计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
[doi: 10.15888/j.cnki.csa.010119] [CSTR: 32024.14.csa.010119]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于改进扩散模型的遗址类建筑物生成①

张旭欣[1], 吴　萌[1,3], 赵怀栋[2,3], 王　璐[3]

[1](西安建筑科技大学 信息与控制工程学院, 西安 710055)
[2](西安建筑科技大学 艺术学院, 西安 710055)
[3](西安建筑科技大学 交叉创新研究院, 西安 710055)
通信作者: 吴　萌, E-mail: wumeng@xauat.edu.cn

摘　要: 遗址类建筑物作为历史文化的重要载体, 具有重要的研究与保护价值. 然而, 由于其数量稀少且持续消亡, 传统重建方法难以实现完整复原, 现有文生图技术虽可借助文本描述重现其外观, 但仍存在细节缺失、图像质量不高等问题. 为此, 本文提出一种基于改进扩散模型的遗址类建筑物生成方法, 通过引入门控残差机制优化信息流动、缓解梯度消失, 提升生成稳定性; 结合通道与空间双重注意力机制以增强局部细节与全局结构建模能力; 并利用 VGG19 作为判别网络, 提取多层次语义特征并引入感知损失以提升对关键视觉特征的建模效果. 实验结果表明, 相比同样基于扩散模型的 KNN-diffusion 与 Simple diffusion, 本文方法 *FID* 下降了 30.39%, *CLIP-score*、*IS* 和 *SSIM* 分别提升了 1.08%、9.01% 和 2.35%. 本研究为高质量遗址类建筑图像生成提供了可行的技术路径, 有助于推动数字文化遗产的可持续研究与智能化保护.

关键词: 遗址类建筑物; 文本生成图像; 门控残差机制; 双重注意力网络; VGG19 判别网络

引用格式: 张旭欣,吴萌,赵怀栋,王璐.基于改进扩散模型的遗址类建筑物生成.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/10119.html

Generation of Heritage Buildings Based on Improved Diffusion Model

ZHANG Xu-Xin[1], WU Meng[1,3], ZHAO Huai-Dong[2,3], WANG Lu[3]

[1](College of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an 710055, China)
[2](College of Art, Xi'an University of Architecture and Technology, Xi'an 710055, China)
[3](Institute for Interdisciplinary Innovate Research, Xi'an University of Architecture and Technology, Xi'an 710055, China)

Abstract: As important carriers of historical and cultural heritage, heritage buildings hold significant value for research and conservation. However, their scarcity and ongoing deterioration make complete restoration through traditional methods challenging. While existing text-to-image generation techniques can reconstruct their appearance from textual descriptions, issues such as missing details and suboptimal image quality persist. To address these limitations, this study proposes a method for generating heritage buildings based on an improved diffusion model. A gated residual mechanism is introduced to optimize information flow, mitigate gradient vanishing, and enhance generation stability. A dual attention network combining channel and spatial attention is incorporated to strengthen the modeling ability of both local details and global structures. Furthermore, VGG19 is employed as a discriminant network to extract multi-level semantic features, and perceptual loss is introduced to improve the modeling effect of key visual features. Experimental results show that, compared with other diffusion-based models (KNN-diffusion and Simple diffusion), the proposed method reduces *FID* by 30.39% and improves *CLIP-score*, *IS*, and *SSIM* by 1.08%, 9.01%, and 2.35%, respectively. This study provides a feasible technical approach for generating high-quality images of heritage buildings, contributing to the sustainable research and intelligent conservation of digital cultural heritage.

① 基金项目: 国家重点研发计划 (2023YFC3803903); 西安建筑科技大学前沿交叉领域培育专项 (X20230085)
收稿时间: 2025-09-30; 修改时间: 2025-10-27; 采用时间: 2025-11-07; csa 在线出版时间: 2026-02-06

Key words: heritage building; text-to-image generation; gated residual mechanism; dual attention network; VGG19 discriminant network

1 引言

遗址类建筑物作为人类文明演进的物质载体, 承载着特定历史时期的社会形态、技术工艺与地域文化特征, 是考古学、建筑史学、文化遗产保护等领域研究的核心对象. 这类建筑不仅是历史的“活化石”, 更是地域身份认同与文化记忆延续的重要媒介, 其科学复原与数字化保存对揭示文明发展脉络、传承地域特色文化具有不可替代的学术价值与社会意义.

然而, 受自然侵蚀、人为破坏等因素影响, 大量遗址类建筑正面临不可逆的损毁甚至消失, 现存实体数量稀少且分布零散; 加之部分遗址仅存残垣断壁或基础轮廓, 传统依赖测绘图纸或局部遗存的物理重建方法, 常因信息缺失难以还原其完整形态与精细结构.

近年来, 文生图技术的快速发展为这一难题提供了新思路. 早期的文生图技术主要依赖生成对抗网络(GAN)[1]和自回归模型 (auto regressive, AR)[2], 这些模型虽然可以有效生成图像, 但会出现生成图像质量低[3]、训练易崩溃[4]等问题. 相比传统的文生图技术, 扩散模型通过迭代去噪生成高保真、语义贴合的图像. 2021年, OpenAI 团队摒弃传统的固定类别监督信号[5], 采用图文对比学习 (*CLIP*) 作为预训练目标, 通过计算文本特征和图像特征的相似性将其融合, 极大地推动了零样本学习的发展. 2022 年, Saharia 等[6]提出了 Imagen, 使用级联扩散模型, 采用多阶段生成策略 (如 64×64→256×256→1024×1024), 逐步提升分辨率, 解决高分辨率图像生成的稳定性问题, 并引入预训练的 T5-XXL 编码文本提取深层文本语义, 有效增强了文本描述与生成内容之间的语义一致性. 2022 年, Rombach 等[7]提出了稳定扩散模型, 该模型将扩散和去噪过程从像素空间转移到潜在空间中, 大幅降低了计算成本, 提高了模型训练效率. 2023 年, Ruiz 等[8]提出了 DreamBooth 方法, 通过在预训练扩散模型中引入唯一文本标识符绑定少量主题样本并进行微调, 解决了通用文生图模型无法根据用户少量图像定制特定对象细节的问题, 实现了小样本驱动的高保真个性化图像生成.

尽管现有的方法已经取得了显著成果, 但仍然存在一些关键问题: (1) 在训练过程中, 深层神经网络常常面临梯度消失或梯度爆炸的问题[9], 特别是在像 Stable diffusion 这样基于深层 U-Net 架构的扩散模型中, 反向传播的梯度需经过多级下采样与上采样层, 容易导致梯度在深层网络中过度衰减 (消失) 或放大 (爆炸), 这种现象会使得模型参数更新不稳定, 影响训练收敛性, 进而降低生成图像的质量与效率. (2) 传统的扩散模型在生成复杂的建筑物图像时, 缺乏足够的能力来捕捉图像的局部细节和全局结构[10]. 尤其是在处理具有复杂细节的遗址类建筑物时, 模型无法充分理解和表现这些细节, 使得生成图像在质量上有所欠缺. (3) 传统的像素级损失函数 (如 L1 或 L2 损失) 通过直接最小化生成图像与真实图像在像素值上的差异来优化生成模型[11], 但其本质上是一种低层次的、局部对比的监督信号, 无法有效捕捉图像的高层语义信息 (如建筑物类别、场景结构、语义关系) 和感知细节 (如纹理、边缘清晰度、色彩协调性). 这导致生成图像虽然在像素统计上可能与真实图像接近, 但在主观视觉质量、整体结构合理性以及与文本描述的语义一致性上往往存在明显差距, 生成结果模糊、细节缺失、色彩失真, 与预期视觉效果不符.

针对上述问题, 本文设计了一种基于改进扩散模型的遗址类建筑物生成方法, 该方法使用遗址类建筑物的图像信息和对应的文本描述作为模型的输入信息, 通过在 U-Net 的残差块中引入门控机制, 提升模型的稳定性和训练效果, 并通过加入双重注意力机制, 帮助模型更好地理解图像局部细节和全部结构, 同时结合 VGG19 特征提取网络作为判别器, 引导模型调整生成方向, 提升模型的泛化能力. 本研究采用对比学习法, 通过迁移学习将预训练的 Stable diffusion 模型权重初始化到遗址类建筑物生成任务中, 从而加速模型在该任务上的学习过程, 然后使用少量数据对模型微调, 进一步提升模型的泛化能力, 本文的主要贡献如下.

- 在 U-Net 的每个残差块中加入门控残差机制, 使网络自适应地调整每个残差块的贡献, 不仅能有效避免梯度在深层网络中的过度衰减或放大, 提升训练稳定性, 还能根据输入内容动态选择关键特征 (如细节、全局结构或语义相关通道), 增强模型对重要信息

的建模能力.

● 在噪声估计网络中嵌入双重注意力网络, 其中空间注意力能够帮助模型聚焦于图像的关键区域, 增强图像的局部细节. 通道注意力使模型能够更加关注图像中的关键特征通道, 增强图像的全局语义表示能力. 通过结合这两种注意力机制, 使模型能够更精准地聚焦图像关键区域并捕捉不同通道间的关联关系, 进一步提升图像的生成质量.

● 结合 VGG19 特征提取网络作为判别器, 提取生成图像与真实图像的深层语义特征, 并通过计算感知损失引导模型不断调整生成结果, 提高模型的灵活性和控制能力, 使得整个模型从多角度、多维度加深对文本信息的理解.

结果表明, 与目前主流模型相比, 本文模型在定性和定量指标上有着较大的提高, 在生成遗址类建筑物时, 本文模型呈现出更好的视觉一致性.

2 本文方法

如图 1 所示, 本文所提出的方法由跨模态语义对齐模块、生成模块以及感知优化模块这 3 部分构成.

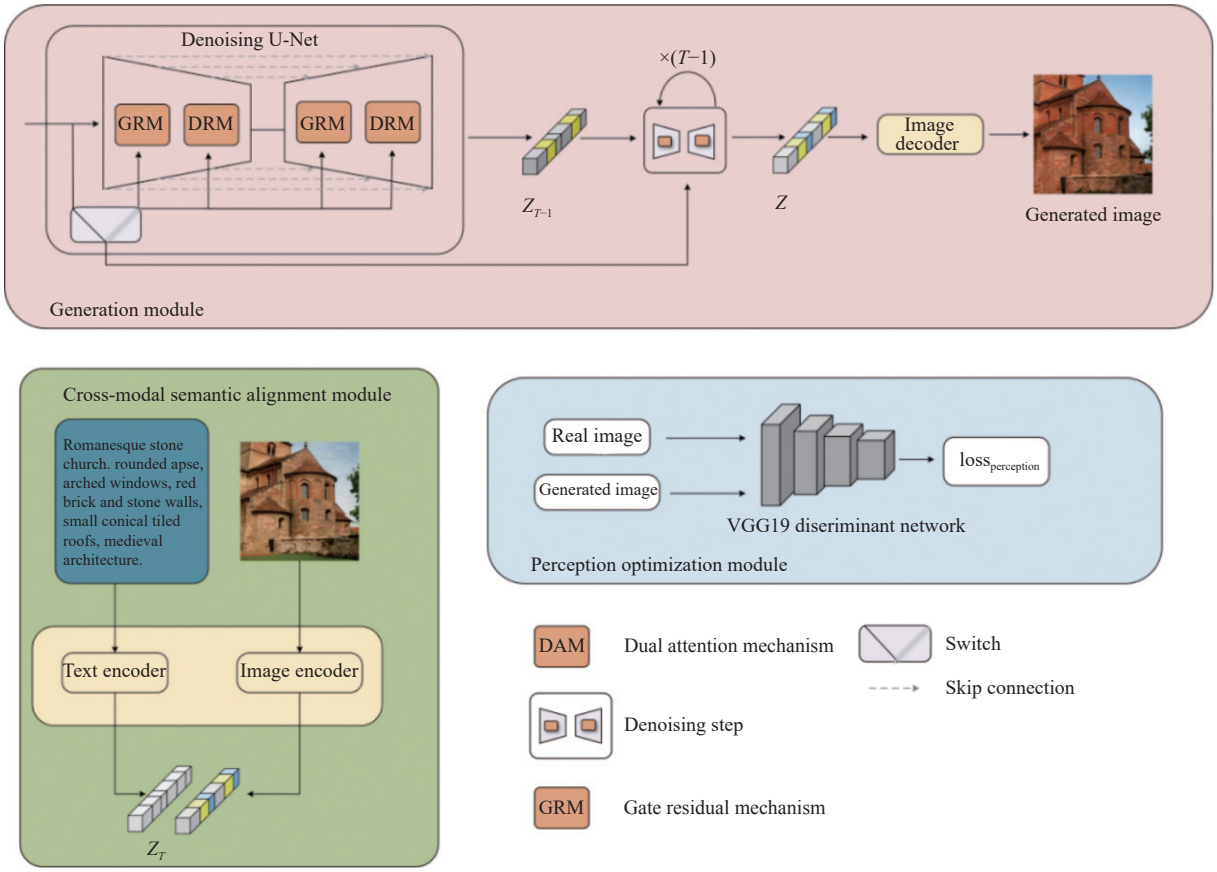


图 1 模型总体架构

在跨模态语义对齐模块中, 训练阶段使用大量真实建筑图像及其对应的文本描述, 通过 *CLIP* 预训练模型[12]实现图文语义空间对齐. *CLIP* 模型由文本编码器与图像编码器组成, 前者将文本转换为固定维度的语义向量, 后者提取建筑图像的视觉特征并映射为向量表示. 通过对比学习, 模型最大化相关图文对之间的相似度, 最小化无关图文对之间的相似度, 从而实现跨模态特征在统一语义空间中的对齐. 所得图文语义向量作为语义引导, 输入至 U-Net 网络, 引导后续图像生成过程. 与此同时, 原始图像也以像素形式输入生成模块, 在其训练阶段逐步加入高斯噪声, 构建正向扩散过程的马尔可夫链路径, 为模型学习反向去噪重建提供监督.

生成模块在图像合成中发挥核心作用, 其主要机制在于融合跨模态语义特征, 引导模型通过多轮反向扩散迭代, 从随机噪声中逐步还原出与输入文本语义一致的建筑图像. 该模块在 U-Net 结构中嵌入门控残差

机制 (GRM), 以增强网络对关键建筑特征的关注力并抑制背景干扰, 从而稳定训练并提升生成图像在细节及结构复杂性方面的表现[13,14]. 此外, 模块引入双重注意力机制 (DAM): 通道注意力提升语义显著通道特征的表达能力[15], 而空间注意力则引导模型聚焦于图像中具有关键语义的区域[16], 二者协同作用, 使模型在生成过程中兼顾建筑的局部构件与整体形态[17]. 图像生成通过多步去噪过程逐渐提升图像质量, 最终生成结构丰富、细节清晰且语义一致的高质量建筑图像.

在感知优化模块中, 所生成的图像与对应的真实图像共同输入至作为辅助判别器的 VGG19 网络中, VGG19 通过深层卷积网络提取图像的高级语义特征, 并基于感知损失计算生成图像与真实图像之间的语义距离. 感知损失从高层语义角度评估图像相似性, 弥补传统像素级损失对结构与内容把握的不足, 提升生成质量评估的感知一致性. 基于 VGG19 的反馈, 模型不断调整生成策略, 优化图像的自然度与细腻度, 从而提升遗址类建筑物的整体生成质量.

2.1 门控残差机制 GRM

在生成建筑物的过程中, 模型需要从噪声图像中恢复出复杂的图像细节, 比如某些建筑物的雕刻花纹、窗户的形状、门的设计等, 而传统残差网络在训练过程中由于会遭遇梯度消失或爆炸问题, 导致模型难以捕捉到细节信息, 因此本研究引入门控残差机制, 通过门控信号灵活调整每一层的激活信息, 防止信息丢失, 确保网络能够有效训练并捕捉到建筑物图像中的重要细节.

门控残差机制通过在每个最小残差单元中引入可学习的门控权重α和β, 来调整每一层的残差影响, 减小网络加深时可能出现的数值误差. 图 2(a) 中展示了最小残差堆叠单元, 在每一层中, 该单元通过残差连接和可学习的门控参数来调整残差的贡献; 图 2(b) 中展示了添加了门控残差机制的 U-Net 网络, 其中每一层都应用门控残差机制, 修正传播误差, 并控制模型中的均值和方差, 动态调整残差信号的贡献, 增强生成模型中的去噪能力, 改善网络的稳定性和生成质量.

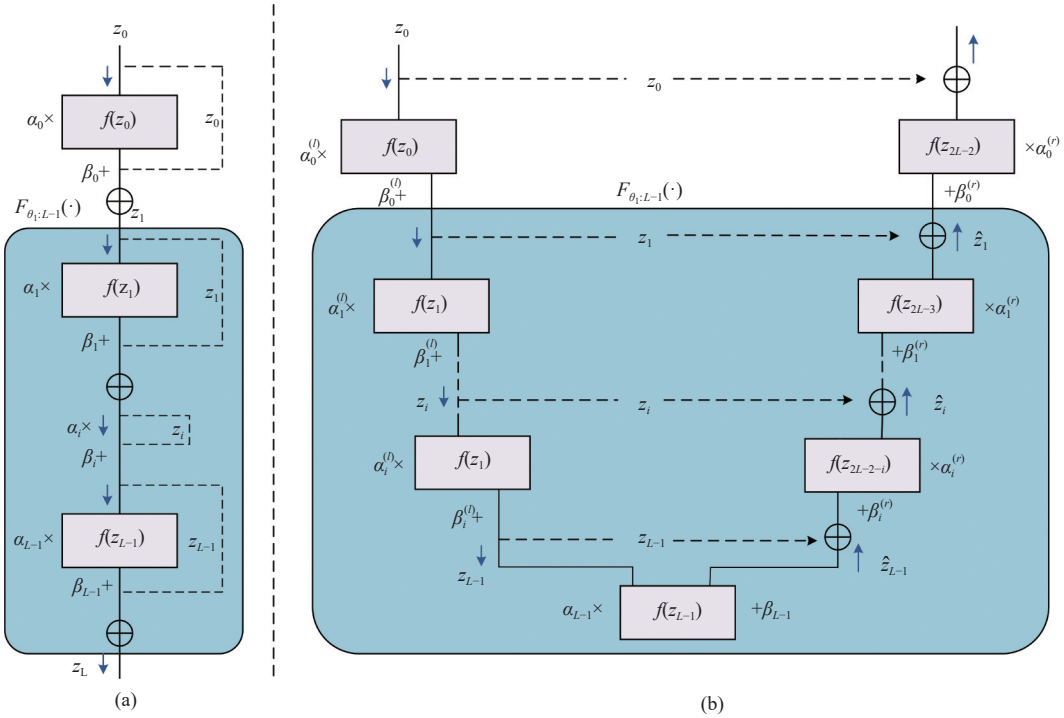


图 2 门控残差机制

假设F_{θ_i} 表示第 i 个最小残差单元块 (图 2(a)), $f(\cdot)$ 表示F_{θ_i}中的任意特征映射器, 本文不再通过普通的神经变化$\hat{z}=f_\theta(z)$传递信号 z, 而是引入一种基于门控的残差连接来处理 z, 该连接依赖两个可学习的权重$\hat{\alpha}$和$\hat{\beta}$, 用于调节非平凡变换$F_{\theta_i}(z_i)$, 具体如下:

$$\hat{z}_i = z_i + \hat{\alpha}_i \cdot F_{\theta_i}(z_i) + \hat{\beta}_i \tag{1}$$

式 (2) 展示了 U-Net 添加门控残差机制后的工作流程, 其中每个最小残差单元包含两个对称分支, 其中

左侧分支接收前一个残差单元左分支的输出 z_i 作为输入, 称为读入分支, 而右分支则继续执行用于读出的非线性变换, 称为读出分支, 便于在不同分支之间传递特征信息, 这种结构使得网络能够在每一层处理来自不同分支的信息, 增强了对不同特征的表达能力, 特别是在图像生成任务中, 能够提高细节的恢复能力和图像质量.

$$\begin{aligned}\hat{z}_i &= \alpha_i^{l}\cdot f_{\theta_i^{(l)}}(z_i)+\beta_i^{(l)}\\ &\to z_i+\alpha_i^{(r)}\cdot f_{\theta_i^{(r)}}(z_{2L-2-i})+\beta_i^{(r)}\\ &= z_i+\hat{\alpha}_i\cdot F_{\theta_i}(z_i)+\hat{\beta}_i\end{aligned} \tag{2}$$

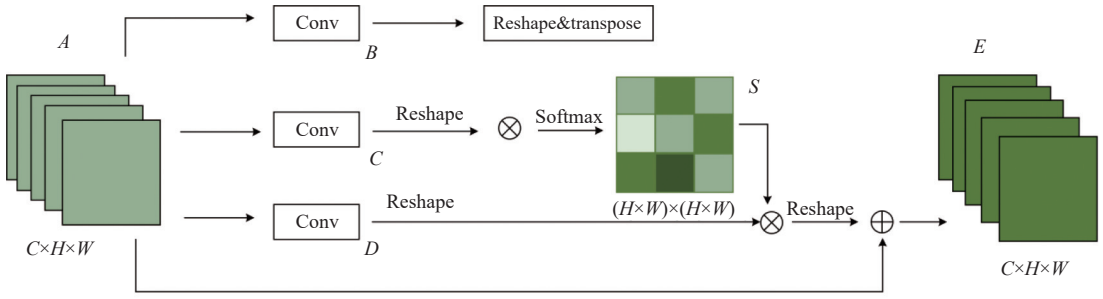
其中, $\hat{\alpha}_i$和$\hat{\beta}_i$共同表示来自左侧和右侧分支的门控权重, F_{θ_i}是 U-Net 中第 i 个最小残差单元, $\to$表示跳跃连接, 用于将z_{i+1}递归传递至z_{2L-2-i}, 该连接通过递归计算得到.

2.2 双重注意力机制 DAM

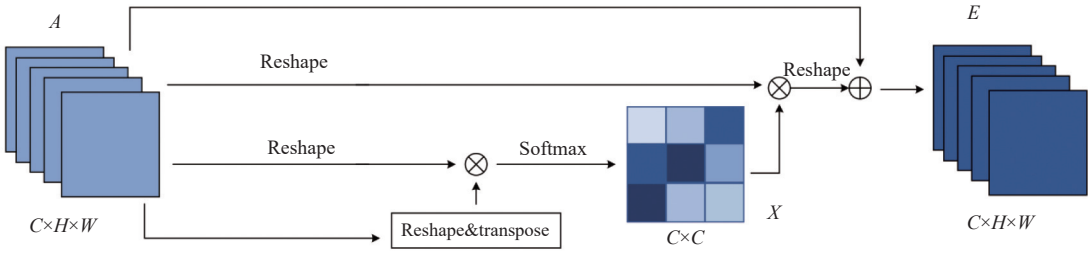
由于建筑物的生成要求高质量的视觉效果, 为了进一步增强模型对建筑物细节的刻画, 优化图像的全局和局部特征, 本研究引入双重注意力模块, 空间注意力模块 (PAM) 和通道注意力模块 (CAM) 分别从空间和通道维度对图像特征进行加权, 帮助模型更好地关注图像中的关键区域和通道, 提升模型的特征建模能力和生成图像质量. 如图 3 所示.

空间注意力模块旨在通过学习特征图上不同空间位置的重要性权重, 对特征图进行空间维度的加权, 从而突出关键区域、抑制次要区域, 增强模型对重要空间信息的感知与表达能力. 如图 3(a) 所示, 对于局部特征 $A\in R^{C\times H\times W}$ 来说, 首先将其送入卷积层分别生成两个新的特征映射 B 和 C, 其中$\{B, C\}\in R^{C\times H\times W}$, 然后将这两个特征重塑为 $R^{C\times N}$, 其中像素数 $N=H\times W$, 之后对B^{T} 和 C 做矩阵乘法, 并通过 Softmax 层计算空间注意力图 $S\in R^{N\times N}$.

$$s_{ji}=\frac{\exp(B_i\cdot C_j)}{\sum_{i=1}^{N}\exp(B_i\cdot C_j)} \tag{3}$$



(a) Spatial attention module



(b) Channel attention module

图 3 双重注意力机制

s_{ji}衡量第 i 个位置对第 j 个位置的影响, 两个位置的特征表示越相似, 它们之间的相关性就越大. 同时将原始特征 A 送入卷积层生成特征图 $D\in R^{C\times H\times W}$ 并将其重塑为 $R^{C\times N}$, 然后对特征图 D 和S^{T} 做矩阵乘法并将结果重塑为 $R^{C\times H\times W}$, 最后将其与尺度参数 α 相乘, 并与特征 A 逐个元素相加得到最终输出, 如式 (4) 所示:

$$E_j = \alpha \sum_{i=1}^{N} (s_{ji} D_i) + A_j \tag{4}$$

其中, α初始化为 0 并逐渐学习更多权重, 从式 (4) 可以看出来, 最终得到的 E 特征是所有位置上的特征和原始特征的加权和, 因此其融合了全局上下文感知能力, 使模型能够更加关注图像中的重要空间区域, 提升对关键视觉信息的建模能力.

通道注意力模块通过学习输入特征图中各通道的重要性权重, 对通道特征进行加权调整, 从而增强模型对关键语义特征的响应能力, 提升特征的语义表达能力. 具体过程与位置注意力模块相似, 如图 3(b) 所示, 不同的是通道注意力模块直接从原始特征 $A \in R^{C\times H\times W}$ 计算出通道注意力图 $X \in R^{C\times C}$. 具体来说, 首先将 A 重塑为 $R^{C\times N}$, 然后将 A 与 A^{T} 做矩阵乘法, 最后通过 Softmax 层得到通道注意力图.

$$x_{ji} = \frac{\exp(A_i \cdot A_j)}{\sum_{i=1}^{C} (x_{ji} A_i) + A_j} \tag{5}$$

其中, x_{ji} 衡量第 i 个通道对第 j 个通道的影响. 此外, 对 X 和 A^{T} 做矩阵乘法并将其结果重塑为 $R^{C\times H\times W}$, 最后将结果与比例参数 β 相乘并与 A 逐个元素相加得到最终输出 $E \in R^{C\times H\times W}$:

$$E_j = \beta \sum_{i=1}^{C} (x_{ji} A_i) + A_j \tag{6}$$

其中, β从 0 开始逐渐学习权重, 式 (6) 显示最终得到的特征图 E 是原始特征图中每个通道特征与其对应通道重要性权重的乘积, 使模型能够更加关注图像中的关键特征通道, 增强图像的全局语义表示能力.

2.3 判别网络 VGG19

为了使模型可以更好地理解建筑物图像的高层次语义信息, 提升生成图像的视觉质量和细节表现, 本文以 VGG19 网络的卷积层 (即特征提取模块) 作为判别器, 利用其强大的特征提取能力, 指导扩散模型不断调整生成结果, 使其更加符合真实建筑物的视觉特征.

VGG19 网络由 16 个卷积层、3 个全连接层以及 5 个池化层组成[18]. 本文在计算感知损失时, 仅使用其卷积层部分. 如图 4 所示, VGG19 的特征提取模块包含 5 个子模块, 每个子模块由若干卷积层 (通常为 2–3 个) 与 1 个最大池化层堆叠组成. 在特征提取过程中, 卷积操作通过滑动窗口计算[19] (通常采用 stride=1 和 padding=1 的配置), 从输入图像的多个空间位置提取局部特征, 并将这些特征整合到同一张特征图中. 在 VGG19 的卷积层中, 为维持特征图的空间维度, 卷积操作采用‘same’填充 (padding), 并将步幅 (stride) 设为 1, 使得输出特征图的高度和宽度与输入保持一致, 通道数则与所使用的卷积核数量相等. 通过卷积提取的特征包含从底层 (如边缘、纹理等基础视觉信息) 到顶层 (如物体部件、结构等高层语义信息) 的多层次语义内容. 每个子模块中的最大池化层通常采用 2×2 的窗口和步长为 2 的设置, 将特征图的空间尺寸 (高和宽) 缩减为原始大小的一半, 在降低计算量的同时保留关键特征信息, 从而逐步实现特征的多尺度抽象.

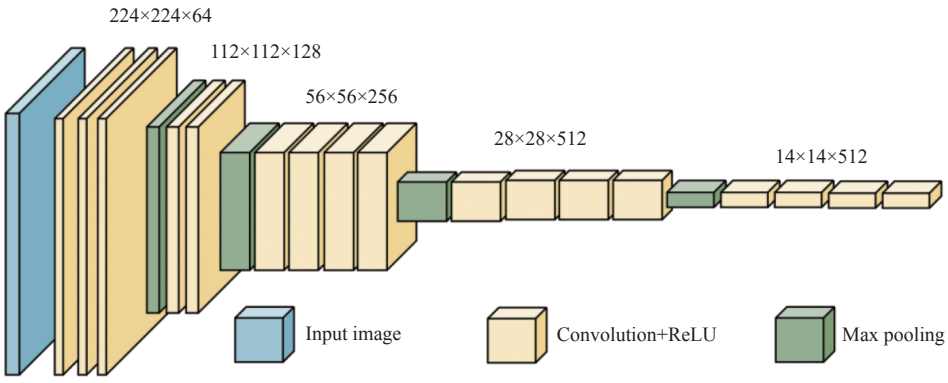


图 4 VGG19 判别网络

选择 VGG19 进行图像深层语义特征提取的关键原因在于, VGG19 相较于其他特征提取网络, 将原本采用的 5×5 和 7×7 大尺寸卷积核替换为 3×3 的小卷积核[20]. 这一改进使得 VGG19 在保持与 5×5 或 7×7 大卷积核相同感受野的前提下, 通过堆叠多个 3×3 小卷积核 (而非直接使用大卷积核) 来增加网络深度. 小卷

积核能更聚焦于图像局部区域的特征提取, 并通过分层堆叠逐步捕捉从底层边缘/纹理到高层语义的复合特征, 从而在提升网络特征提取能力的同时, 有效减少了因大卷积核参数冗余或计算平滑效应导致的局部细节信息损失.

2.4 损失函数

对于该建筑物生成任务, 模型采用联合损失进行训练, 包括去噪损失和感知损失, 以获得语义一致和视觉逼真的结果.

去噪损失主要来自反向去噪过程, 通过训练 U-Net 模型来预测在不同的时间步骤下图像的噪声部分, 计算预测噪声与真实噪声之间的误差. 具体来说, 在训练时, 给定真实图像 x_0, 首先通过 VAE 编码器将 x_0 压缩到潜在空间, 得到潜变量 z_0, 之后按照前向扩散过程对 z_0 逐步加噪, 得到噪声图像 x_t, 再训练 U-Net 模型预测噪声, 得到预测噪声与真实噪声之间的均方误差, 其定义为:

$$L_{\text{denosie}} = E_q[\|\hat{\varepsilon}_\theta(x_t,t)-\varepsilon\|^2] \tag{7}$$

其中, x_t 是第 t 步添加噪声后的图像, $\hat{\varepsilon}_\theta(x_t,t)$ 是模型预测的噪声, ε 是真实噪声, $\|\cdot\|^2$ 是均方误差 (MSE) 损失, 用来衡量预测噪声和真实噪声之间的差异.

为了量化生成图像和真实图像在高层语义特征层面的具体差异, 本文引入感知损失, 弥补了传统像素级损失的局限性, 它表示在预训练的 VGG19 模型上定义的 I_{out} 和 I_{real} 之间的 L2 范数, 其定义为:

$$L_{\text{perception}} = \sum_{i=1}^{P} \frac{\|\phi_i(I_{\text{out}})-\phi_i(I_{\text{real}})\|_2^2}{N_i} \tag{8}$$

其中, ϕ_i 表示 VGG19 第 i 层激活映射的特征图, N_i 为 $\phi_i(I_{\text{real}})$ 中的元素个数, I_{out} 表示生成图像, I_{real} 表示真实图像.

综上所述, 模型总损失是上述所有损失的综合, 其定义为:

$$L_{\text{loss}} = \lambda_1 L_{\text{denoise}} + \lambda_2 L_{\text{perception}} \tag{9}$$

在实验中, 根据经验设置 $\lambda_1=6$, $\lambda_2=0.05$, 以平衡各个损失函数的贡献.

3 实验和结果分析

3.1 数据集与实验环境

为了保证更改后模型的有效性, 本研究先在 CUB-bird 数据集上进行了实验, 之后在自有遗址类建筑物数据集进行实验, 该数据集为经过筛选后的全球范围内遗址类建筑物, 总共筛选出 5 000 张符合遗址特征的建筑物图像, 并按照 6:2:2 的比例划分为训练集、验证集和测试集, 每组数据均由建筑物图像和其对应的文本描述构成.

本研究所使用的实验操作系统为 Ubuntu, 显卡型号为 NVIDIA RTX 3090, 使用 Python 3.11.7, PyTorch 2.8.0 深度学习框架, CUDA 11.8 版本, 基于扩散模型采用 stable-diffusion-v1.5.

3.2 定性评价

为了验证本文提出算法生成图像的有效性, 本研究选用了 4 种有代表性的文生图算法在鸟类数据集和建筑物数据集上进行对比实验, 包括基于扩散模型的 KNN-diffusion[21]、Simple diffusion[22], 基于自回归模型的 CogView2[23]和基于生成对抗网络的 AttnGAN[24], 实验过程中, 所有模型的输入数据、文本条件都相同, 实验结果如图 5、图 6 所示.

从图 5 中可以看出, KNN-diffusion 生成的鸟类整体符合文本描述, 但对于鸟的脚这种细节方面生成模糊, Simple diffusion 模型没有生成鸟的尾巴, 并且没有展现出来“站在树枝”这样的描述, CogView2 生成图像质量较低, 生成图中有斑影影响图像观感, AttnGAN 生成的图像整体较模糊, 不容易看出细节, 而使用本文模型生成的图像在整体细节、图像质量上都表现较好.

从图 6 可以看出, KNN-diffusion 生成图像存在细节问题, 如第 1 行第 2 列, 石头地面的亭子上长出了柳树, 树根细小与实际明显不符, 并且细节处的光影做得不好; 第 5 行第 2 列天坛的蓝色顶部没有展现出来, 且生成的前景比例过大. Simple diffusion 生成的图像也缺少细粒度控制, 如第 1 行第 3 列中的亭子怪异, 且天空中没有展现出夕阳; 第 3 行第 3 列中教堂上花纹凌乱, 且整体图像意境不符合文本描述. CogView2 生成的图像普遍质量较低, 如第 1 行第 4 列生成的柱子明显歪曲, 不符合常理; 第 4 行第 4 列图像整体颜色明显过深, 与其他图像形成鲜明对比; 第 5 行第 4 列生成的天坛过于平滑. AttnGAN 生成图像质量不稳定, 可以明显看出第 1 行第 5 列天空中出现了斑影, 第 5 行第 5 列天坛周围出现了没有提及的柱子. 而通过改进后的扩散模型生成的图像在图像质量、细节处理、光影效果和语义一致性方面都优于对比算法.

图 5 不同模型对鸟类生成结果对比

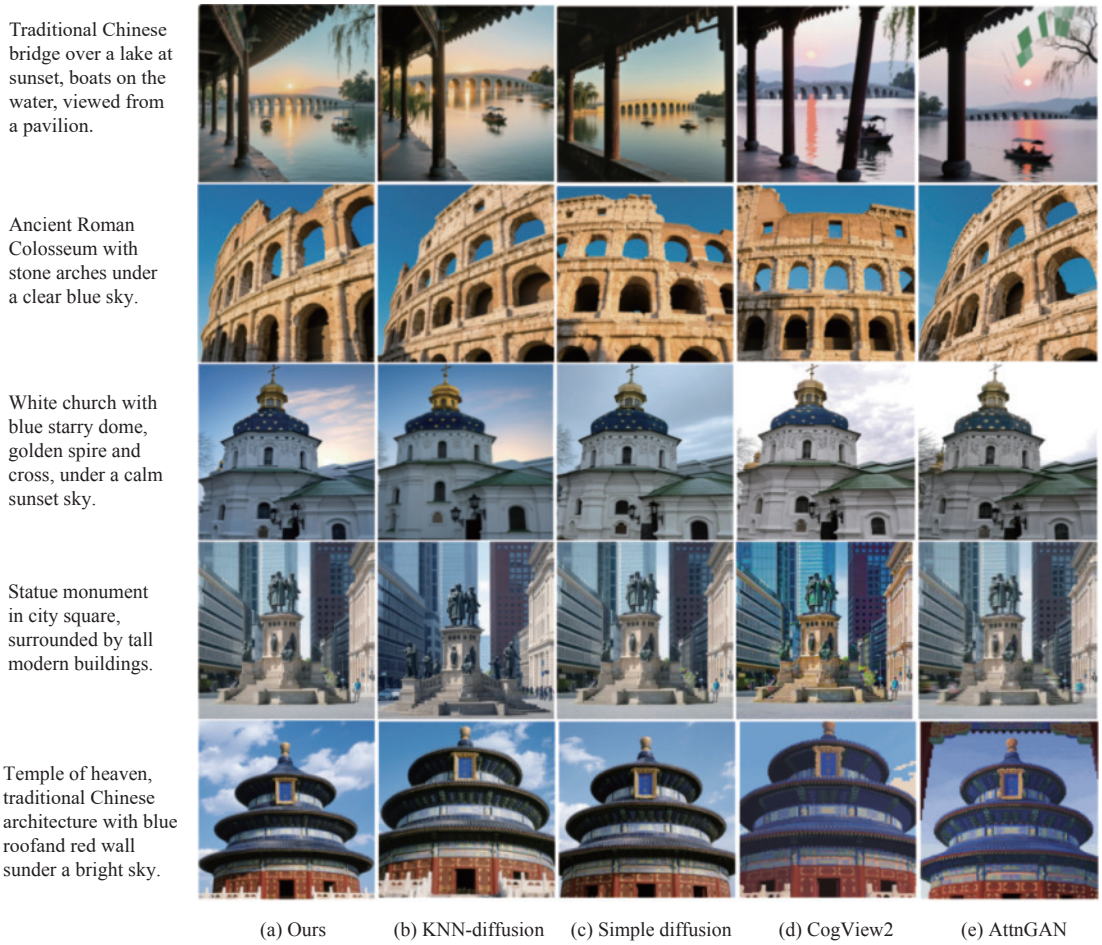


图 6 不同模型对建筑物生成结果对比

3.3 定量评价

为了进一步客观分析本研究中改进的扩散模型生成图像的质量, 本文分别采用评价指标费雷彻特初始距离 (Frechet inception distance, *FID*)、*CLIP* 分数 (*CLIP*-

score)、初始分数 (inception score, *IS*) 和结构性相似指数 (structural similarity, *SSIM*) 来比较生成图像与真实图像和文本描述的相似度.

FID 计算的是真实图像和生成图像特征分布之间的距离, 使用预训练的 Inception V3 网络从真实图像和生成图像中提取特征, 分别计算它们在特征空间中的均值和协方差矩阵, 进而通过比较分布差异来评估生成质量. 通常情况下, *FID* 越低, 说明生成图像与真实图像越接近, 生成图像质量越好, 其计算公式为:

$$FID = \left\|\mu_r \mu_g\right\|^2 + \mathrm{tr}(\sigma_r + \sigma_g - 2(\sigma_r \sigma_g)^{\frac{1}{2}}) \tag{10}$$

其中, μ_r、σ_r 分别为真实图像特征的均值和协方差, μ_r、σ_r 分别为生成图像特征的均值和协方差.

CLIP 分数用来评估图像和文本之间的相似度, 使用 *CLIP* 模型将目标图像和文本描述转化为对应的特征向量, 通过计算特征向量之间的余弦相似度衡量文本和图像之间的匹配度. 通常, *CLIP* 值越大, 表示文本和图像之间相似度越高, 数学表示为:

$$CLIP\text{-}score = \max(\cos(E_i, E_t), 0) \tag{11}$$

其中, E_i、E_t 分别表示图像和文本的特征向量.

IS 主要用于衡量生成模型中生成图像的质量和多样性, 其核心思想是通过评估生成图像的分类置信度和多样性衡量质量. 通常, *IS* 值越高, 生成图像的质量越高, 图像多样性越好, 计算公式如下:

$$IS = \exp(E_{x\sim P_g} D_{\mathrm{KL}}(p(y|x)p(y))) \tag{12}$$

SSIM 基于图像的亮度、对比度和结构来衡量生成图像与真实图像的相似度, 更符合人类视觉的感知特点. 通常, *SSIM* 值越大, 表示生成图像越接近真实图像, 计算公式如下:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{13}$$

根据上述评价指标, 将本文模型与 KNN-diffusion、CogView2、AttnGAN 及 Simple diffusion 模型进行对比, 结果如表 1、2 所示. 本文表格中“↑”表示对应指标数值越大, 模型性能越优 (如 *CLIP-score* 衡量图文语义一致性, 值越高表示生成图像与文本描述匹配度越高); “↓”表示对应指标数值越小, 模型性能越优 (如 *FID* 为生成图像与真实图像分布的 Frechet 距离, 值越小表示两者分布越接近, 生成质量越高).

表 1 不同模型在 CUB-bird 上的定量对比

Model	*FID* (↓)	*CLIP-score* (↑)	*IS* (↑)	*SSIM* (↑)
KNN-diffusion	11.27	0.341	3.81±0.02	0.945
CogView2	15.69	0.326	3.46±0.04	0.889
AttnGAN	18.31	0.284	3.39±0.03	0.854
Simple diffusion	10.94	0.396	3.93±0.05	0.953
Ours	**7.96**	**0.490**	**4.41±0.05**	**0.973**

表 2 不同模型在建筑物数据集上的定量对比

Model	*FID* (↓)	*CLIP-score* (↑)	*IS* (↑)	*SSIM* (↑)
KNN-diffusion	12.50	0.325	3.74±0.05	0.930
CogView2	17.05	0.319	3.27±0.04	0.872
AttnGAN	19.84	0.256	3.08±0.03	0.865
Simple diffusion	12.47	0.326	3.82±0.04	0.942
Ours	**8.69**	**0.329**	**4.12±0.05**	**0.958**

表 1 展现的是不同模型在 CUB-Bird 数据集上的定量比较结果. 可以看出, 与目前主流模型相比, 本文方法的 *FID* 平均下降了 40.60%, *CLIP-score*、*IS*、*SSIM* 分别平均提升了 47.57%、21.38%、7.11%, 各项指标均有显著改善, 证明了本文提出的方法在文本生成图像任务中的有效性.

表 2 展示的是不同模型在遗址类建筑物数据集上的定量比较结果. 由表 2 可知, 本文模型与同样基于扩散模型的 KNN-diffusion 和 Simple diffusion 相比, *FID* 分别从 12.50 和 12.47 降低到 8.69, 平均降低 30.39%; *CLIP-score* 分别从 0.325 和 0.326 上升到 0.329, 平均提高 1.08%; *IS* 分别从 3.27 和 4.08 上升到 4.12, 平均提高 9.01%; *SSIM* 分别从 0.930 和 0.954 上升到 0.958, 平均提高 1.72%, 这些结果表明, 本文方法在生成图像质量、文本-图像对齐、多样性和结构保真度等方面均取得了显著改进, 进一步证明了本文方法在遗址类建筑物生成任务中的有效性.

3.4 消融实验

为验证基础模型在加入残差块、双重注意力机制和判别网络后的有效性, 本文设计并开展了一系列消融实验. 实验设计为: (1) 使用原始扩散模型 (SD) 进行生成图像实验; (2) 使用仅添加门控残差机制的扩散模型 (SD+GRM) 进行实验; (3) 使用仅添加双重注意力模块的扩散模型 (SD+DAM) 展开实验; (4) 扩散模型仅结合 VGG19 判别器 (SD+VGG19) 进行生成实验; (5) 使用本文提出的方法 (Ours) 进行实验.

实验结果如表 3 显示, 门控残差机制的主要优势是改善信息流动和避免梯度消失, 提高训练稳定性, 但从最终生成图像质量的角度来看, 尽管相比单一扩散

模型来说其各项指标均有提升, 但提升幅度不大, 不如双重注意力网络和 VGG19 网络; 双重注意力网络通过空间注意力机制和通道注意力机制能够在细节层面和全局层面增强模型对图像内容的感知能力, 从表 3 中可以看出它对生成图像质量和细节的提升作用最大, 各项指标也有 1.50%–4.94% 的优化; VGG19 帮助模型在生成过程中更注重图像的长尾特征, 使得生成的图像更接近真实图像, 但相比双重注意力网络来说, 其提升效果略低, 改善范围在 1.28% 到 1.9% 之间; 同时添加这 3 个模块后, 各项指标改善更显著. 其中, *FID* 降低了 26.29%, *CLIP-score*、*IS* 和 *SSIM* 分别提高了 4.44%、7.01%、2.35%, 说明本文新增模块发挥了作用, 有效提升了生成图像质量.

表 3 消融实验结果表

Model	*FID* (↓)	*CLIP-score* (↑)	*IS* (↑)	*SSIM* (↑)
SD	11.79	0.315	3.85±0.03	0.936
SD+GRM	11.20	0.318	3.88±0.05	0.943
SD+DAM	9.89	0.324	4.04±0.05	0.951
SD+VGG19	10.49	0.321	3.96±0.04	0.948
Ours	**8.69**	**0.329**	**4.12±0.05**	**0.958**

4 结论

本文设计了一种基于改进扩散模型的遗址建筑物生成模型, 针对现有文本生成图像模型在复杂建筑物生成中的不足, 提出了新的解决方案. 通过引入门控残差机制, 优化了信息流动, 减轻了梯度消失问题, 从而提高了模型的生成稳定性和恢复能力; 结合双重注意力网络, 增强了模型对图像局部和全局特征的关注, 显著提升了生成图像的细节表现; 此外, 采用 VGG19 网络作为判别网络, 帮助模型更好地捕捉图像的长尾特征并提升图像质量. 实验结果表明, 所提方法与传统基于扩散模型的 KNN-diffusion 和 Simple diffusion 方法相比, *FID* 值平均降低了 30.39%, *CLIP-score*、*IS* 和 *SSIM* 指标分别提高了 1.08%、9.01% 和 2.35%, 表明本方法在图像生成质量和细节还原方面具有显著优势.

尽管如此, 生成图像与复杂文本描述之间的契合度仍有提升空间. 当前, *CLIP-score* 和 *SSIM* 分别提高了 1.08% 和 2.35%, 但模型在处理复杂文本时仍存在理解局限, 导致生成图像未能完全捕捉文本中的细节. 为进一步提高生成图像与文本的匹配度, 未来考虑引入自监督学习模块, 结合图像和文本的多维度信息进行训练, 这将有助于模型更好地理解图像和文本之间的关系, 提升生成图像的精度与质量.

参考文献

1 Goodfellow I, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial networks. Communications of the ACM, 2020, 63(11): 139–144. [doi: 10.1145/3422622]

2 Chen M, Radford A, Child R, *et al*. Generative pretraining from pixels. Proceedings of the 37th International Conference on Machine Learning. JMLR.org, 2020. 158.

3 Xu T, Zhang PC, Huang QY, *et al*. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 1316–1324.

4 Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 574.

5 Radford A, Kim JW, Hallacy C, *et al*. Learning transferable visual models from natural language supervision. Proceedings of the 38th International Conference on Machine Learning. PMLR, 2021. 8748–8763.

6 Saharia C, Chan W, Saxena S, *et al*. Photorealistic text-to-image diffusion models with deep language understanding. Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 2643.

7 Rombach R, Blattmann A, Lorenz D, *et al*. High-resolution image synthesis with latent diffusion models. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 10674–10685.

8 Ruiz N, Li YZ, Jampani V, *et al*. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 22500–22510.

9 Jiang ZY, Po LM, Xu XY, *et al*. RMP-adapter: A region-based Multiple Prompt Adapter for multi-concept customization in text-to-image diffusion model. Expert Systems with Applications, 2025, 274: 126936. [doi: 10.1016/j.eswa.2025.126936]

10 Ramírez-Juidias E, Antón D. Geospatial analysis of the Roman site of Munigua based on RGB airborne imagery.

Remote Sensing, 2025, 17(18): 3224. [doi: 10.3390/rs17183224]

11 Xu ZT, Zeng L, Zhao JL, *et al*. Sketch123: Multi-spectral channel cross attention for sketch-based 3D generation via diffusion models. Computer-Aided Design, 2025, 185: 103896. [doi: 10.1016/j.cad.2025.103896]

12 刘杰, 乔文昇, 朱佩佩, 等. 基于图像-文本大模型 CLIP 微调的零样本参考图像分割. 计算机应用研究, 2025, 42(4): 1248–1254. [doi: 10.19734/j.issn.1001-3695.2024.06.0254]

13 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

14 Khan S, Naseer M, Hayat M, *et al*. Transformers in vision: A survey. ACM Computing Surveys (CSUR), 2022, 54(S10): 200. [doi: 10.1145/3505244]

15 马天帅, 杨燕. 结合双重注意力与 Transformer 的雾天图像复原算法. 电光与控制, 2025, 32(9): 61–67.

16 Yu WB, Li Y, Yang HT, *et al*. The centerline extraction algorithm of weld line structured light stripe based on pyramid scene parsing network. IEEE Access, 2021, 9: 105144–105152. [doi: 10.1109/ACCESS.2021.3098833]

17 Ding HH, Jiang XD, Shuai B, *et al*. Semantic segmentation with context encoding and multi-path decoding. IEEE Transactions on Image Processing, 2020, 29: 3520–3533. [doi: 10.1109/TIP.2019.2962685]

18 Dey N, Zhang YD, Rajinikanth V, *et al*. Customized VGG19 architecture for pneumonia detection in chest X-rays. Pattern Recognition Letters, 2021, 143: 67–74. [doi: 10.1016/j.patrec.2020.12.010]

19 Awan MJ, Masood OA, Mohammed MA, *et al*. Image-based malware classification using VGG19 network and spatial convolutional attention. Electronics, 2021, 10(19): 2444. [doi: 10.3390/electronics10192444]

20 Karacı A. VGGCOV19-NET: Automatic detection of COVID-19 cases from X-ray images using modified VGG19 CNN architecture and YOLO algorithm. Neural Computing and Applications, 2022, 34(10): 8253–8274. [doi: 10.1007/s00521-022-06918-x]

21 Rao JH, Shan ZF, Liu LP, *et al*. Retrieval-based knowledge augmented vision language pre-training. Proceedings of the 31st ACM International Conference on Multimedia. Ottawa: ACM, 2023. 5399–5409. [doi: 10.1145/3581783.3613848]

22 Hoogeboom E, Heek J, Salimans T. Simple diffusion: End-to-end diffusion for high resolution images. Proceedings of the 40th International Conference on Machine Learning. Honolulu: PMLR, 2023. 13213–13232.

23 Ding M, Zheng WD, Hong WY, *et al*. CogView2: Faster and better text-to-image generation via hierarchical transformers. Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 1229.

24 Chen KY, Zhou ZW. Research on AttnGAN text image generation method based on CLIP enhancement and diffusion optimization. Engineering Letters, 2025, 33(9): 3801–3808.

(校对责编: 李慧鑫)