计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
[doi: 10.15888/j.cnki.csa.010078] [CSTR: 32024.14.csa.010078]

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

语义感知交互扩散图像超分辨率重建①

王　军[2,3], 杨立文[1]

[1](南京信息工程大学 计算机学院、网络空间安全学院, 南京 210044)
[2](南京信息工程大学 科技产业处, 南京 210044)
[3](南京信息工程大学 软件学院, 南京 210044)
通信作者: 杨立文, E-mail: 202412490786@nuist.edu.cn

摘　要: 针对真实世界图像超分辨率任务中图片退化类型多样与细节恢复困难的问题, 现有方法在结构保持与语义一致性方面仍存在不足. 为此, 本文提出一种语义感知交互扩散图像超分辨率重建方法 (semantic-aware interactive diffusion method for image super-resolution reconstruction, SISRM), 引入语义分割信息作为先验以增强重建过程的结构理解与语义引导. 具体而言, 该方法首先设计并训练分割感知提示提取器, 通过分割掩码编码器和标签文本生成器, 从退化低分辨率图像中高效提取分割掩码嵌入与语义标签; 其次, 引入交互式文本到图像控制器, 结合分割交叉注意力模块和可训练图像编码器, 通过多模态语义条件引导扩散过程增强局部细节与全局结构感知; 最后, 提出掩码特征融合机制缓解局部条件控制与全局潜在分布差异, 提高生成图像的一致性和视觉质量. 在 DIV2K-Val 和 RealSR 数据集上, 所提方法在无参考图像质量评估和跨模态图像质量评估最高分别达到 0.6121 和 0.7274, 感知质量提高明显, 验证了其在细节还原、语义一致性及视觉质量方面的综合优势.

关键词: 超分辨率重建; 语义感知; 扩散模型; 掩码融合; 真实世界图像

引用格式: 王军,杨立文.语义感知交互扩散图像超分辨率重建.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/10078.html

Semantic-aware Interactive Diffusion for Image Super-resolution Reconstruction

WANG Jun[2,3], YANG Li-Wen[1]

[1](School of Computer Science & School of Cyber Science and Engineering, Nanjing University of Information Science & Technology, Nanjing 210044, China)
[2](Science and Technology Industry Division, Nanjing University of Information Science & Technology, Nanjing 210044, China)
[3](School of Software, Nanjing University of Information Science & Technology, Nanjing 210044, China)

Abstract: Diverse degradation types and the difficulty of detail recovery make real-world image super-resolution challenging, with existing methods still struggling with structural preservation and semantic consistency. This study proposes a semantic-aware interactive diffusion method for image super-resolution reconstruction (SISRM) method. Semantic segmentation information is introduced as prior knowledge, enhancing structural understanding and providing semantic guidance during reconstruction. Specifically, a segmentation-aware prompt extractor is designed and trained to efficiently obtain segmentation mask embeddings and semantic labels from degraded low-resolution images using a segmentation mask encoder and a label text generator. An interactive text-to-image controller is then introduced, integrating a segmentation-guided cross-attention module with a trainable image encoder. The diffusion process is guided under multi-modal semantic conditions to enhance local detail and global structure awareness. Finally, a mask feature fusion mechanism is proposed to mitigate the mismatch between local conditional control and the global latent distribution, improving the consistency and visual quality of the generated images. Experimental results on the DIV2K-

① 基金项目: 国家自然科学基金 (41975183)
收稿时间: 2025-08-01; 修改时间: 2025-08-22, 2025-09-05; 采用时间: 2025-09-15; csa 在线出版时间: 2025-12-26

Val and RealSR datasets show that the proposed method achieves the highest scores of 0.612 1 and 0.727 4 in no-reference and cross-modal image quality assessment, respectively. These results demonstrate notable improvements in detail restoration, semantic consistency, and overall perceptual quality.

Key words: super-resolution reconstruction; semantic awareness; diffusion model; mask fusion; real-world image

早期研究通常基于简单且已知的退化假设 (如双三次下采样) 来解决图像超分辨率 (image super-resolution, ISR) 问题, 并提出多种成功的模型[1−7]. 这类方法以优化保真度为主要目标, 因此在细节恢复方面表现有限, 往往产生过度平滑的结果. 为提升视觉感知质量, 生成对抗网络 (generative adversarial network, GAN)[8]被引入 ISR 任务[9], 通过在训练过程中引入对抗损失, 使模型能够生成更加逼真的视觉细节. GAN 方法[10−13]在感知质量提升方面取得显著进展, 但由于对抗训练的不稳定性, 这类方法难以在感知质量与保真度之间取得良好平衡, 结果中常出现伪影或细节失真.

近年来, 去噪扩散概率模型 DDPM[14]在逼近复杂分布与生成高质量图像方面展现出强大能力[15]. 特别是随着大规模预训练文本到图像 (text-to-image, T2I) 模型 (如 Stable Diffusion[16]) 的兴起, 图像生成技术进入新的发展阶段. 在此背景下, 研究者逐渐探索如何借助 Stable Diffusion 的强大生成能力改进超分辨率重建性能. 在现有基于扩散的超分辨率方法中, StableSR[17]将低质量图像的潜在表示作为条件输入, 引导 Stable Diffusion 进行超分辨率重建; DiffBIR[18]则通过先行恢复低质量图像并结合生成先验, 在扩散过程中平衡重建质量与保真度. 这些方法凸显出生成先验在超分辨率任务中的潜力, 但由于仅利用低质量图像信息, 缺乏高层次语义指导, 易导致内容重建不准确.

鉴于文本提示在预训练 T2I 模型中对生成过程中所展示的引导优势, 近期相关研究[13,19−21]开始尝试通过从低质量图像中提取语义描述, 来进一步控制 Stable Diffusion 的生成过程, 以提升图像恢复的语义保真度. 例如, PASD[22]和 SeeSR[23]分别利用预训练的图像描述生成模型[24]或标签模型[25]提取图像内容信息, 作为文本提示引导生成. 然而, 这类方法仍有局限: 一方面, 预训练模型难以完整覆盖真实场景中多样化的细节与语义, 提示信息可能不够准确或遗漏关键内容; 另一方面, 纯文本提示缺乏对局部区域的细粒度控制, 易在复杂场景中产生语义不一致的结果.

针对上述问题, 本文提出方法框架 SISRM, 结合分割感知提示提取器 (segmentation-aware prompt extractor, SAPE) 与交互式文本到图像控制器 (interactive text-to-image controller, IT2IC), 实现细粒度语义引导. SISRM 方法的主要贡献如下.

1) 设计并训练提示提取器, 可直接从退化低分辨率 (low-resolution, LR) 图像中预测用于语义控制的分割掩码嵌入与标签文本, 无需显式生成像素级分割掩码, 以指导 T2I 模型生成语义保真的超分辨率图像. 该语义嵌入作为紧凑的高维特征表示, 有效弥补了传统基于全局文本提示方法中提示信息不准确或不完整的问题.

2) 基于 ControlNet[26]构建交互式控制器, 引入分割交叉注意力模块以学习精准的语义指导, 并集成可训练图像编码器处理 LR 图像特征, 在训练时仅更新新增模块, 以提高训练效率.

3) 提出掩码特征融合 (mask feature fusion, MFF) 机制, 以缓解局部条件控制与全局潜在分布之间的差异, 有效提升图像整体质量与视觉一致性.

实验结果表明, SISRM 在生成高质量图像的同时, 可有效保留语义信息, 并在多个真实世界图像超分辨率任务中取得优异性能.

1 相关工作

图像超分辨率旨在从低分辨率图像中恢复高分辨率细节, 广泛应用于遥感、安防、医疗等领域. 传统方法多采用双三次插值, 虽计算简单, 但难以有效重建细节. 随着深度学习的发展, 卷积神经网络成为超分辨率重建任务的主流方法, 代表性工作如 SRCNN、VDSR 等[1,2], 虽然可提升重建性能, 但结果过于平滑且感知质量有限. 为增强视觉感知效果, ESRGAN[9] 等引入 GAN 及感知损失, 改善纹理细节, 但训练不稳定, 容易引入伪影, 难以在保真度与感知质量间取得平衡. 近年来, 基于 Transformer 架构的方法, 如 IPT[1]、SwinIR[3] 等, 通过自注意力机制提升全局上下文建模能力, 进一步

改善细节还原和视觉一致性. 但这些方法普遍假设理想化退化过程, 面对真实世界中复杂退化场景时, 性能仍有限, 尤其在细节恢复与语义一致性方面存在不足. 为改善这类问题, 结合语义引导的扩散模型用于图像超分辨率的方法正逐渐引起广泛关注.

1.1 语义分割先验

语义分割本质上是像素级分类任务, 传统方法[27,28]多采用编码器-解码器架构, 为图像中的不同区域分配语义标签. 不同于类无关的通用分割方法, 开放词汇分割方法 (如 X-decoder[29]、SEEM[30]) 能够结合用户提供的文本输入生成分割掩码. 近年来, DDPS[31]利用离散扩散模型[32]建立分割掩码先验以提升分割效果. 由于语义分割能提供丰富的语义与空间信息, 已有研究将其作为先验引导更精准的图像复原, 例如: SFTGAN[33]恢复符合语义类别的纹理细节, SSG-RWSR[34]通过语义分割损失约束超分学习过程, SAM-DiffSR[35]以 SAM[36]生成的掩码调节扩散噪声分布以改善复原质量. 然而, 真实场景中的低质量图像往往存在复杂未知退化, 如何实现精准语义分割并充分利用分割先验提升 Real-ISR 表现, 仍是亟待解决的重要问题.

1.2 扩散概率模型

受非平衡热力学和序贯蒙特卡罗方法的启发, Sohl-Dickstein 等人[37]最早提出扩散模型, 用于建模复杂数据分布. 随后, 扩散模型被广泛应用于图像生成任务, 特别是自 DDPM[14]提出以来得到快速发展. Rombach 等人[16]进一步将 DDPM 训练迁移到潜在空间中, 显著推动了大规模预训练文图生成模型 (如 Stable Diffusion 和 Imagen) 的进步. 近年来, 研究者将扩散模型应用于图像超分辨率任务, 弥补传统方法在感知质量与细节恢复上的不足. StableSR[17]以低质量图像潜在表示为条件, DiffBIR[18]通过引入生成先验, 在视觉真实感与像素保真间取得平衡, 展现出良好的适应性与性能. 近期 PASD[22]、SeeSR[23]等工作进一步引入语义提示以提升语义一致性与细节恢复.

总的来说, 现有图像超分辨率方法在感知质量与细节恢复方面取得一定进展, 但在真实场景中仍难以同时保证细节还原和语义一致性. 现有研究尝试结合语义提示与扩散模型缓解这一问题, 但仍缺乏对局部区域的精准控制. 本文提出的语义感知交互扩散超分辨率方法, 旨在通过细粒度语义指导与交互式扩散生成, 有效提升真实场景下的重建质量与语义一致性.

2 本文方法

2.1 方法概述

本文提出 SISRM 方法的整体框架如图 1 所示, 主要有 2 个核心组件: 分割感知提示提取器 (SAPE) 和交互式文本到图像控制器 (IT2IC). 并在推理阶段引入掩码特征融合机制 (MFF). 设定低分辨率图像 I_{LR}、超分辨率重建图像 I_{SR} 为方法的输入和输出, 高分辨率图像 I_{HR} 为理想图像.

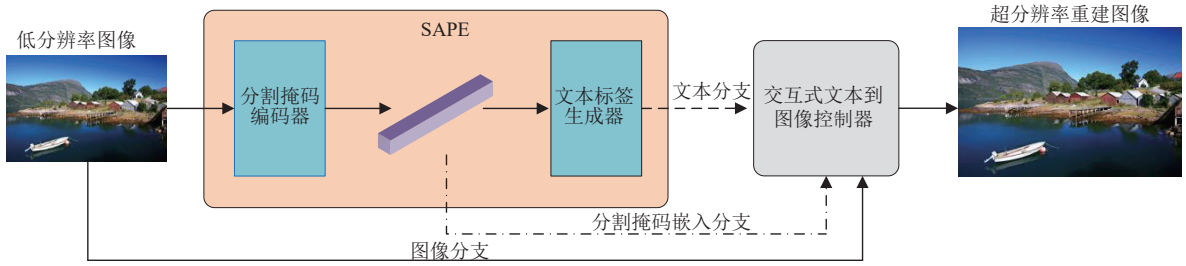

图 1 SISRM 方法的整体框架

首先利用分割掩码编码器从 I_{LR} 提取分割掩码嵌入:

$$E_m = F_{\mathrm{conv}}(F_{\mathrm{seg}}(I_{\mathrm{LR}})) \tag{1}$$

其中, F_{seg} 表示语义分割网络, F_{conv} 为卷积编码器. 由语义分割网络得到像素级掩码, 再按“局部聚合→通道重映射”的两步进行编码, 得到语义嵌入 E_m. 随后将 E_m 传入文本标签生成器生成文本标签, 文本标签生成器的输出可表示为:

$$t = G(E_m) \tag{2}$$

其中, G 表示文本标签生成器, 对 E_m 进行全局平均池化获得紧凑向量, 经全连接层得到类别分数并以 Softmax 归一化形成概率分布, 随后选取 Top-k 的类别索引组合作为标签集合 t. 利用交互式文本到图像控制器将 t

和 I_{LR} 转化为对应的特征编码 E_t 和 E_i:

$$E_t = F_{\mathrm{conv}}(t) \tag{3}$$

$$E_i = F_{\mathrm{conv}}(I_{\mathrm{LR}}) \tag{4}$$

其中, E_t 表示文本标签的特征编码, 由标签集合 t 输入到文本编码器得到. E_i 表示图像的特征编码, 由可训练图像编码器将 I_{LR} 映射到潜在空间得到. 最后将 E_m、E_t 和 E_i 进行特征融合指导扩散模型进行图像重建. 则超分辨率重建图像 I_{SR} 重构过程为:

$$I_{\mathrm{SR}} = Diffusion(F_{\mathrm{fuse}}(E_m, E_t, E_i)) \tag{5}$$

其中, F_{fuse} 表示融合机制, *Diffusion* 表示 T2I 扩散模型.

2.2 分割感知提示提取器 (SAPE)

SAPE 模块是在预训练的提示提取器 (即 DAPE)[23] 基础上进行重构而得, 旨在从退化的 LR 图像中提取高质量语义信息, 以指导预训练的 T2I 扩散模型实现语义保真的超分辨率图像生成. 其核心结构包括封装的分割掩码编码器与文本标签生成器.

如图 2 所示, 分割掩码编码器由语义分割网络 (SegFormer) [29]和卷积编码器组成: SegFormer 先从退化 LR 图像中预测像素级分割掩码. 随后该掩码通过卷积编码器进行特征编码: 先使用 3×3 卷积对局部邻域进行上下文聚合, 再通过 ReLU 激活引入非线性表示, 最后利用 1×1 卷积完成通道映射与降维, 输出分割掩码嵌入 E_m. 该嵌入保留了空间语义结构, 并与后续扩散模型的特征空间对齐, 用作局部语义提示.

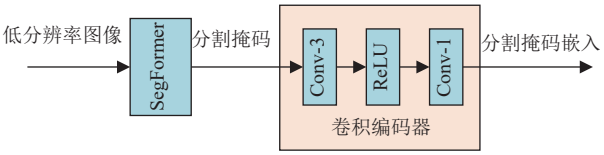

图 2 分割掩码编码器

标签生成器的具体架构如图 3 所示, E_m 作为标签生成器的输入, 首先通过全局平均池化 (GAP), 对空间维度求平均得到全局语义向量. 随后, 通过全连接层 (FC) 将全局语义向量投影至类别数维度, 得到类别得分向量. 对得分向量施加 *Softmax* 函数以获得各类别的概率分布, 并从中选取概率最高的前 k 个类别索引, 映射为对应的文本标签集合 t. 该文本提示作为全局语义条件, 与 E_m 协同用于后续扩散生成.

SAPE 利用上述双路径形成语义提示条件, 以精准引导 T2I 模型在超分辨率重建过程中生成视觉上逼真且语义一致的图像.

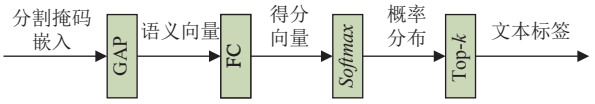

图 3 文本标签生成器

在训练阶段, 如图 4 所示, 对输入的 I_{HR} 图像 x 进行随机退化操作生成 I_{LR} 图像 y. 将 I_{HR} 图像 x 输入到冻结的提取器中生成分割掩码嵌入 f_x^{mask} 和标签预测得分嵌入 f_x^{logits}. 将 I_{LR} 图像 y 输入到可训练的提取器中生成分割掩码嵌入 f_y^{mask} 和标签预测得分嵌入 f_y^{logits}. 以 I_{HR} 图像 x 输出的分割掩码嵌入 f_x^{mask} 和标签预测得分嵌入 f_x^{logits} 作为锚点来监督 SPAE 的训练. 为使 SAPE 对图像退化更具鲁棒性, 强制 L_r 分支的分割掩码嵌入和标签预测得分嵌入接近 I_{HR} 分支的分割掩码嵌入和标签预测得分嵌入. 则训练 SAPE 的损失函数如下:

$$L_{\mathrm{SAPE}} = L_r(f_y^{\mathrm{mask}}, f_x^{\mathrm{mask}}) + \lambda L_l(f_y^{\mathrm{logits}}, f_x^{\mathrm{logits}}) \tag{6}$$

其中, L_r 是均方误差 (MSE) 损失, L_l 是交叉熵损失[38], λ 是平衡参数. 通过对齐 I_{HR} 和 I_{LR} 分支的分割掩码嵌入和标签预测得分嵌入, SAPE 模块能够学习从退化的低分辨率图像中提取高质量的分割掩码嵌入和标签文本, 从而为后续的图像生成提供准确的语义信息.

在推理阶段, SAPE 模块用于提取分割掩码嵌入和标签文本, 以引导预训练的 T2I 模型生成语义保真的超分辨率图像. 标签文本被传递至 T2I 模型中冻结的文本编码器, 以增强其局部理解能力. 文本提示的数量由预设阈值控制: 当阈值较高时, 虽可提升类别预测的准确性, 但可能导致召回率下降; 而阈值较低时, 则可能出现相反的情况. 为克服这一限制, 引入分割掩码嵌入, 该方法无需依赖阈值设定, 同时避免传统独热编码类别信息熵偏低的问题[39].

2.3 交互式文本到图像控制器 (IT2IC)

图 5 为 IT2IC 的详细结构, 本文将 ControlNet[26] 作为 T2I 模型的控制器, 用于实现超分辨率重建.

本方法将预训练 SD 模型中 U-Net 的编码器克隆为可训练的副本, 用以初始化 ControlNet. 为将分割掩码嵌入融入扩散过程, 采用 PASD[22]中所提出的交叉注意力机制来学习语义指导. 在 U-Net 中添加分割交叉注意力 (segmentation cross-attention, SCA) 模块, 并将其放置于文本交叉注意力 (text cross-attention, TCA)

模块之后. 需要注意的是, 随机初始化的 SCA 模块与编码器同时进行克隆. 除文本分支和分割掩码嵌入分支, 图像分支也在重建目标 HR 图像中发挥作用. 通过可训练的图像编码器处理 LR 图像, 以获得其潜在表示, 并将其输入至 ControlNet 中. 该可训练图像编码器的结构与文献[26]中保持一致.

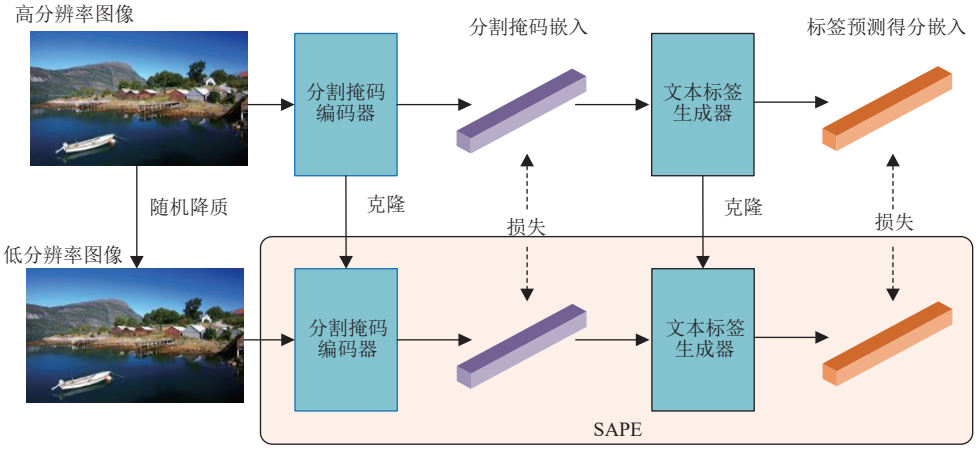

图 4 SAPE 训练过程

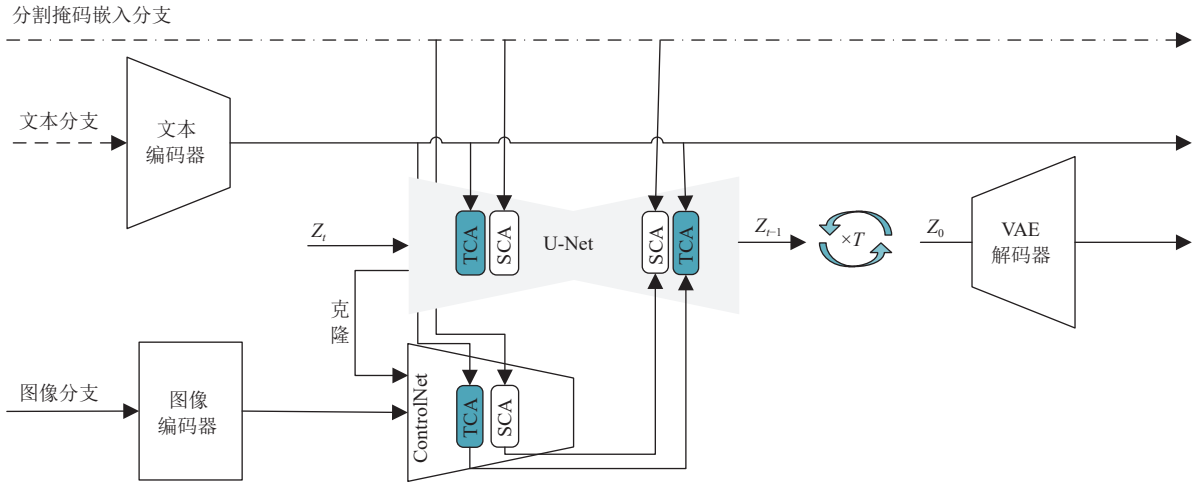

图 5 IT2IC 的内部结构

具体实现方面, IT2IC 的输入包括: (1) 扩散过程中的潜变量 Z_t; (2) SD 文本编码器输出的文本特征 E_t; (3) SAPE 提取的分割掩码 E_m, 经线性投影后形成 L_m 个“语义锚点” K_m、V_m; (4) 可训练图像编码器提取并与 ControlNet 对齐的低分辨率潜变量 E_i. 在每个跨注意力层首先计算文本交叉注意力:

$$Attn_t = Softmax\left(\frac{QK_t^{\mathrm{T}}}{\sqrt{d}}\right)V_t \tag{7}$$

其中, Q 来自当前 U-Net 层特征, K_t、V_t 为文本提示键值对; 随后在叠加文本注意力的基础上计算分割交叉注意力:

$$Attn_m = Softmax\left(\frac{(Q+Attn_t)K_m^{\mathrm{T}}}{\sqrt{d}}\right)V_m \tag{8}$$

最终层输出为 $h = Q + Attn_t + Attn_m$ 并加上对应尺度的 ContronNet 特征进行融合. 通过此设计, 模型能在保持全局语义一致性的同时, 对局部区域进行精细重建.

IT2IC 模型的训练过程如下: 首先, 通过预训练 VAE[16]的编码器获取高分辨率图像的潜在表示, 记作 Z_0. 随后, 在扩散过程中逐步向 Z_0 添加噪声, 得到带噪声的潜在表示 Z_t, 其中 t 表示随机采样的扩散步长. 利用扩散步长 t、LR 图像的潜在表示 Z_{lr}、标签文本提示 P_t 以及分割掩码嵌入提示 P_m 训练 IT2IC 网络 (记作 ε_θ), 目标为使该网络能预测出当前带噪潜变量 Z_t 中添加的噪声. 优化目标如下:

$$L = E_{Z_0,Z_{\mathrm{lr}},P_t,P_m,\varepsilon\sim N}\left[\| \varepsilon - \varepsilon_\theta(Z_t,Z_{\mathrm{lr}},t,P_t,P_m)\|_2^2\right] \tag{9}$$

为提高训练效率, 保持 Stable Diffusion 主体模型参数不变, 仅训练新加入的图像编码器、ControlNet 和 U-Net 中的 SCA 模块.

2.4 掩码特征融合机制 (MFF)

在去噪过程中, 局部控制只携带局部控制信息, 其余区域为空白控制信息, 控制区域内外存在显著的差异. 因此, 局部控制条件下得到的潜在分布与全局控制条件下得到的潜在分布不同, 可能影响图像质量. 为解决这个问题, 从 ControlNet 模块的输出中提取局部特征映射, 并将局部特征加入至 U-Net 中. 假设 U-Net 架构中的神经块为$F(\cdot:\theta)$, x_t 表示第 t 次去噪中的输入特征, 空间条件输入特征为 c_f, ControlNet 的架构中的神经块为$G(\cdot:\cdot)$, 每个块输出的融合特征表示为 y_t:

$$y_t = F(x_t:\theta) + M \cdot G(x_t:c_f) \tag{10}$$

其中, $F(x_t:\theta)$表示由主干 U-Net 基于全局潜在分布生成的重建特征, $G(x_t:c_f)$表示由 ControlNet 分支提取并映射的局部细粒度特征, 其中 $M\in\{0,1\}$表示从局部控制区域生成的二值空间掩码. 当 M_{ij}=1 时, 对应位置会叠加 ControlNet 的局部特征以增强该区域的语义和纹理细节; 当 M_{ij}=0 时, 该位置保持主干输出 $F(x)$ 的全局生成结果. 通过这种空间位置选择, 模型能够在掩码覆盖的局部区域注入额外的高频信息, 而在非控制区域保持全局一致性, 从而有效缓解局部控制与整体潜在分布冲突.

与直接全图拼接或简单加权不同, 这种掩码加性融合方式明确分离了受控区域和非受控区域, 提升了细节注入的可控性和稳定性, 并避免了全局范围内无关特征叠加所带来的伪影.

3 实验结果与分析

3.1 数据集

本文采用 DIV2K[40]和 RealSR[41]两个数据集来验证所提方法的有效性. DIV2K 数据集包含 1 000 张高质量图像, 其中 800 张用于训练, 100 张用于验证, 100 张用于测试. 该数据集采用标准的双三次下采样方式生成低分辨率图像, 并引入多种复杂退化类型, 真实地模拟实际应用中的图像退化, 为超分辨率算法的训练与评估提供了丰富支持. RealSR 是一个面向真实世界图像超分辨率的数据集, 旨在解决现有模型在合成数据与真实场景之间存在的性能差异问题. 该数据集通过调整数码相机焦距拍摄同一场景的低分辨率 (LR) 和高分辨率 (HR) 图像对, 并提出专门的图像配准算法进行对齐, 以支持端到端训练. 数据集包含使用佳能 5D3 和尼康 D810 相机拍摄的丰富多样的场景图像, 涵盖建筑、风景、动植物等类别, 且纹理细节丰富清晰.

3.2 实验设置

采用 LoRA 方法[42] (r=8), 对来自 DAPE[23]的整个 SAPE 模块进行 10k 次迭代的微调. 批量大小设置为 32, 学习率设置为 10^{-4}. 对 DIV2K 数据集采取与 Real-ESRGAN [11] 相同的退化处理流程以及与 SeeSR [23]相同的退化设置来合成 LQ-HQ 训练对, 用来训练 IT2IC. 整个训练使用 3 块 NVIDIA A10 34 GB GPU, 在 512×512 分辨率图像上进行.

为评估 SISRM 的性能, 在 DIV2K-Val 和 RealSR 数据集上进行测试. DIV2K-Val 数据集图像被调整为 512 像素的最短边, 然后被中心裁剪为 512×512, 作为地面真实值. 然后应用与 SeeSR[23]相同的降级流水线来生成 LQ 图像. 对于真实世界的基准测试, 实验使用 RealSR 数据集, 基准测试的分辨率为 128×128. 另外, 所有实验均在缩放因子×4 下进行.

3.3 评价指标

在实验中, 采用多种评价指标对超分辨率结果进行定量评估, 包括峰值信噪比 (peak signal-to-noise ratio, PSNR)、结构相似性指数 (structural similarity index measure, SSIM)[43]、学习感知图像块相似性 (learned perceptual image patch similarity, LPIPS)[44]、转换空间距离 (distance in transformation space, DISTS)[45]、基于元分析的无参考图像质量评估 (meta-analysis-based no-reference image quality assessment, MANIQA)[46]以及跨模态图像质量评估 (cross-modal learning for image quality assessment, CLIPIQ)[47]. 其中, PSNR 衡量重建图像与参考图像在像素亮度上的差异, 值越高表示质量越好; SSIM 评估图像在亮度、对比度和结构方面的相似性, 值越接近 1 表示越相似; LPIPS 通过深度网络建模人类视觉感知差异, 值越低代表质量越高; DISTS 通过深度特征衡量感知相似性, 值越低表示更好; MANIQA 无需参考图像即可评估质量, 值越高表示质量更优; CLIPIQA 则利用跨模态特征进行评估, 值越高表明感知质量越好.

3.4 对比实验分析

将 SISRM 与多个当前最先进的真实图像超分辨

率 (real-ISR) 方法进行对比, 包括: 基于 GAN 的方法: BSRGAN[10]、Real-ESRGAN[11]、DASR[12]; 基于扩散模型的方法: StableSR[17]、ResShift[48]、DiffBIR[18]、PASD[22]、SeeSR[23]. 对比实验采用这些方法公开的官方实现和预训练模型进行测试.

对表 1 结果进行分析, 实验结果表明, 所提方法在无参考指标 (MANIQA、CLIPIQA) 上表现优异, 表明恢复结果质量较高. GAN 方法在 PSNR、SSIM 等指标上表现较好, 说明保真度高, 但缺乏真实细节. 扩散模型则更注重真实感恢复, 但在全参考指标上 (PSNR、SSIM) 得分较低, 可能是因为生成了真实图像中不存在的细节.

表 1 不同数据集上与不同方法的定量比较

数据集	指标	BSRGAN	Real-ESRGAN	DASR	StableSR	ResShift	PASD	DiffBIR	SeeSR	SISRM
DIV2K-Val	PSNR↑ (dB)	21.43	21.51	21.33	20.72	**21.55**	20.52	20.87	20.51	20.43
	SSIM↑	**0.5234**	0.5231	0.5120	0.4773	0.5223	0.4876	0.5016	0.5058	0.4659
	LPIPS↓	0.4133	0.3861	0.4279	0.4042	0.4124	0.4378	0.4661	**0.3723**	0.3758
	DISTS↓	0.2747	0.2599	0.2826	0.2340	0.2583	0.2394	0.2183	0.2037	**0.2029**
	MANIQA↑	0.4295	0.4657	0.3447	0.5378	0.4420	0.5803	0.5570	0.6018	**0.6121**
	CLIPIQA↑	0.5147	0.5371	0.4573	0.6658	0.5713	0.6629	0.7002	0.7111	**0.7274**
RealSR	PSNR↑ (dB)	26.92	26.22	**27.58**	25.79	26.93	26.28	24.75	25.65	26.13
	SSIM↑	0.7773	0.7736	**0.7831**	0.7387	0.7688	0.7411	0.6756	0.7325	0.6967
	LPIPS↓	**0.2706**	0.2761	0.3140	0.2973	0.3221	0.3195	0.3541	0.3067	0.3454
	DISTS↓	0.2139	**0.2074**	0.2217	0.2189	0.2453	0.2321	0.2311	0.2241	0.2207
	MANIQA↑	0.3901	0.3885	0.2571	0.4222	0.4096	0.5281	0.5102	0.5207	**0.5233**
	CLIPIQA↑	0.5016	0.4441	0.3149	0.5709	0.5432	0.5832	0.6831	0.6625	**0.6836**

注: 最佳结果用加粗标出; 次优结果用下划线标出; ↑表示指标值越大, 图像质量越好; ↓表示指标值越小, 图像质量越好

为直观呈现所提 SISRM 方法在细节恢复及语义一致性方面的优势, 选取 DIV2K-Val 与 RealSR 数据集中的典型样例进行可视化对比, 结果见图 6. 各列依次展示 LR 输入、扩散过程的中间生成状态 (t=0.5T、t=0.2T, T 表示总扩散步数)、SISRM 最终重建结果、对比方法 SeeSR 的输出, 以及 DIV2K-Val 样例对应的 HR 参考, 红/蓝框标示局部放大区域位置. 可视化结果显示, SISRM 在扩散过程中逐步恢复高频纹理并增强边缘语义, 最终输出在清晰度和结构保持方面均优于现有方法; 在窗格、胡须等细节区域, SISRM 重建结果表现出更锐利的纹理和更符合真实语义的结构, 而对比方法存在不同程度的模糊或伪影, 上述结果与定量实验趋势一致, 进一步验证了所提出语义感知交互扩散策略在视觉质量提升方面的有效性.

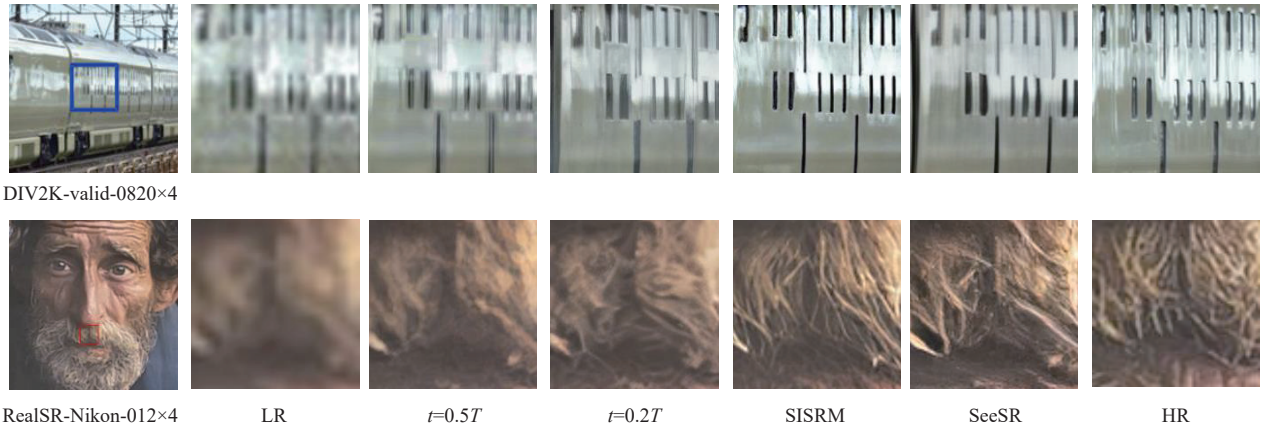

图 6 在 DIV2K-Val (上) 和 RealSR (下) 数据集上的可视化对比

图 7 为 DIV2K-Val 数据集中自然人像图像的定性对比. 对比结果表明, Real-ESRGAN 重建结果存在明显的细节模糊问题, 纹理恢复不清晰; StableSR 和 DiffBIR 虽在高频细节恢复方面有所提升, 但皮肤区域过度锐化并伴随明显伪影, 导致视觉不自然; PASD 和 SeeSR 可借助语义信息一定程度上改善细节恢复的效果, 但眼部纹理仍然存在明显的模糊与不一致性问题. 相比之下, 本文提出的 SISRM 能够清晰地重建眼睛区域的

精细结构与皮肤纹理细节, 边缘锐利且更接近真实图像效果, 展现出更佳的细节重建性能与更高的语义一致性. 这表明 SISRM 在真实世界的人像场景中具有一定的优势.

图 7 DIV2K-Val 数据集中的定性对比

图 8 为 RealSR 数据集中真实拍摄场景图像的定性对比. 对比结果表明, Real-ESRGAN、StableSR 和 DiffBIR 方法在细线条区域表现欠佳, 存在模糊或线条扭曲现象, 难以准确恢复高频结构. PASD 和 SeeSR 在高频细节恢复方面有所改善, 但仍有一定程度的模糊. 相比之下, SISRM 可准确重建分辨率卡上的细线条, 线条间隔与锐利度均与 HR 参考图像高度吻合, 视觉感知质量显著优于其他方法. 这进一步验证了 SISRM 在真实世界超分辨率任务中对结构细节的强大恢复能力及整体视觉质量的提升作用.

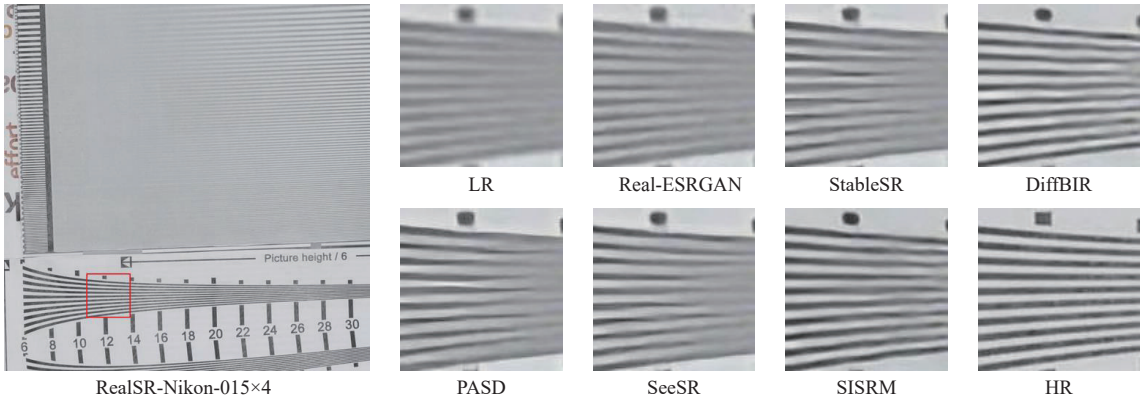

图 8 RealSR 数据集中的定性对比

3.5 消融实验分析

为分析 SISRM 方法中各核心模块对整体性能的贡献, 表 2 给出了在 DIV2K-Val 和 RealSR 数据集上的逐模块消融实验结果. 结果显示, 完整 SISRM 模型在所有指标上均优于移除任何模块的版本, 表明 SAPE、SCA 与 MFF 均对性能提升具有积极作用.

具体而言, 移除 SAPE 导致 PSNR 和感知质量指标显著下降, DIV2K-Val 下降约 2.07 dB, RealSR 下降约 2.15 dB, 说明其提供的语义提示对全局结构和语义一致性至关重要. 移除 SCA 后, PSNR 下降约 1.4 dB, MANIQA 与 CLIPIQA 也明显降低, 表明跨模态语义交叉注意力有效缓解了语义与图像特征间的信息鸿沟. 移除 MFF 会对 DIV2K-Val 的感知指标产生一定影响, LPIPS 与 MANIQA 分别出现恶化, 而在 RealSR 上作用相对较弱, 可能与真实退化分布下的掩码精度有限有关. 完全移除 SCA+MFF 时, 性能下降幅度最大, 验证了两者在细粒度语义控制上的协同作用. 综上, SISRM 方法中各核心模块在不同指标上的提升幅度存在差异, 但整体趋势一致, 且组合使用可显著增强 SISRM 在真实和合成场景下的重建能力, 充分证明所提方法设计

的合理性.

3.6 模型复杂度分析

为系统评估 SISRM 的部署可行性, 本文在统一硬件与输入设置下, 对其与当前主流超分辨率方法进行了模型复杂度对比. 输入图像分辨率设为 128×128, 缩放因子设置为×4. 对比指标包括参数量、计算复杂度(FLOPs)、显存占用及单张图像的推理时间. 具体实验结果如表 3 所示.

尽管 SISRM 构建于扩散模型框架之上, 其设计充分考虑了轻量化与效率的平衡. 在保持较高重建质量的前提下, SISRM 降低了模型规模与计算开销, 推理速度快、显存使用合理, 整体在同类扩散方法中表现出较强的硬件适配性与部署效率, 可见该方法在资源受限场景下具备良好的实用性.

表 2 消融实验结果

方案	数据集	PSNR↑ (dB)	SSIM↑	LPIPS↓	DISTS↓	MANIQA↑	CLIPIQA↑
Baseline	DIV2K-Val	17.12	0.3232	0.4221	0.2898	0.5121	0.6511
	RealSR	23.85	0.5011	0.4335	0.2972	0.4894	0.5989
w/o SAPE	DIV2K-Val	18.36	0.3651	0.4181	0.2656	0.5875	0.6892
	RealSR	23.98	0.5429	0.4093	0.2711	0.4961	0.6054
w/o SCA	DIV2K-Val	19.01	0.3827	0.4011	0.2532	0.5922	0.7023
	RealSR	24.67	0.5593	0.3968	0.2648	0.5012	0.6492
w/o MFF	DIV2K-Val	19.53	0.3981	0.3935	0.2115	0.6068	0.7156
	RealSR	25.94	0.5757	0.3752	0.2231	0.5075	0.6721
w/o SCA+MFF	DIV2K-Val	18.96	0.3771	0.4074	0.2589	0.5793	0.6962
	RealSR	24.73	0.5573	0.3914	0.2451	0.4984	0.6603
SISRM	DIV2K-Val	20.43	0.4659	0.3758	0.2029	0.6121	0.7274
	RealSR	26.13	0.6967	0.3454	0.2207	0.5133	0.6836

注: ↑表示指标值越大, 图像质量越好; ↓表示指标值越小, 图像质量越好

表 3 不同方法的模型复杂度

指标	BSRGAN	Real-ESRGAN	DASR	StableSR	ResShift	PASD	DiffBIR	SeeSR	SISRM
参数量 (M)	16	43	32	114	57	93	89	129	48
FLOPs (G)	86	149	121	563	284	497	471	612	198
显存 (MB)	1600	3100	2500	9100	6300	8200	7500	9800	5500
推理时间 (ms)	28	41	36	115	88	102	98	120	73

4 结论

本文提出一种基于 SISRM 的真实世界图像超分辨率重建方法, 该方法创新性地融合语义分割感知机制与扩散模型, 以实现更有效的语义引导和结构感知图像重建. 具体而言, 本方法是以分割感知提示提取器(SAPE) 为核心的语义引导框架, 并集成交互式文本到图像控制器 (IT2IC) 与掩码特征融合机制 (MFF), 以提升模型对局部细节与全局结构的感知能力. 通过引入语义提示作为附加控制条件, SISRM 在真实场景下能够在保持语义一致性的同时, 准确恢复物体细节和纹理结构, 实验结果在多个指标上全面优于现有方法, 验证了其在真实世界图像超分辨率任务中的有效性.

参考文献

1 Chen HT, Wang YH, Guo TY, *et al*. Pre-trained image processing Transformer. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 12294–12305. [doi: 10.1109/CVPR46437.2021.01212]

2 Chen XY, Wang XT, Zhou JT, *et al*. Activating more pixels in image super-resolution Transformer. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 22367–22377. [doi: 10.1109/CVPR52729.2023.02142]

3 Liang JY, Cao JZ, Sun GL, *et al*. SwinIR: Image restoration using Swin Transformer. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 1833–1844. [doi: 10.1109/ICCVW54120.2021.00210]

4 Ma JQ, Guo S, Zhang L. Text prior guided scene text image super-resolution. IEEE Transactions on Image Processing, 2023, 32: 1341–1353. [doi: 10.1109/TIP.2023.3237002]

5 Sun LC, Liang J, Liu SZ, *et al*. Perception-distortion balanced super-resolution: A multi-objective optimization perspective. IEEE Transactions on Image Processing, 2024, 33: 4444–4458. [doi: 10.1109/TIP.2024.3434426]

6 Zhang XD, Zeng H, Guo S, *et al*. Efficient long-range attention network for image super-resolution. Proceedings of

the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022. 649–667. [doi: 10.1007/978-3-031-19790-1_39]

7 Zhang ZQ, Li RH, Guo S, *et al*. TMP: Temporal motion propagation for online video super-resolution. IEEE Transactions on Image Processing, 2024, 33: 5014–5028. [doi: 10.1109/TIP.2024.3453048]

8 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]

9 Wang XT, Yu K, Wu SX, *et al*. ESRGAN: Enhanced super-resolution generative adversarial networks. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 63–79. [doi: 10.1007/978-3-030-11021-5_5]

10 Zhang K, Liang JY, Van Gool L, *et al*. Designing a practical degradation model for deep blind image super-resolution. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 4771–4780. [doi: 10.1109/ICCV48922.2021.00475]

11 Wang XT, Xie LB, Dong C, *et al*. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 1905–1914. [doi: 10.1109/ICCVW54120.2021.00217]

12 Liang J, Zeng H, Zhang L. Efficient and degradation-adaptive network for real-world image super-resolution. Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022. 574–591. [doi: 10.1007/978-3-031-19797-0_33]

13 Qi CY, Tu ZZ, Ye KR, *et al*. SPIRE: Semantic prompt-driven image restoration. Proceedings of the 18th European Conference on Computer Vision. Milan: Springer, 2024. 446–464. [doi: 10.1007/978-3-031-73661-2_25]

14 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.

15 Dhariwal P, Nichol A. Diffusion models beat GANs on image synthesis. Proceedings of the 35th International Conference on Neural Information Processing System. Curran Associates Inc., 2021. 672.

16 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. [doi: 10.1109/CVPR52688.2022.01042]

17 Wang JY, Yue ZS, Zhou SC, *et al*. Exploiting diffusion prior for real-world image super-resolution. International Journal of Computer Vision, 2024, 132(12): 5929–5949. [doi: 10.1007/s11263-024-02168-7]

18 Lin XQ, He JW, Chen ZY, *et al*. DiffBIR: Toward blind image restoration with generative diffusion prior. Proceedings of the 18th European Conference on Computer Vision. Milan: Springer, 2024. 430–448. [doi: 10.1007/978-3-031-73202-7_25]

19 Fan YT, Liu CX, Yin NZ, *et al*. AdaDiffSR: Adaptive region-aware dynamic acceleration diffusion model for real-world image super-resolution. Proceedings of the 18th European Conference on Computer Vision. Milan: Springer, 2024. 396–413. [doi: 10.1007/978-3-031-73254-6_23]

20 Yu FH, Gu JJ, Li ZY, *et al*. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 25669–25680. [doi: 10.1109/CVPR52733.2024.02425]

21 Qu YP, Yuan K, Zhao K, *et al*. XPSR: Cross-modal priors for diffusion-based image super-resolution. Proceedings of the 18th European Conference on Computer Vision. Milan: Springer, 2024. 285–303. [doi: 10.1007/978-3-031-73247-8_17]

22 Yang T, Wu RY, Ren PR, *et al*. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. Proceedings of the 18th European Conference on Computer Vision. Milan: Springer, 2024. 74–91. [doi: 10.1007/978-3-031-73247-8_5]

23 Wu RY, Yang T, Sun LC, *et al*. SeeSR: Towards semantics-aware real-world image super-resolution. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 25456–25467. [doi: 10.1109/CVPR52733.2024.02405]

24 Li JN, Li DX, Savarese S, *et al*. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. Proceedings of the 40th International Conference on Machine Learning. Honolulu: PMLR, 2023. 19730–19742.

25 Zhang YC, Huang XY, Ma JY, *et al*. Recognize anything: A strong image tagging model. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle: IEEE, 2024. 1724–1732.

26 Zhang LM, Rao AY, Agrawala M. Adding conditional control to text-to-image diffusion models. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023. 3813–3824. [doi: 10.1109/ICCV51070.2023.00355]

27 Guo MH, Lu CZ, Hou QB, *et al*. SegNeXt: Rethinking convolutional attention design for semantic segmentation.

Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 84.

28 Xie EZ, Wang WH, Yu ZD, *et al*. SegFormer: Simple and efficient design for semantic segmentation with Transformers. Proceedings of the 35th International Conference on Neural Information Processing Systems. Curran Associates Inc., 2021. 924.

29 Zou XY, Dou ZY, Yang JW, *et al*. Generalized decoding for pixel, image, and language. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 15116–15127. [doi: 10.1109/CVPR52729.2023.01451]

30 Zou XY, Yang JW, Zhang H, *et al*. Segment everything everywhere all at once. Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023. 868.

31 Lai ZQ, Duan YC, Dai JF, *et al*. Denoising diffusion semantic segmentation with mask prior modeling. arXiv: 2306.01721, 2023.

32 Hoogeboom E, Nielsen D, Jaini P, *et al*. Argmax flows and multinomial diffusion: Learning categorical distributions. Proceedings of the 35th International Conference on Neural Information Processing Systems. Curran Associates Inc., 2021. 953.

33 Wang XT, Yu K, Dong C, *et al*. Recovering realistic texture in image super-resolution by deep spatial feature transform. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 606–615. [doi: 10.1109/CVPR.2018.00070]

34 Aakerberg A, Johansen AS, Nasrollahi K, *et al*. Semantic segmentation guided real-world super-resolution. Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2022. 449–458.

35 Wang CC, Hao ZW, Tang YH, *et al*. SAM-DiffSR: Structure-modulated diffusion model for image super-resolution. arXiv:2402.17133, 2025.

36 Kirillov A, Mintun E, Ravi N, *et al*. Segment anything. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023. 3992–4003. [doi: 10.1109/ICCV51070.2023.00371]

37 Sohl-Dickstein J, Weiss EA, Maheswaranathan N, *et al*. Deep unsupervised learning using nonequilibrium thermodynamics. Proceedings of the 32nd International Conference on Machine Learning. Lille: PMLR, 2015. 2256–2265. [doi: 10.48550/arXiv.1503.03585]

38 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

39 Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv:1503.02531, 2015.

40 Agustsson E, Timofte R. NTIRE 2017 challenge on single image super-resolution: Dataset and study. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu: IEEE, 2017. 1122–1131.

41 Cai JR, Zeng H, Yong HW, *et al*. Toward real-world single image super-resolution: A new benchmark and a new model. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 3086–3095. [doi: 10.1109/ICCV.2019.00318]

42 Hu EJ, Shen YL, Wallis P, *et al*. LoRA: Low-rank adaptation of large language models. Proceedings of the 10th International Conference on Learning Representations. OpenReview.net, 2022.

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

44 Zhang R, Isola P, Efros AA, *et al*. The unreasonable effectiveness of deep features as a perceptual metric. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 586–595. [doi: 10.1109/CVPR.2018.00068]

45 Ding KY, Ma KD, Wang SQ, *et al*. Image quality assessment: Unifying structure and texture similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(5): 2567–2581. [doi: 10.1109/TPAMI.2020.3045810]

46 Yang SD, Wu TH, Shi SW, *et al*. MANIQA: Multi-dimension attention network for no-reference image quality assessment. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 1190–1199. [doi: 10.1109/CVPRW56347.2022.00126]

47 Wang JY, Chan KCK, Loy CC. Exploring clip for assessing the look and feel of images. Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington: AAAI, 2023. 2555–2563. [doi: 10.1609/aaai.v37i2.25353]

48 Yue ZS, Wang JY, Loy CC. Efficient diffusion model for image restoration by residual shifting. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(1): 116–130. [doi: 10.1109/TPAMI.2024.3461721]

(校对责编: 张重毅)