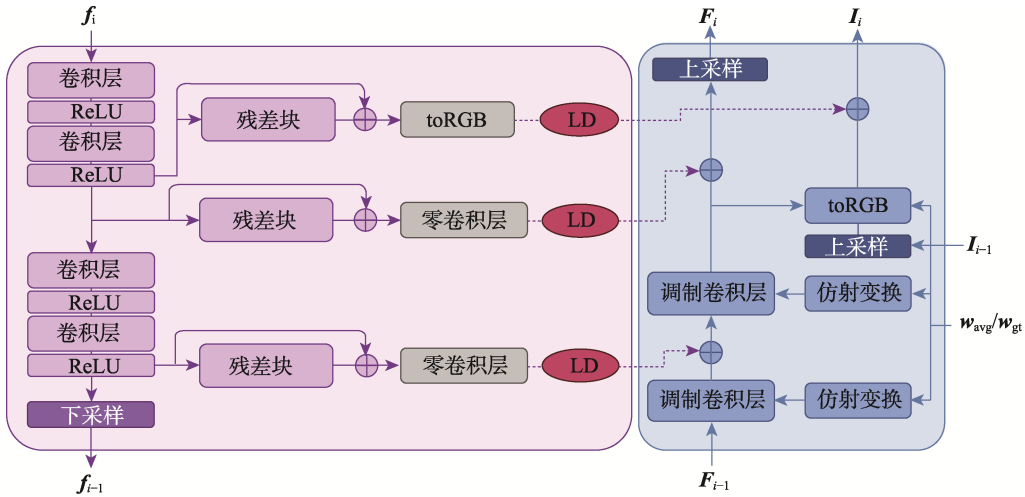


图 4 RS 块结构

了弥补这一缺陷,在推理时,对于未连接的层数部分,本文采用目前效果较好的 W 编码器^[14] $E(\cdot)$ 将图像映射到合成网络的 W 空间,既减少过拟合的风险,又尽可能地保留了图像的色彩信息,从而提高编辑效果的整体质量.反演得到的隐码为

$$\mathbf{w}_{\text{invert}} = E(\mathbf{I}_{\text{input}}).$$

完成 ABEditor 的训练流程后,编辑操作在推理阶段可通过调整输入隐码 $\mathbf{w}$ 来灵活实现.本文沿用 InterFaceGAN^[27] 的方法,系统地实施了属性编辑策略.如图 5 所示,首先利用 StyleGAN 模型生成大规模的图像数据集,并借助属性预测模型对每幅图像进行属性量化评分;然后,分别挑选出得分最高和最低的 1000 个正负样本,确保样本的代表性和准确性.基于这些样本及其对应的隐码 $\mathbf{w}$,训练一个支持向量机(support vector machine, SVM)二分类模型,得到特定属性的决策边界,此边界的法向量 $\mathbf{d}$ 作为关键参数,用于指导属性编辑的方向.通过向原始隐码 $\mathbf{w}$ 添加编辑向量 $\sigma \mathbf{d}$,计算出修改后的 $\mathbf{w}$ 向量 $\mathbf{w}_{\text{edit}} = \mathbf{w}_{\text{avg}} + \sigma \mathbf{d}$;其中, σ 表示控制编辑的强度.最终得到编辑后的图像

$$\mathbf{I}_{\text{output}} = G(R(\mathbf{I}_{\text{input}}), \mathbf{w}_{\text{invert}}, \mathbf{w}_{\text{edit}}).$$

2.3 损失函数

由于 ABEditor 的目标是预测图像映射到 StyleGAN 空间的残差特征,同时保持加入此空间残差后的预训练 StyleGAN 的生成质量,因此需要精确地将编码特征映射到原始合成过程中的对应部分.训练过程的目标函数表述为

$$L_{\text{total}} = \lambda_{L_2} L_2 + \lambda_{L_{\text{lp}}} L_{\text{lp}} + \lambda_{L_{\text{grad}}} L_{\text{grad}} + \lambda_{L_{\text{spatial}}} L_{\text{spatial}} + \lambda_{L_{\text{adv}}} L_{\text{adv}}.$$

其中, $L_2 = \|\mathbf{I}_{\text{input}} - \mathbf{I}_{\text{output}}\|_2$ 表示 L_2 范数损失函数,用于测量像素相似度; $L_{\text{lp}} = \|F_{\text{lp}}(\mathbf{I}_{\text{input}}) - F_{\text{lp}}(\mathbf{I}_{\text{output}})\|_2$ 表示全局感知损失; $L_{\text{grad}} = \|\nabla(\mathbf{I}_{\text{input}}) - \nabla(\mathbf{I}_{\text{output}})\|_2$ 表示图像梯度损失,其可以很好地表示图像中物体的边缘,保留图像良好的梯度属性,在一定程度上保证了图像的真实感和保真度^[28]. F_{lp} 表示感知特征提取器^[29];使用预训练的 AlexNet^[30] 版本; $\nabla(\cdot)$ 表示图像梯度.

另外,为了使图像在每个特征层可以得到良好的表达,在目标函数中增加了一个空间匹配损失,公式为

$$L_{\text{spatial}} = \frac{1}{N} \sum_l \|G^l(\mathbf{w}_{\text{gt}} - G^l(\mathbf{w}_{\text{avg}}, R^l(\mathbf{I}_{\text{output}})))\|_2.$$

其中, $G^l(\cdot)$ 表示预训练 StyleGAN 合成网络的第 l 个卷积输出的空间特征或空间图像; $R^l(\cdot)$ 表示与 $G^l(\cdot)$ 对应维度的残差空间.

原始的 StyleGAN 使用对抗损失指导网络收敛,使用预训练的 StyleGAN 的判别器指导编码器收敛到原始的中间空间.对抗损失 L_{adv} 和判别器损失 L_D 分别定义为

$$L_{\text{adv}} = -E_{\mathbf{I}_{\text{output}}} [\lg(D(\mathbf{I}_{\text{output}}))],$$

$$L_D = -E_{\mathbf{I}_{\text{input}}} [\lg(D(\mathbf{I}_{\text{input}}))] - E_{\mathbf{I}_{\text{output}}} [\lg(D(1 - \mathbf{I}_{\text{output}}))] + E_{\mathbf{I}_{\text{input}}} [\|\nabla_{\mathbf{I}_{\text{input}}} D(\mathbf{I}_{\text{input}})\|_2^2];$$

其中, D 为判别器,使用预训练 StyleGAN 鉴别器权值进行初始化,其与 RSN 一起以对抗的方式进行训练.最后,进一步对 L_D 施加 R_1 正则化^[31].

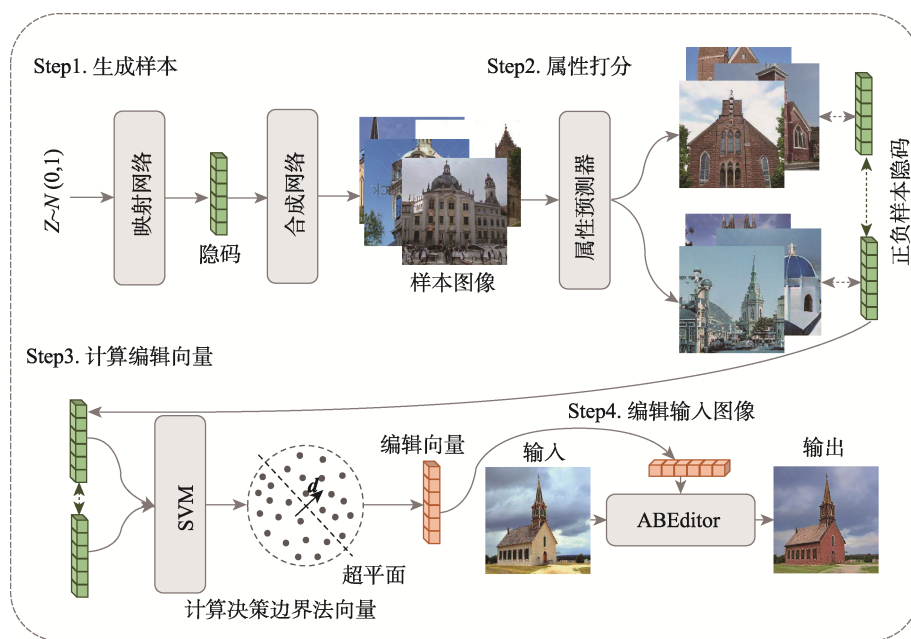


图 5 对砖块纹理编辑属性示例 ABEditor 的推理过程

3 实验及结果分析

3.1 数据集

本文使用 LSUN Church 数据集^[32]训练 StyleGAN 的合成网络部分,并将随机采样的 $\mathbf{z}$ 输入到合成网络得到的生成教堂数据作为 RSN 的训练集. LSUN Church 数据集包含 126 226 幅不同分辨率的教堂图像,为了便于训练,本文将图像尺寸统一为 256×256 像素.由于缺少编辑前与编辑后成对的效果图,因此采用 StyleGAN 生成的假教堂数据集作为编辑后的真值图.

3.2 实验设置

本文使用 Adam^[33]优化器,训练时,设置学习率为 0.000 1, $\beta_1 = 0.5$, $\beta_2 = 0.999$, 训练批大小为 8, 迭代次数为 200 000 次. 损失函数超参数设置为 $\lambda_{L_2} = 1.0$, $\lambda_{L_p} = 0.8$, $\lambda_{L_{grad}} = 0.3$, $\lambda_{L_{spatial}} = 1.0$ 和 $\lambda_{L_{adv}} = 0.15$. 每 i 层的 Layer Dropout 发生的概率为 0.8. 使用网格搜索验证 $\lambda_{L_{grad}}$ 和 $\lambda_{L_{spatial}}$ 的设置,其他超参数和学习率沿用了 E4E^[14]的数值,同样使用网格搜索验证它们的有效性.

在训练中,使用 1 张 Tesla V100 SMX3 显卡,显存为 32 GB. 在推理时,使用 1 张 3060 Laptop 显卡,显存为 6 GB. 模型的训练时间为 24 h,推理时间为 0.11 s.

3.3 评价指标

因为 ABEditor 遵从先重建后编辑的原则,所以从评价指标上分为重建质量和编辑效果. 评估重

建质量时,采用像素级 L_2 距离,峰值信噪比(peak signal-to-noise ratio, PSNR)^[34], 结构相似性指数(structural similarity index, SSIM)^[35], 学习感知图像块相似度(learned perceptual image patch similarity, LPIPS)^[29]和弗雷歇初始距离(Fréchet inception distance, FID)^[36]指标.

评估编辑效果时,采用属性改变量 ΔS ^[18]和 APEDS 指标. ΔS 指标旨在评估建筑图像编辑前和编辑后建筑属性类别分值的改变量,该指标在其他任务中被广泛采用^[18],用于评估编辑的幅度,其值越大,意味着编辑效果越显著. 本文通过预训练的建筑属性分类器^[37]对生成图像进行属性预测, ΔS 可以在一定程度上反映对图像中建筑属性的修改程度,但由于 ΔS 是对整幅图像进行计算,因此很容易受非建筑属性的影响. 例如,当生成图像中,天空区域包含很多砖块信息,该属性改变量依然很高,忽略了建筑结构语义信息,很难客观地评估编辑效果.

为了解决上述问题,本文提出 APEDS 评价指标借助语义分割模型^[38]把建筑和天空进行分割,在此基础上,计算天空和建筑的像素平均值. 首先用 Oneformer 分割模型^[38]将图像分割成建筑、天空、地面、树和其他背景 5 类,提取出每个类别 c 的像素总和 $p_c = \sum_i \sum_j S_c(I_{ij}) \cdot I_{ij}$. 其中, $S(\cdot)$ 表示分割函数,结果为 0 或 1; c 表示对应像素的分割分类; I_{ij} 表示输入图像对应的第 i 行第 j 列像素. 然

后分别求出真实图像、真实编辑图像、重建图像以及重建编辑图像的各类像素总和 $p_c^g, p_c^{ge}, p_c^r, p_c^{re}$; 真实编辑后的移动向量为真实像素与真实编辑像素的差值, 重建的编辑移动向量为重建像素与重建编辑像素的差值, 则平均像素编辑距离为重建的编辑移动向量和真实的编辑移动向量之间的距离. 距离越近, 说明编辑效果越接近真实值. APEDS 表示为

$$APEDS = \frac{1}{n} \sum_c \sqrt{\left((p_c^g - p_c^{ge}) - (p_c^r - p_c^{re}) \right)^2}.$$

其中, n 表示所有类别像素的数量和.

3.4 建筑图像编辑结果

3.4.1 重建质量

(1) 定性评估. 在 LSUN Church 数据集上, 将 ABEditor 与 PSP^[7], E4E^[14], HyperStyle^[17], HyperInverter^[18], HFGI^[15]和 CLCAE^[25]方法进行实验, 7 种方法对于建筑图像的重建结果如图 6 所示. 可以看出, 对于图像 1 城堡上的小窗户, ABEditor 可以保留更多的细节, 很好地还原城堡上的几个小窗户, PSP^[7]和 E4E^[14]只复原了建筑的大致轮廓, 窗户数量和位置都完全错误, 而 HyperInverter^[18]则将窗户歪七扭八地杂糅在了一起; 在图像 3 的木房子中, ABEditor 很好地还原了房顶和房间主体的分



图 6 在 LSUN Church 数据集上 7 种方法的重建质量定性对比

界的直线, HFGI^[15]和 CLCAE^[25]虽然也都做到了这些, 但图像较模糊, 并且在一些细节方面未能完好地重建, 如第 1 列城堡顶部的线条处。

(2) 定量评估. 针对 7 种方法的重建质量进行实验, 其评价指标对比结果如表 1 所示. 可以看出, ABEditor 的各项指标明显优于其他基于编码器的方法: ABEditor 在 PSNR, SSIM, LPIPS 和 FID 上均高于其他 5 种方法; 其中, SSIM 和 FID 相较于对比方法中最好的方法分别提升 28.64%和 20.74%; L_2 , LPIPS 和 PSNR 相较于对比方法中最好的方法分别提升 13.67%, 1.67%和 10.01%. 定量评估结果证明了 ABEditor 在建筑图像重建质量方面的有效性。

表 1 在 LSUN Church 数据集上 7 种方法重建质量指标对比

方法	$L_2\downarrow$	PSNR/dB $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$
PSP ^[7]	0.1714	17.2509	0.4124	0.2493	53.0281
E4E ^[14]	0.2061	15.7616	0.3891	0.3393	55.1521
HyperStyle ^[17]	0.1374	18.9064	0.4617	0.1438	51.8011
HyperInverter ^[18]	0.1788	16.5647	0.4155	0.1949	47.2026
HFGI ^[15]	0.1457	18.2558	0.4731	0.2119	49.7725
CLCAE ^[25]	0.1456	18.9818	0.5289	0.1601	65.7148
ABEditor	0.1257	20.8814	0.6804	0.1414	37.4121

注: 粗体表示最优值。

3.4.2 编辑效果

本文选择砖块、植被、石头、混凝土和时间这 5 种具有代表性的属性进行编辑效果的评估. 其中, 前 4 种使用属性预测器^[37], 时间属性利用 ResNet18^[39]进行二分类, 分别筛选出所有属性得分最高及最低的 1000 个正负样本。

(1) 定性评估. 在 LSUN Church 和 StyleGAN 生成图像这 2 个数据集上, 对 7 种方法的编辑效果进行实验。

在 LSUN Church 数据集上, 对时间属性和砖块属性这 2 个样例进行编辑, 定性结果如表 2 所示, 其中, 增幅表示不同方法采用相同的 σ 值. 从第 1 个样例可以看出, 输入和重建结果中, 将图像映射到高维隐空间, 使得 ABEditor 在重建质量上更多关注到建筑结构信息, 重建效果优于其他方法; 输入、增幅 1 和增幅 2 结果中, ABEditor 仅在网络的后期层映射到低维空间, 因此仍能保证较好的编辑效果, 并且在编辑方面考虑到建筑结构的线条细节, 保证了建筑结构的完整性和合理性, 使整

体生成效果更加逼真。

使用 StyleGAN 生成图像作为数据集时编辑效果的定性结果如表 3 所示. 可以看出, 与其他方法相比, ABEditor 的编辑效果更接近于真实值; 即使忽略重建质量, 其编辑效果也与其他方法中编辑效果最好的 E4E^[14]相似, 甚至改变更加真实, 如将图像转换到夜晚时. 这是因为加入的残差风格特征可以更好地在特征空间对齐, 学习到更准确的风格。

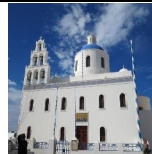
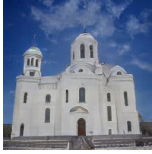
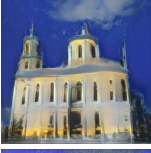
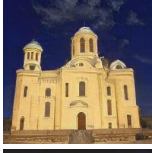
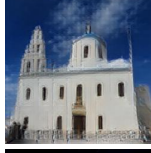
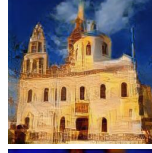
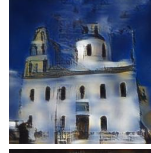
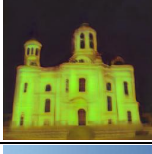
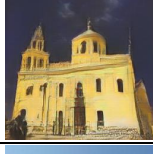
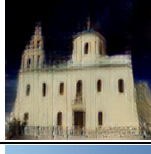
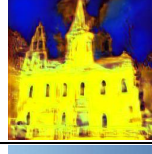
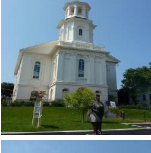
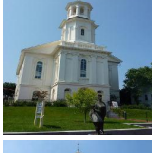
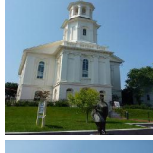
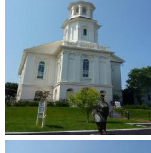
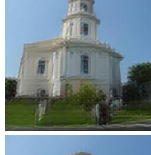
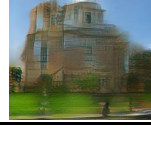
(2) 定量评估. 对砖块、混凝土、石头、时间和植被 5 种属性的 ΔS 和 APEDS 指标进行定量评估, 结果如表 4 所示. 可以看出: a. 在 ΔS 指标中, 除时间外的 4 个属性中, E4E^[14]得到 2 项最优, HFGI^[15]和 ABEditor 分别得到 1 项最优; 尽管 HFGI^[15]的编辑得分都不低, 但在表 3 增加砖块属性后, 虽然 HFGI^[15]获得更高的得分, 但它的结构保留十分不稳定; 在其他属性中, ABEditor 的编辑得分与 E4E^[14]相近, 远高于 PSP^[7], HyperStyle^[17]和 HyperInverter^[18]. b. 在 APEDS 指标中, ABEditor 在 4 个属性中都获得了最佳的结果; 对于未编辑的图像, 计算出的 APEDS 是最差结果, 编辑石头属性时, PSP^[7], HyperStyle^[17]和 HyperInverter^[18]的 APEDS 分别为 0.025 9, 0.025 7 和 0.025 0, 与最差结果(未编辑)0.024 2 相近; 结合表 3 也可以发现, 这 3 种方法编辑效果不明显; 同样, CLCAE^[25]对属性编辑较为敏感, 虽然得到较好的 ΔS 评估结果, 但其在编辑中也会导致其他信息的改变, 造成图像大规模失真和扭曲, 使得 APEDS 得分较高; 只有 ABEditor 有较大的编辑幅度; E4E^[7]在编辑时间属性时的效果最优, 幅度略微大于 ABEditor, 天空和建筑的颜色更接近真实值; 所有属性的编辑中, ABEditor 的 APEDS 基本都是未编辑时 APEDS 的一半左右, 其他方法只有在几个属性有较小的改变; 总体而言, 与相关方法相比, 其在 APEDS 中提升约 21.26%。

3.5 消融实验

3.5.1 使用不同层数的 RSN 对生成结果的影响

RSN 输出 13 个风格空间特征, 并各自嵌入合成网络中对应维度的特征中. 表 5 所示为 2 幅图像每层连接的可视化效果, 并通过仅连接 n 层和连接 0~ n 层 2 种方式生成图像的结果进行定量评估, 分析连接不同层数的重建质量和编辑效果. 可以看出: (1) 第 1 幅输入图像中, RSN 连接层为 2~5 时, 对图像结构影响最大; 连接层数为 9~13 时,

表 2 在 LSUN Church 数据集上 7 种方法编辑效果对比

样例	PSP ^[7]	E4E ^[14]	HyperStyle ^[17]	HyperInverter ^[18]	HFGI ^[15]	CLCAE ^[25]	ABEditor
1	输入						
	重建						
	增幅 1						
	增幅 2						
2	输入						
	重建						
	增幅 1						
	增幅 2						

对建筑纹理、细节影响最大, 出现了一定的伪影。
 (2) 第 2 幅输入图像中, 1~6 层改变了图像的结构, 如建筑布局、窗户形状等; 7~13 层影响了云的数量, 建筑的颜色, 以及植物、建筑纹理的细节。Tov 等^[14]发现, 重建效果越好时, 编辑效果越差。如图 7 所示。ABEditor 对不同重建层数的重建质量和编辑效果也体现了这一特点。可以看出, 当风格空间特征连接层数增多时, 纹理重建效果变好, 纹理编辑效果变差; 如果连接全部层数, 则仅改变了建筑颜色。图 8 所示为不同层数归一化后的重建质量(归一化 PSNR)和编辑效果(归一化 ΔS)。可以看出,

在 3~12 层时, 随着风格空间特征层数连接的增多, 重建质量越来越好; 在 0~8 层以及 9~13 层时, 随着风格空间特征层数连接的增多, 编辑效果和编辑幅度越来越差; 连接层数为 9 时编辑效果最好, 并且拥有较好的重建质量, 这是因为 10~13 层为高细粒度特征, 可以更好地还原图像细节, 但这种高维特征往往会附着在图像上, 导致图像难以编辑。综合考虑, 本文选择使用效果最出色的第 9 层, 同时可以在推理时自由选择层数; 若想要得到更好的重建质量, 则可以增加连接层数, 但会牺牲一定的编辑效果。

表 3 使用 StyleGAN 生成图像作为数据集时 7 种方法的编辑效果对比

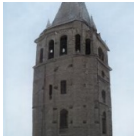
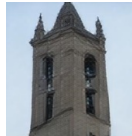
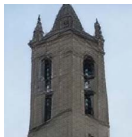
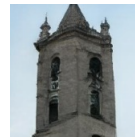
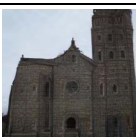
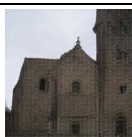
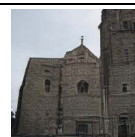
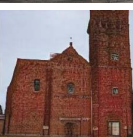
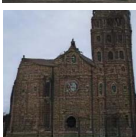
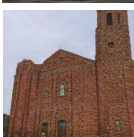
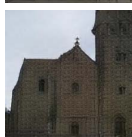
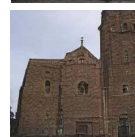
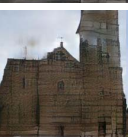
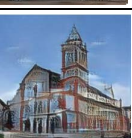
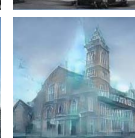
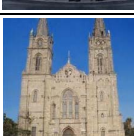
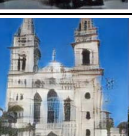
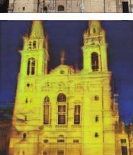
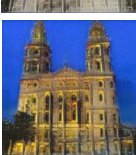
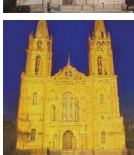
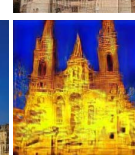
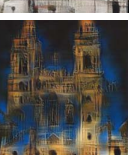
样例	真实值	PSP ^[7]	E4E ^[14]	HyperStyle ^[17]	Hyper-Inverter ^[18]	HFGI ^[15]	CLCAE ^[25]	ABEditor	
1	重建								
	重建+石头								
2	重建								
	重建+砖块								
3	重建								
	重建-砖块								
4	重建								
	重建-时间								

表 4 5 种属性编辑效果指标对比

编辑元素	评估指标	未编辑	PSP ^[7]	E4E ^[14]	HyperStyle ^[17]	HyperInverter ^[18]	HFGI ^[15]	CLCAE ^[25]	ABEditor
砖块	APEDS↓	0.0729	0.0660	0.0424	0.0623	0.0532	0.1181	0.0927	0.0411
	ΔS↑	0.0000	0.1328	0.1700	0.0130	0.0942	0.1941	0.1508	0.1810
混凝土	APEDS↓	0.0350	0.0350	0.0293	0.0342	0.0284	0.0511	0.0451	0.0202
	ΔS↑	0.0000	0.0105	0.0587	0.0003	0.0098	0.0561	0.0439	0.0557
石头	APEDS↓	0.0242	0.0259	0.0239	0.0257	0.0250	0.0409	0.0312	0.0200
	ΔS↑	0.0000	0.0129	0.1395	0.0024	0.0442	0.0831	0.0759	0.1319
时间	APEDS↓	0.2086	0.1373	0.0724	0.1202	0.1256	0.2930	0.2267	0.0760
	ΔS↑								
植被	APEDS↓	0.0872	0.0809	0.0498	0.0809	0.0676	0.1000	0.0935	0.0337
	ΔS↑	0.0000	0.0241	0.1755	0.0034	0.0159	0.2246	0.1737	0.2325

注. 粗体表示最优值.

表 5 使用不同层数的 RSN 对生成结果的影响

输入		输出为仅连接 n 层时的结果						
图像	层数							
	0~6							
	7~13							
输入		输出为连接 0~ n 层时的结果						
图像	层数							
	0~6							
	7~13							

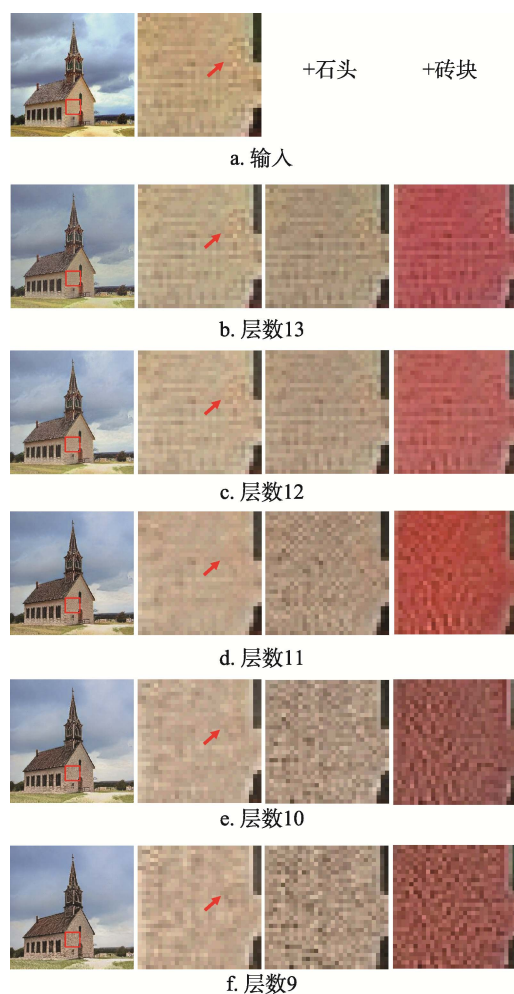


图 7 不同重建层数的重建质量和编辑效果

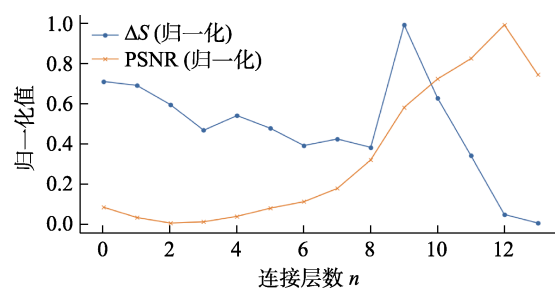


图 8 RSN 连接各层的重建质量和编辑效果对比

3.5.2 各组件对生成结果的影响

通过 5 组消融实验, 验证各组件对生成结果的影响. (1) 只使用 W 编码器, 缺少 RSN 的原始模型 (简称为 w/o RSN); (2) 缺少 Layer Dropout 的模型 (简称为 w/o L Dropout); (3) 缺少图像梯度损失的模型 (简称为 w/o Grad Loss); (4) 缺少 W 编码器的模型 (简称为 w/o W 编码器); (5) 带有全部 RSN 组件的模型 (即 ABEditor). 消融实验重建图像的定性结果如图 9 所示, 消融实验的定量结果如表 6 所示. 可以看出, w/o RSN 中, 仅剩将图像映射到 W 空间的 W 编码器, 建筑结构失真明显, 除了建筑主体位置和颜色, 其他属性均不能重建; w/o L Dropout 中, 在连接第 9 层风格空间特征时损失了大量的几何结构, 如窗户形状、数量; w/o Grad Loss 中, 尽管窗户的位置和数量得到了一定的保证, 但窗户边缘有一定程度的扭曲, 并且建筑材质、纹理有很大的模糊; w/o W 编码器中, 由于高维层数上使用



图 9 消融实验重建图像对比

表 6 消融实验结果对比

方法	PSNR/dB↑	SSIM↑	FID↓
w/o RSN	15.2134	0.3778	55.2675
w/o L Dropout	16.8691	0.4466	56.6596
w/o Grad Loss	18.1327	0.5232	45.0103
w/o W 编码器	20.1065	0.6676	39.9969
ABEditor	20.8814	0.6804	37.4121

注: 粗体表示最优值.

的隐码是 w_{avg} , 因此生成的图像在高层次细节上, 如天气和建筑颜色等, 更偏向于使用平均隐码 w_{avg} 生成图像时产生的属性风格, 不能很好地还原原图的天气和建筑颜色.

4 结 语

为了解决建筑属性编辑领域中存在的问题, 本文提出 ABEditor. 与传统的 GAN 反演编码器将图像编码到低维 W 空间不同, 本文引入 RSN, 将

图像直接编码到 StyleGAN 的高维风格特征 F 空间中, 解决传统方法空间信息缺失、重建质量差的问题, 平衡图像的重建质量和属性编辑效果, 显著地提高了模型的性能; 针对已有方法对编辑效果评估的不足, 提出 APEDS 指标, 更好地分离了重建质量对编辑效果的影响. 实验结果表明, ABEditor 的结果明显优于其他编码器方法, 不仅能够保留图像细节, 还能实现出色的编辑效果.

ABEditor 也有一定的局限性, 虽然在中后期层对应的属性编辑中具有很好的编辑质量, 如改变建筑材质、建筑颜色、环境(植被、天气、时间等), 但编辑网络早期层对应的属性时往往效果较差, 如改变建筑结构, 改变窗户大小. 未来将进一步改进编辑效果, 探索新的训练策略和网络结构, 以提高模型的性能.

参考文献(References):

- [1] Jiang S D, Yan Y Y, Lin Y M, et al. Sketch to building:

- architecture image translation based on GAN[J]. *Journal of Physics: Conference Series*, 2022, 2278(1): Article No.012036
- [2] Nauata N, Hosseini S, Chang K H, *et al.* House-GAN++: generative adversarial layout refinement network towards intelligent computational agent for professional architects[C] // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2021: 13627-13636
 - [3] Brock A, Donahue J, Simonyan K. Large scale GAN training for high fidelity natural image synthesis[C] // *Proceedings of the 7th International Conference on Learning Representations*. Washington DC: ICLR, 2019: 1-35
 - [4] Luan F J, Paris S, Shechtman E, *et al.* Deep photo style transfer[C] // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2017: 6997-7005
 - [5] Sangkloy P, Lu J W, Fang C, *et al.* Scribbler: controlling deep image synthesis with sketch and color[C] // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2017: 6836-6845
 - [6] Karras T, Laine S, Aittala M, *et al.* Analyzing and improving the image quality of StyleGAN[C] // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2020: 8107-8116
 - [7] Richardson E, Alaluf Y, Patashnik O, *et al.* Encoding in style: a StyleGAN encoder for image-to-image translation[C] // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2021: 2287-2296
 - [8] Zhu J P, Shen Y J, Zhao D L, *et al.* In-domain GAN inversion for real image editing[C] // *Proceedings of the 16th European Conference on Computer Vision*. Heidelberg: Springer, 2020: 592-608
 - [9] Gatys L A, Ecker A S, Bethge M. Image style transfer using convolutional neural networks[C] // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2016: 2414-2423
 - [10] Gan Yibo, Tan Zhiyi, Bao Bingkun. Joint intra-domain and inter-domain information modeling for image-to-image translation[J]. *Journal of Computer-Aided Design & Computer Graphics*, 2022, 34(10): 1489-1496(in Chinese)
(甘益波, 谭智一, 鲍秉坤. 联合图像域间和域内信息建模的图像风格转换[J]. *计算机辅助设计与图形学学报*, 2022, 34(10): 1489-1496)
 - [11] Chen Y S, Vu T A, Shum K C, *et al.* Time-of-day neural style transfer for architectural photographs[C] // *Proceedings of the IEEE International Conference on Computational Photography*. Los Alamitos: IEEE Computer Society, 2022: 1-12
 - [12] Liang Haorang, Zhu Zehao, Liang Ronghua. Scene text editing based on style transfer[J]. *Journal of Computer-Aided Design & Computer Graphics*, 2024, 36(9): 1362-1374(in Chinese)
(梁浩然, 朱泽浩, 梁荣华. 融合风格迁移的图像文本编辑方法[J]. *计算机辅助设计与图形学学报*, 2024, 36(9): 1362-1374)
 - [13] Abdal R, Qin Y P, Wonka P. Image2StyleGAN: how to embed images into the StyleGAN latent space?[C] // *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Los Alamitos: IEEE Computer Society Press, 2019: 4431-4440
 - [14] Tov O, Alaluf Y, Nitzan Y, *et al.* Designing an encoder for StyleGAN image manipulation[J]. *ACM Transactions on Graphics*, 2021, 40(4): Article No.133
 - [15] Wang T F, Zhang Y, Fan Y B, *et al.* High-fidelity GAN inversion for image attribute editing[C] // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2022: 11369-11378
 - [16] Roich D, Mokady R, Bermano A H, *et al.* Pivotal tuning for latent-based editing of real images[J]. *ACM Transactions on Graphics*, 2022, 42(1): Article No.6
 - [17] Alaluf Y, Tov O, Mokady R, *et al.* HyperStyle: StyleGAN inversion with HyperNetworks for real image editing[C] // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2022: 18490-18500
 - [18] Dinh T M, Tran A T, Nguyen R, *et al.* Hyperinverter: improving StyleGAN inversion via hypernetwork[C] // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2022: 11379-11388
 - [19] Liu F K, Shao M W, Wang F, *et al.* High-fidelity GAN inversion by frequency domain guidance[J]. *Computers & Graphics*, 2023, 114: 286-295
 - [20] Rombach R, Blattmann A, Lorenz D, *et al.* High-resolution image synthesis with latent diffusion models[C] // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2022: 10674-10685
 - [21] Kavar B, Zada S, Lang O, *et al.* Imagic: text-based real image editing with diffusion models[C] // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2023: 6007-6017
 - [22] Xia W H, Zhang Y L, Yang Y J, *et al.* GAN inversion: a survey[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, 45(3): 3121-3138
 - [23] Wu Z Z, Lischinski D, Shechtman E. StyleSpace analysis: disentangled controls for StyleGAN image generation[C] // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2021: 12858-12867
 - [24] Zhu P H, Abdal R, Femiani J, *et al.* Barbershop: GAN-based image compositing using segmentation masks[J]. *ACM Transactions on Graphics*, 2021, 40(6): Article No.215
 - [25] Liu H Y, Song Y B, Chen Q F. Delving StyleGAN inversion for image editing: a foundation latent space viewpoint[C] // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2023: 10072-10082
 - [26] Li H, Huang M Q, Zhang L, *et al.* Gradual residuals alignment: a dual-stream framework for GAN inversion and image attribute editing[C] // *Proceedings of the 38th AAAI Conference on Artificial Intelligence*. Palo Alto: AAAI Press, 2024: 3064-3072
 - [27] Shen Y J, Gu J J, Tang X O, *et al.* Interpreting the latent space of GANs for semantic face editing[C] // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*

- dition. Los Alamitos: IEEE Computer Society Press, 2020: 9240-9249
- [28] Mechrez R, Shechtman E, Zelnik-Manor L. Photorealistic style transfer with screened Poisson equation[C] // Proceedings of the British Machine Vision Conference. Durham: BMVA Press, 2017: 153.1-153.12
- [29] Zhang R, Isola P, Efros A A, *et al.* The unreasonable effectiveness of deep features as a perceptual metric[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 586-595
- [30] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C] // Proceedings of the 25th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2012: 1097-1105
- [31] Mescheder L M, Geiger A, Nowozin S. Which training methods for GANs do actually converge?[C] // Proceedings of the 35th International Conference on Machine Learning. New York: PMLR, 2018: 3478-3487
- [32] Yu F, Seff A, Zhang Y D, *et al.* LSUN: construction of a large-scale image dataset using deep learning with humans in the loop[OL]. [2023-10-27]. <https://arxiv.org/abs/1506.03365>
- [33] Kingma D P, Ba J. Adam: a method for stochastic optimization[OL]. [2023-10-27]. <https://arxiv.org/abs/1412.6980>
- [34] Wang Z, Bovik A C, Sheikh H R, *et al.* Image quality assessment: from error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612
- [35] Almohammad A, Ghinea G. Stego image quality and the reliability of PSNR[C] // Proceedings of the 2nd International Conference on Image Processing Theory, Tools and Applications. Los Alamitos: IEEE Computer Society Press, 2010: 215-220
- [36] Heusel M, Ramsauer H, Unterthiner T, *et al.* GANs trained by a two time-scale update rule converge to a local nash equilibrium[C] // Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2017: 6629-6640
- [37] Zhou B L, Lapedriza A, Khosla A, *et al.* Places: a 10 million image database for scene recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(6): 1452-1464
- [38] Jain J, Li J C, Chiu M T, *et al.* OneFormer: one transformer to rule universal image segmentation[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2023: 2989-2998
- [39] He K M, Zhang X Y, Ren S Q, *et al.* Deep residual learning for image recognition[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 770-778