文章编号:1673-0291(2026)04-0119-10 **DOI**:10.11860/j.issn.1673-0291.20250138

面向工务巡检系统评定的钢轨扣件图像合成

罗 伟[1],贺 雨[2,3],邬 俊[1]

(1.北京交通大学 计算机科学与技术学院,北京 100044;2.中国铁道科学研究院集团有限公司 基础设施检测研究所,北京 100081;3.轨道交通基础设施智能检测运维北京市重点实验室,北京 100081)

摘 要:铁路工务巡检系统性能评定依赖于大量高质量异常扣件图像,而真实异常样本获取困难,使得图像合成成为解决数据不足问题的重要手段.图像和谐化作为图像合成的关键步骤,现有扩散模型方法未充分考虑铁路巡检图像中前景扣件与背景轨面的结构差异,同时噪声预测与图像和谐化目标耦合,限制了生成图像质量.针对上述问题,提出一种基于随机微分方程(Stochastic Differential Equation,SDE)的双分支钢轨扣件图像和谐化方法.首先,采用变分自编码器(Variational Autoencoder,VAE)将输入图像映射至隐空间,在降低高分辨率图像计算开销的同时保留关键结构信息;然后,设计内容匹配注意力模块,通过建立前景扣件与背景区域之间的内容匹配关系,引导模型重点关注背景中与前景结构相似的扣件区域,抑制轨面等无关背景区域带来的干扰,提高前景与背景的视觉一致性;最后,构建由重建分支和和谐化分支组成的双分支扩散网络,其中重建分支负责学习稳定的噪声预测能力,并通过多尺度特征约束引导和谐化分支训练,缓解去噪目标与和谐化目标之间的优化冲突,进一步提升图像生成质量.在自建铁路异常扣件图像数据集上开展的实验结果表明:所提方法相比现有主流图像和谐化方法,在均方误差(Mean Squared Error,MSE)和前景均方误差(foreground Mean Squared Error,fMSE)上分别提升 14.37% 和 8.91%,同时在峰值信噪比(Peak Signal-to-Noise Ratio,PSNR)和结构相似性指数(Structural Similarity Index Measure,SSIM)上取得最优性能.主观评价实验及下游异常扣件检测实验进一步验证了所生成图像具有较高的真实性和视觉一致性,可为铁路工务巡检系统性能评定提供高质量的数据支撑.

关键词:铁路运输;工务巡检;图像合成;图像和谐化;扩散模型

中图分类号:U216;TP391.41 **文献标志码**:A

Rail-fastener image synthesis for evaluating track inspection systems

Luo Wei[1], *He Yu*[2,3], *Wu Jun*[1]

(1.School of Computer Science and Technology, Beijing Jiaotong University, Beijing 100044, China;2.Infrastructure Inspection Institute, China Academy of Railway Sciences Corporation Limited, Beijing 100081, China;3.Beijing Key Laboratory of Intelligent Inspection & Maintenance for Rail Transit Infrastructure, Beijing 100081, China)

Abstract: Evaluating railway track inspection systems requires a large number of high-quality images

收稿日期:2025-09-08;**修回日期**:2026-03-22

基金项目:北京市自然科学基金(L232033);中央高校基本科研业务费专项资金(2025JBZY011);中国铁道科学研究院集团有限公司科技研究开发计划(2024YJ392)

Foundation items: Beijing Natural Science Foundation (L232033); Fundamental Research Funds for the Central Universities (2025JBZY011); Science and Technology Research and Development Program of China Academy of Railway Sciences Corporation Limited (2024YJ392)

第一作者:罗伟(2001—),男,江苏溧阳人,硕士生.研究方向为计算机视觉.**email**:23120389@bjtu.edu.cn.

通信作者:邬俊(1981—),男,辽宁鞍山人,博士,教授.**email**:wuj@bjtu.edu.cn.

引用格式:罗伟,贺雨,邬俊.面向工务巡检系统评定的钢轨扣件图像合成[J].北京交通大学学报,2026,50(4):119-128.

Luo Wei, He Yu, Wu Jun. Rail-fastener image synthesis for evaluating track inspection systems[J]. Journal of Beijing Jiaotong University, 2026, 50(4): 119-128.(in Chinese)

of defective fasteners. Because real defective samples are difficult to obtain, image synthesis provides an effective means of addressing data scarcity. Image harmonization is a critical step in this process. However, existing diffusion-based methods do not adequately account for the structural differences between foreground fasteners and background rail surfaces in railway inspection images, and the coupling between noise prediction and image harmonization limits the quality of the generated images. To address these issues, this study proposes a dual-branch rail-fastener image harmonization method based on Stochastic Differential Equations (SDEs). First, a Variational Auto-Encoder (VAE) maps the input image into a latent space, reducing the computational cost of processing high-resolution images while preserving essential structural information. Next, a Content-Matching Attention Module (CMAM) establishes correspondences between foreground fasteners and background regions. This module directs the model toward background fasteners that are structurally similar to the foreground fastener while suppressing interference from irrelevant rail-surface regions, thereby improving visual consistency between the foreground and background. Finally, a dual-branch diffusion framework comprising reconstruction and harmonization branches is constructed. The reconstruction branch learns stable noise prediction and guides the training of the harmonization branch through multiscale feature constraints, alleviating the optimization conflict between the denoising and harmonization objectives and further improving image quality. Experiments on a self-constructed dataset of defective railway fasteners show that the proposed method outperforms state-of-the-art image harmonization methods, reducing Mean Squared Error (MSE) and foreground Mean Squared Error (fMSE) by 14.37% and 8.91%, respectively, while achieving the best Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM). Subjective evaluations and downstream fastener-detection experiments further confirm the high realism and visual consistency of the generated images, demonstrating that the proposed method can provide high-quality data for evaluating railway track inspection systems.

Keywords: railway transportation; railway track inspection; image synthesis; image harmonization; diffusion model

扣件是铁路系统的核心设施之一,扣件一旦出现诸如移位、变形等异常情况,可能对列车运行安全构成严重威胁,甚至引发脱轨事故,进而导致重大经济损失.

深度学习的快速发展,推动多种计算机视觉模型被应用到工务巡检系统之中,以实现扣件异常自动检测[1-3].然而,铁路扣件巡检系统的研发由各铁路局负责,其性能评定则由铁路检测中心承担,这使得如何客观评估各系统的检测性能成为铁路检测中心亟需解决的问题.有效的性能评估依赖于充足的高质量数据,但该需求受限于铁路线路实际运行中扣件异常的低发生率,获取高质量、多样化的异常扣件图像存在较大困难.因此,构建一套铁路扣件异常图像合成框架势在必行.广义的图像合成分为3个步骤:图像拼贴、图像和谐化和阴影生成.鉴于扣件异常图像的尺寸统一且采用垂直拍摄方式,在合成此类图像时无需考虑图像拼贴带来的前背景尺寸适配问题,亦无需涉及阴影生成问题.因此,目前方法主要聚焦于图像和谐化技术.该技术的核心作用是调整前景光照,使其与背景图像在视觉上保持一致性.

学者提出了大量图像和谐化算法[4-18],在过去十余年间,低分辨率方法将图像和谐任务视为通过一个或多个编码器-解码器网络(通常为U-Net)进行像素级转换.文献[4]提出一种基于U-Net的和谐网络,增加了语义分支,这是首个端到端的深度图像和谐网络.文献[5]使用领域验证判别器,强制前景与背景转换到同一领域.文献[6]提出一种简单高效的即插即用模块,根据背景统计调整前景风格.文献[7]将任务表述为背景引导的领域转换任务,采用三元组损失.文献[8]引入一个预训练网络,提取前景对象的高层次语义表示,并将其添加到U-Net主干中.文献[9]提出一种新的注意力模块,该模块包含的注意力机制根据前景和背景补丁的相似性计算每个前景补丁对应的背景统计特征,并增加了一个自洽的风格对比学习任务.除了使用U-Net,文献[10-11]设计了一个Transformer模块来提高建模长距离上下文依赖的能力.

近年来,一些高分辨率方法被提出.基于低分辨率深度网络的应用,这些方法引入了一些低计算量的颜色转换方法,以减少高分辨率图像的计算负担.文献[12]通过U-Net获取低分辨率和谐图像,通过可学习的3D-LUTs获取高分辨率和谐图像,然后通过额外的卷积神经网络(Convolutional Neural Network,CNN)将它们结合.文献[13]通过在色相、饱和度、亮度颜色空间(Hue-Saturation-Value,HSV)中对原始分辨率图像应用滤波器进行补丁级转换,而滤波器参数来自低分辨率和谐网络.文献[14]通过引入通常在图像处理中使用的白盒滤波器来执行图像级别的转换.文献[15]设计了一个白盒图像与视频和谐化网络,通过显式建模前景与背景之间的颜色、纹理和光照差异,实现对合成前景的自然融合与视觉一致性校正.此外,随着扩散模型的爆火,文献[16-17]尝试使用扩散模型进行图像和谐化:文献[16]在扩散模型之后级联一个微调网络对扩散模型结果进行精调;文献[17]使用真实(Ground Truth,GT)图像对变分自编码器(Variational Auto-Encoder,VAE)的解码器进行微调,以求解码出GT图像.

虽然扩散模型的出现带来了显著的性能提升,但已有方法大多基于通用场景进行设计[19-23],默认假设背景区域在光照分布上与前景目标具有较强一致性,因而可作为稳定、有效的先验信息加以利用.然而,该假设在工务巡检图像的扣件和谐化任务中并不成立.具体而言,工务巡检图像中扣件区域与轨面区域在纹理结构以及光照反射特性等方面存在显著差异.在现有方法中,通常将整个背景区域(包括背景中的扣件区域与大面积轨面区域)整体作为先验信息引入模型,用以引导前景扣件的和谐化.然而,在该特定任务中,背景区域中仅有少量与前景结构相似的扣件区域能够为和谐化过程提供有效参考,其他占据主导比例的轨面区域不仅难以提供有价值的先验约束,反而由于其与前景扣件在光照反射特性上的显著差异,极易在特征对齐与生成过程中引入无关甚至误导性的干扰信号.

鉴于此,本文提出一种基于内容匹配注意力机制的双分支图像和谐化方法(Dual-Branch image harmonization method based on Content-Matching Attention Mechanism,DB-CMAM).该方法以铁路巡检图像中"前景扣件-背景轨面"显著异构的成像特点为出发点,强调对有效背景先验的精准筛选与利用.该方法的核心是设计了一个内容匹配注意力模块(Content-Matching Attention Module,CMAM).该模块将前背景特征划分为图像块,以前景特征块作为查询(Query),背景特征块作为键(Key)和值(Value),计算注意力权重,以此引导模型自适应地聚焦于背景中与前景内容高度相似(即表观相近的扣件)的图像块,有效滤除背景噪声的干扰,从而显著提升生成图像中前景与背景的视觉一致性.同时,为克服扩散模型在图像和谐化任务中存在的"去噪目标"与"和谐化目标"耦合所导致的优化困难,DB-CMAM引入双分支协同机制,其包含重建分支和和谐化分支:重建分支专注噪声预测;和谐化分支在重建分支的噪声预测引导下,专注生成具有真实感且与背景和谐融合的前景区域.

1　DB-CMAM设计

假设复合灰度图像$I_{\mathrm{LQ}}\in\mathbb{R}^{H\times W\times 1}$的宽度为$W$、高度为$H$、通道数为1,其对应的二值前景掩码记为$M_{\mathrm{I}}\in\{0,1\}^{H\times W}$,该掩码用于精确指示图像中的前景区域.图像和谐化任务中,设置复合灰度图像I_{LQ}和二值输入掩码M_{I}作为输入,其核心目标是调整前景区域的光照属性,使前景与背景图像在视觉风格上保持一致性.该任务的最终目标可定义为最小化预测图像$I\in\mathbb{R}^{H\times W\times 1}$与真实图像$I_{\mathrm{GT}}\in\mathbb{R}^{H\times W\times 1}$之间的误差.

为更好地利用背景的上下文信息,本文提出一个内容匹配注意力引导的双分支图像和谐化方法,即DB-CMAM.该方法主要由两个核心部分组成:内容匹配注意力模块和双分支扩散机制.模型结构如图1所示.图1中:y_0为重建分支的原始输入图像;y_t为重建分支加噪到t时间步的图像;x_0为和谐化分支的输入图像;x_t为和谐化分支加噪到t时间步的图像;$\mathrm{d}x_t$为依据噪声得到的退化偏移量;x_{t-1}与y_{t-1}分别为x_t与y_t推理得到的前一时间步状态;L_{fea}为特征损失.内容匹配注意力模块在和谐化分支中发挥核心作用,通过引导模型关注背景中与前景语义相似的结构区域,从而提升生成图像的语义一致性与视觉协调性.双分支扩散机制则包含一个重建分支和一个和谐化分支,前者仅执行噪声预测任务,提供稳定的训练目标,引导模型学习更可靠的去噪能力;后者则在重建分支的辅助下,完成图像的和谐化任务.

1.1　双分支网络

在文献[24]的设计中,模型需要同时完成噪声预测与退化图像内容生成两项任务.这种多任务优化带来了明显的问题:一方面,噪声预测要求模型精准估计噪声分布;另一方面,图像生成对纹理构建、细节生成提出更高要求.这种设计不仅增加了模型的复杂性,同时还可能抑制模型潜在性能的充分发

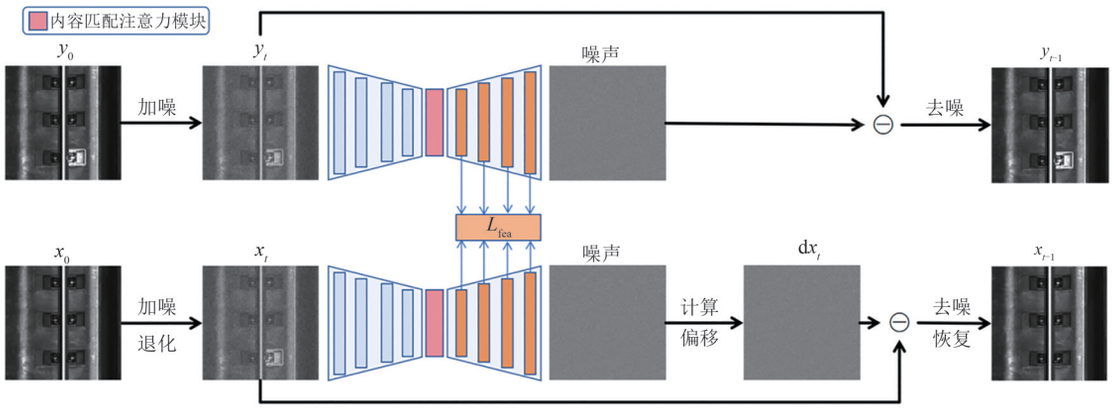


图 1 网络结构图

Fig.1 Network architecture

挥.针对这一问题,提出一个双分支架构,其包含一个和谐化分支和一个重建分支.

在和谐化分支中,将前向加噪退化过程定义为从高质量图像逐步加噪退化至低质量图像的过程.具体而言,在具有T个时间步的扩散过程中,将初始状态x_0设置为I_{GT},最终状态x_{T}设置为低质图像$\mu=I_{\text{LQ}}$与高斯噪声ε的组合.对于任意时间步$t\in[0,T]$,该时间步状态x_t通过均值回归随机微分方程(Stochastic Differential Equation,SDE)定义为

$$\mathrm{d}x_t=-\theta_t(\mu-x_t)\mathrm{d}t+\sigma_t\mathrm{d}w \tag{1}$$

式中:w为标准维纳过程;θ_t和σ_t是时间依赖的正参数,分别为均值回归速度和随机波动率.这些参数受到$\eta_t^2/\theta_t=2\lambda^2$的约束,以实现平稳方差,其中$\lambda$为应用于$x_{\text{T}}$的固定噪声强度,$\eta_t$为时间相关的参数.

反向去噪恢复过程定义为从加噪后的低质图像逐步去噪恢复至高质量图像的过程.为了从T时间步状态y_{T}恢复至原始图像y_0,可通过反转随机微分方程得到逆扩散过程,其表达式为

$$\mathrm{d}x_t=\left[\theta_t(\mu-x_t)-\sigma_t^2\nabla_{x_t}\log p_t(x_t)\right]\mathrm{d}t+\sigma_t\mathrm{d}\hat{w} \tag{2}$$

式中:$\nabla_{x_t}\log p_t(x_t)$为当前状态的分数函数;$\mathrm{d}\hat{w}$为标准布朗运动,式(1)和式(2)的详细推导过程见文献[25].

在重建分支中,前向过程被定义为对低质量图像进行逐步加噪.具体而言,在具有T个时间步的扩散过程中,将初始状态x_0设置为I_{LQ},最终状态x_{T}设置为低质量图像I_{LQ}与高斯噪声ε的组合.对于任意时间步$t\in[0,T]$,该时间步状态x_t通过均值回归随机微分方程定义为

$$\mathrm{d}x_t=\sigma_t\mathrm{d}w \tag{3}$$

反向去噪过程定义为从加噪后的低质量图像逐步去噪恢复至低质量图像的过程.为了从T时间步状态x_{T}恢复至原始图像x_0,通过反转随机微分方程得到逆扩散过程,其形式为

$$\mathrm{d}x_t=-\sigma_t^2\nabla_{x_t}\log p_t(x_t)\mathrm{d}t+\sigma_t\mathrm{d}\hat{w} \tag{4}$$

另外,采用分阶段训练策略.首先,对重建分支进行独立训练,使其充分学习扩散过程中的噪声预测能力,从而获得稳定的多尺度中间特征表示.然后,进入和谐化分支的训练阶段,此时冻结重建分支的全部参数,以保证其特征的稳定性与一致性,在和谐化分支训练过程中,针对U-Net解码阶段的多个尺度层,将重建分支与和谐化分支在对应分辨率下的特征图进行一一配对.最后,通过在这些配对特征之间引入特征对齐损失,显式约束和谐化分支的中间表示向重建分支的特征空间靠拢,该对齐损失作为一种结构化的引导信号,被联合纳入整体训练目标中,从而在不直接共享参数的前提下,促进和谐化分支在噪声预测任务中学习到更加稳定的表示.通过这种双分支特征对齐机制,重建分支中蕴含的噪声建模先验能够有效迁移至和谐化分支,辅助和谐化分支在噪声预测任务上的学习.

1.2 内容匹配注意力模块

在异常扣件图像中,背景中的扣件部分可以给予很好的引导,这些部分应该得到更多的关注.因此设计内容匹配注意力模块,如图2所示.x为输入特征图;O为输出特征图;x_{f}和x_{b}分别为通过前景掩码和背景掩码分离得到的前景特征和背景特征.$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$分别表示查询(Query)、键(Key)和值(Value)特征,其中$\boldsymbol{Q}$由前景特征映射得到,用于表示待匹配区域;$\boldsymbol{K}$和$\boldsymbol{V}$由背景特征映射得到,用于提供参考信息,Attn表示计算得到的注意力权重矩阵,$\boldsymbol{S}$表示基于余弦相似度计算得到的空间相似度矩阵.该模块通过引导注意力集中于背景中与前景相似的区域,促进模型更好地学习背景中的有用信息,让图像更加和谐.

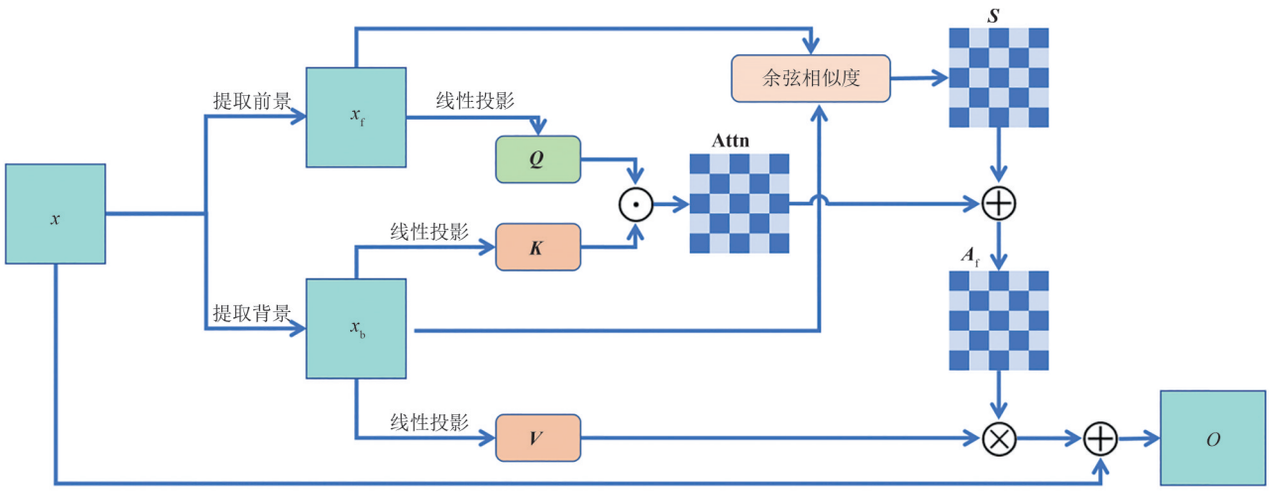


图 2　内容匹配注意力模块

Fig.2　Content-matching attention module

具体来说,给定输入特征图$x\in\mathbb{R}^{B\times C\times H\times W}$及其对应的二值前景掩码$M\in\{0,1\}^{B\times 1\times H\times W}$,其中,$B$为批大小,$C$为通道数.该模块首先对特征图进行掩码,得到分离的前景特征x_f和背景特征x_b为

$$\begin{cases} x_f = x\cdot M \\ x_b = x\cdot(1-M) \end{cases} \tag{5}$$

将前景图像块映射为$\boldsymbol{Q}$,将背景图像块映射为$\boldsymbol{K}$和$\boldsymbol{V}$,其表达式为

$$\begin{cases} \boldsymbol{Q} = \boldsymbol{W}_q(x_f) \\ \boldsymbol{K} = \boldsymbol{W}_k(x_b) \\ \boldsymbol{V} = \boldsymbol{W}_v(x_b) \end{cases} \tag{6}$$

对$\boldsymbol{Q}$和$\boldsymbol{K}$做归一化,然后计算注意力矩阵为

$$\mathbf{Attn} = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{K}}{\sqrt{d}}\right) \tag{7}$$

式中:Softmax为归一化指数出数;$\boldsymbol{W}_q$为查询矩阵;$\boldsymbol{W}_k$为键矩阵;$\boldsymbol{W}_v$为值矩阵;d为通道维度.

为了让注意力更精确地对齐背景中与前景相似的局部结构,进一步引入空间相似性引导.具体而言,计算背景每个位置与前景区域之间的相似度,这一过程通过计算背景区域与前景特征的余弦相似度矩阵来实现,其表达式为

$$\boldsymbol{S} = \mathrm{consine}(x_b^{\mathrm{T}}, x_f) \tag{8}$$

式中:consine为余弦相似度计算.

通过M对前景区域进行屏蔽,$\boldsymbol{S}$可表示为

$$\boldsymbol{S} = \boldsymbol{S}\cdot(1-M) \tag{9}$$

将生成的余弦相似度矩阵与原始背景-前景注意力结合,经过Softmax归一化后得到最终的注意力分布$\boldsymbol{A}_f$为

$$\boldsymbol{A}_f = \mathrm{Softmax}(\mathbf{Attn} + \gamma\cdot\boldsymbol{S}) \tag{10}$$

式中:γ为可学习的缩放因子,用来控制相似度引导的强度.

最后,用融合后的权重对背景值$\boldsymbol{V}$加权聚合,并通过残差连接生成模块输出为

$$O = \boldsymbol{A}_f\cdot\boldsymbol{V} + x \tag{11}$$

1.3　损失函数设计

总损失L_{total}包含全局损失L_{whole},局部损失L_{local},特征损失L_{fea}共3个关键组件,其表达式为

$$L_{\mathrm{total}} = \lambda_1 L_{\mathrm{whole}} + \lambda_2 L_{\mathrm{local}} + \lambda_3 L_{\mathrm{fea}} \tag{12}$$

式中:λ_1、λ_2、λ_3为预设权重系数,分别用于平衡全局损失L_{global}、局部损失L_{local}与特征损失L_{fea}的贡献.

1.3.1　全局损失

由随机微分方程求解可得$t-1$时间步的最优状态x_{t-1}^*为

$$x_{t-1}^* = \frac{1-e^{-2\bar{\theta}_{t-1}}}{1-e^{-2\bar{\theta}_t}}e^{-\theta'}(x_t-\mu) + \frac{1-e^{-2\theta_t'}}{1-e^{-2\bar{\theta}_t}}e^{-\theta'_{t-1}}(x_0-\mu)+\mu \tag{13}$$

式中:$\theta_t' = \int_{t-1}^{t}\theta_i\mathrm{d}i$,$\bar{\theta}_t = \int_0^t\theta_i\mathrm{d}i$,$\theta_i$为第$i$个时刻的均值回归系数;$\theta_t'$表示第$t-1$时刻至第$t$时刻区间内扩散系数$\theta_i$的积分量;$\bar{\theta}_t$为从初始时刻0至第$t$时刻扩散系数$\theta_i$的累积积分量,推导过程见文献[25].在此基础上,以L_{whole}为目标训练噪声预测网络$\epsilon_\phi(x_t,t)$,其表达式为

$$L_{\mathrm{whole}} = \sum_{t=1}^{T}\beta_t E[\|x_t - \mathrm{d}x_t - x_{t-1}^*\|] \tag{14}$$

式中:β_t为与t相关的参数;$x_t-\mathrm{d}x_t$为模型预测得出的$t-1$步状态;$E[\cdot]$为数学期望算子.

1.3.2　局部损失

由于图像和谐化任务着重于前景区域的调和,为更好地引导模型学习,希望模型的关注点在前景

区域上，通过局部损失来实现，表达式为

$$L_{\text{local}}=\sum_{t=1}^{T}\beta_t E\left[\left\|(x_t-\mathrm{d}x_t)\cdot M-x_{t-1}^{*}\cdot M\right\|\right] \quad (15)$$

1.3.3 特征损失

为了能为和谐化分支提供噪声预测引导，设计重建网络，并通过计算两个网络的特征之间的损失来进行约束，具体损失函数为

$$L_{\text{fea}}=\sum_{j=0}^{n}\omega_j\left(\left\|x_{\text{h}}^{j}-x_{\text{r}}^{j}\right\|\right) \quad (16)$$

式中：ω_j为重建网络中解码器部分的第j层预先定义的参数，用以控制第i层的约束力度；x_{h}^{j}和x_{r}^{j}分别为和谐化网络和重建网络中解码器部分的第j层特征.

2 实验与结果分析

为验证所提和谐化方法在铁路异常扣件图像上的有效性，本文从以下3个方面展开系统实验设计与分析：首先，从明确实验数据、模型参数与评价指标等方面介绍实验设置；然后，将该方法与现有代表性方法进行多维度性能比较，以凸显其在该特定任务中的优势；最后，通过设计消融实验拆解各核心模块，深入分析各模块对模型性能的影响.

2.1 实验设置

使用自建铁路异常扣件图像数据集进行实验，该数据集包括8 265个训练图像对和2 028个测试图像对.每个图像对包括一张合成图像，一张掩码图像和一张真实图像[25].

与文献[17]一致，使用均方误差(Mean Squared Error，MSE)、前景区域均方误差(foreground Mean Squared Error，fMSE)、峰值信噪比(Peak Signal-to-Noise Ratio，PSNR)、结构相似性指数(Structural Similarity Index Measure，SSIM)对模型的和谐化结果进行评估.MSE用于衡量生成图像与真实图像之间的整体像素差异，其计算所有像素位置的均方误差，能够反映模型对图像整体颜色、纹理及亮度信息的恢复能力，MSE值越小表示生成结果与真实图像之间的差异越小，图像质量越高.fMSE是在MSE基础上针对前景区域设计的评价指标，仅计算目标区域(异常扣件区域)内的像素误差，能够避免大面积背景区域对评价结果的影响，更准确地衡量模型对前景目标的和谐化效果；对于铁路扣件图像合成任务而言，fMSE能够更加直观地反映异常扣件区域在颜色、纹理和光照方面与真实图像之间的一致程度，fMSE值越小表示前景区域的恢复效果越好.PSNR通过计算生成图像与真实图像之间的信噪比评价图像重建质量，能够反映图像受到噪声干扰的程度.PSNR值越高，说明生成图像与真实图像之间的失真越小，模型生成结果具有更好的保真能力.SSIM从亮度、对比度和结构信息3个方面衡量2幅图像之间的相似程度，相比于仅考虑像素差异的MSE和PSNR，SSIM更加符合人眼视觉感知特点.SSIM取值范围通常为0～1，其值越接近1，表示生成图像与真实图像在结构信息和视觉感知上的一致性越高.

为了验证所提方法DB-CMAM在铁路异常扣件图像和谐化任务中的有效性，将该方法(高分辨率图像和谐方法)与2种低分辨率图像和谐化方法(RainNet[6]、PHDNet[18])和3种高分辨率图像和谐化方法(DiffHarmony＋＋[17]、CDTNet[12]、DCCF[13])进行性能对比.由于本任务要求和谐化图像与输入图像保持相同的分辨率，若采用低分辨率方法并对其结果进行上采样，将不可避免地引入图像模糊，因此仅选择2种低分辨率方法进行比较.

采用Lion优化器进行训练，初始学习率设为3×10^{-5}.为提升收敛稳定性，使用True Cosine Annealing学习率调度策略，最小学习率设置为1×10^{-7}，并在迭代次数达到2×10^{5}、4×10^{5}和6×10^{5}时分别引入衰减，衰减系数为0.5.优化器动量参数设置为$\beta_1=0.9$和$\beta_2=0.99$.整个训练过程总计进行4×10^{5}次迭代.数据增强包括随机裁剪和随机水平翻转.所有图像均调整为512×512像素大小.

2.2 性能比较

在铁路扣件异常测试集上对比RainNet、PHDNet、CDTNet、DCCF、DiffHarmony＋＋和DB-CMAM的性能，4项关键指标(MSE、fMSE越低越好；PSNR、SSIM越高越好)的定量结果如表1所示.由表1可知：高分辨率方法要整体优于低分辨率方法，这是由于低分辨率方法生成结果在上采样后会存在模糊；在高分辨率情况下，扩散模型方法(DiffHarmony＋＋与DB-CMAM)整体优于传统U-Net方法(CDTNet与DCCF)——扩散模型方法组的平均MSE(1.18)与平均fMSE(55.83)较UNet方法组(MSE 3.40，fMSE 170.01)分别降低65.3%与67.2%，同时平均PSNR(49.52)提升7.0%，验证了扩散模型在图像和谐化任务中要优于UNet方法；进一步地，DB-CMAM与DiffHarmony＋＋相比，其MSE(0.65)与fMSE(39.85)分别降低61.9%与44.5%，验证所提双分支机制和前背景内容匹配机制的有效性.

针对各模型生成效率进行评估.具体而言，在

表 1　定量结果比较

Tab.1　Comparison of quantitative results

方法	MSE	fMSE	PSNR	SSIM
RainNet	7.775 4	310.237 5	43.798 2	0.990 6
PHDNet	6.573 7	298.013 4	44.040 0	0.992 7
CDTNet	5.026 2	246.934 3	44.578 2	0.997 8
DCCF	1.781 2	93.092 7	47.937 4	0.997 9
DiffHarmony＋＋	1.711 0	71.799 2	47.110 2	0.997 1
DB-CMAM	**0.652 3**	**39.851 4**	**51.921 2**	**0.998 5**

相同硬件环境与推理设置下,对不同方法在测试集上的平均单张图像生成时间进行统计与对比,结果显示,CDTNet、DCCF、DiffHarmony＋＋以及DB-CMAM方法的单张图像平均生成时间分别为0.15 s、0.10 s、3.25 s和3.10 s.从结果上来看,由于扩散模型(DiffHarmony＋＋与DB-CMAM)需要多步推理,所以其生成速度要慢于其他模型(CDTNet与DCCF);而DiffHarmony＋＋由于在扩散模型后级联了一个微调网络,其生成效率略慢于DB-CMAM.

同时,为直观评估不同方法在工务巡检扣件图像和谐化任务中的视觉表现,本文从测试集中随机选取若干具有代表性的样本进行主观评价,让被试者在真实图像与网络生成图像中选择看起来最和谐的一张图像.其中所选样本涵盖了不同轨面纹理、光照条件及扣件类型的典型场景,以保证评价结果的全面性 . RainNet、PHDNet、CDTNet、DCCF、DiffHarmony＋＋以及DB-CMAM方法得票率分别为0.20、0.19、0.26、0.38、0.46和0.62.从结果可以看出,低分辨率方法由于生成区域存在明显模糊现象,整体和谐程度较差,因而得票率较低;相比之下,各类高分辨率方法在视觉质量上均有显著提升.其中,DB-CMAM在主观评价中表现最优,且其得票率已超过0.5.

此外,进一步引入下游目标检测任务,对所提出方法的实际应用价值进行了验证.选用轻量级检测器YOLOv8n与两阶段检测器Faster R-CNN作为代表性模型,对基于DB-CMAM生成的图像结果进行扣件检测,结果如表2所示.使用平均精度均值(mean Average Precision,mAP)、召回率(Recall)、精确率(Precision)进行性能评价.mAP50-95是在交并比(Intersection over Union,IoU)阈值为0.50,0.55,…,0.95的多个条件下分别计算平均精度(Average Precision,AP),并对所得AP取平均,用于综合评价目标检测模型的检测精度与定位性能;mAP50是在IoU阈值为0.50时计算各类别AP,再对所有类别的AP取平均值,用于评价目标检测模型在较宽松定位条件下的整体检测性能;Recall为实际存在的目标中被模型正确检测出来的目标所占比例,用于衡量模型对目标的检出能力,召回率越高,说明模型漏检的目标越少;Precision表示模型检测结果中正确检测目标所占的比例,用于衡量模型检测结果的准确程度,精确率越高,说明模型产生的误检越少.由表2可知,DB-CMAM生成的异常扣件几乎和真实采集的异常扣件无异.

表 2　下游检测任务验证

Tab.2　Validation using downstream detection tasks

检测器	mAP50-95	mAP50	Recall	Precision
Fasterrcnn	0.840	0.950	0.886	0.950
Yolov8n	0.977	0.995	0.991	0.999

最后,在模型主观评价方面进行详细实验,并与之前提出的高分辨率图像和谐化方法进行可视化结果比较,如图3所示(清晰大图见文献[25]).图3中,Composite表示初始输入的合成图像,GT表示真实图像.由图3可知,DB-CMAM在视觉效果上优于现有方法,能够更自然地生成异常区域,整体光照更加协调.

2.3　消融实验

为验证所提出的双分支机制和内容匹配注意力模块的有效性,进行一系列消融实验.

在模型结构改进上,相关消融实验结果如表3所示.表3中:基础模型表示同时移除重建分支和内容匹配注意力模块后的模型;＋重建分支表示在基础模型上仅引入重建分支;＋内容匹配注意力表示仅引入内容匹配注意力模块;完整模型表示同时引入两个模块后的最终模型.由表3可知:①基础模型的性能相对较低,其MSE、fMSE、PSNR和SSIM分别为1.130 4、73.938 8、49.633 5和0.998 4,说明仅依靠基础扩散框架难以充分利用背景结构信息实现高质量图像和谐化;在加入重建分支后,模型性能得到明显提升,MSE和fMSE分别降低至0.908 4和58.372 8,PSNR提升至50.603 4,表明重建分支能够为和谐化分支提供更加稳定的特征指导,提高模型的噪声预测能力.②进一步分析内容匹配注意力模块的作用可以发现,仅引入该模块后,模型的MSE和fMSE分别降低至1.059 1和64.267 1,PSNR提升至49.823 6,相比基础模型具有一定改善,验证了利用背景中与前景结构相似区域作为参考信息能够有效提升图像和谐化效果;相比重建分支,内容匹配注意力模块对前景区域的特征匹配具有更直接的作用,能够减少无关背景区域对生成过程的干扰.③当同时引入重建分支和内容匹配注意力模块后,模型取得最佳性能,MSE、fMSE、PSNR

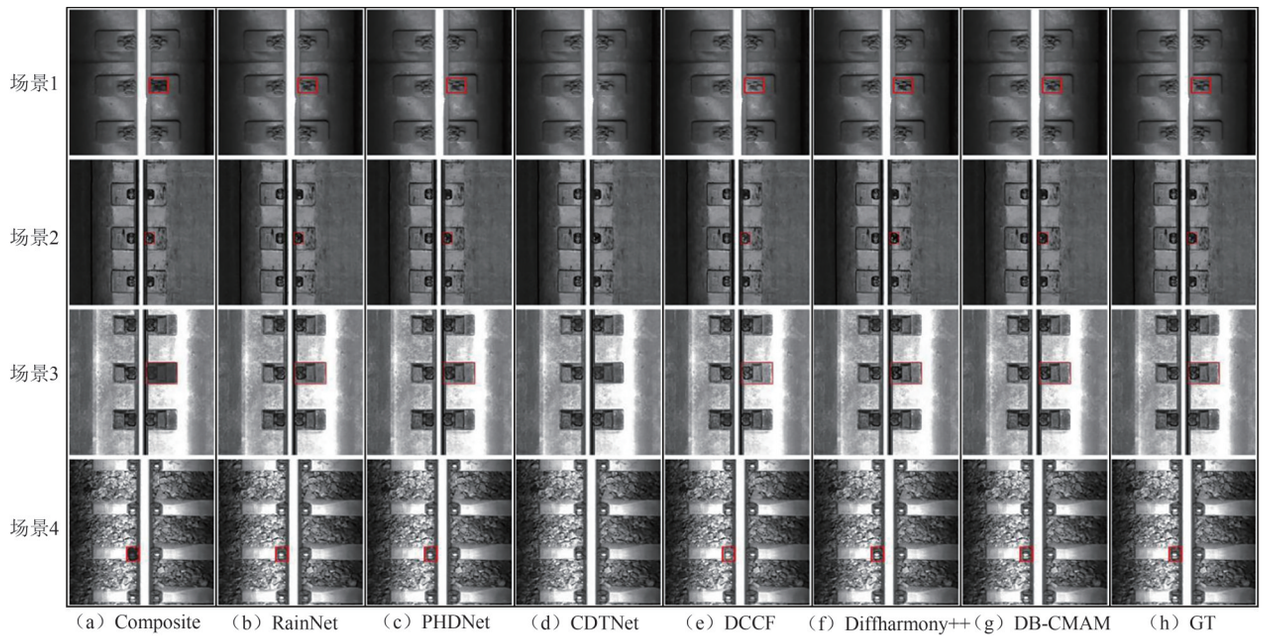


图 3 可视化结果比较

Fig.3 Visual comparison of the results

表 3 移除各模块后模型性能比较

Tab.3 Performance comparison after removing individual modules

实验设置	MSE	fMSE	PSNR	SSIM
基础模型	1.130 4	73.938 8	49.633 5	0.998 4
＋重建分支	0.908 4	58.372 8	50.603 4	0.998 4
＋内容匹配注意力	1.059 1	64.267 1	49.823 6	0.998 3
完整模型	**0.652 3**	**39.851 4**	**51.921 2**	**0.998 5**

和 SSIM 分别达到 0.652 3、39.851 4、51.921 2 和 0.998 5. 相比基础模型，完整模型的 MSE 和 fMSE 分别降低 42.29% 和 46.10%，说明两个模块之间具有良好的互补性. 其中，内容匹配注意力模块增强了模型对有效背景参考区域的利用能力，而重建分支进一步提升了扩散过程中的特征表达稳定性，二者协同共同提高了图像和谐化质量.

在损失函数设计上，相关消融实验结果如表 4 所示. 表 4 中：基础损失表示未引入额外损失约束的模型；＋局部损失 L_{local} 表示在基础损失上加入局部区域约束后的模型；＋特征重建损失 L_{fea} 表示仅加入特征重建约束后的模型；完整损失函数表示同时引入局部损失和特征重建损失后的最终模型.

表 4 有关损失函数的消融研究

Tab.4 Ablation study of the loss functions

设置	MSE	fMSE	PSNR	SSIM
基础损失	10.017 4	720.706 3	43.625 9	0.997 5
＋局部损失 L_{local}	1.130 4	73.938 8	49.633 5	**0.998 4**
＋特征重建损失 L_{fea}	4.038 7	293.659 0	47.735 3	0.998 0
完整损失函数	**0.908 4**	**58.372 8**	**50.603 4**	**0.998 4**

由表 4 可知：①基础损失模型性能较差，其 MSE 和 fMSE 分别达到 10.017 4 和 720.706 3，PSNR 仅为 43.625 9，说明仅依靠基础扩散目标难以有效约束前景区域的和谐化过程；由于铁路扣件图像中前景区域占比较小，全局优化过程容易受到背景区域影响，模型难以充分关注异常扣件区域的颜色、纹理及结构一致性. ②在基础损失上加入局部损失 L_{local} 后，模型性能得到显著提升，MSE 由 10.017 4 降低至 1.130 4，fMSE 由 720.706 3 降低至 73.938 8，PSNR 提升至 49.633 5. 该结果表明，局部损失能够增强模型对前景扣件区域的关注程度，使优化过程更加聚焦于目标区域，从而有效提升图像和谐化效果. ③进一步分析特征重建损失 L_{fea} 的作用可以发现，仅加入该损失后，模型的 MSE、fMSE 和 PSNR 分别达到 4.038 7、293.659 0 和 47.735 3，相比基础损失模型均有所改善. 这说明特征重建损失能够约束重建分支与和谐化分支之间的特征一致性，为噪声预测过程提供有效指导，提高扩散模型生成过程的稳定性. ④当同时引入局部损失 L_{local} 和特征重建损失 L_{fea} 后，模型取得最佳性能，MSE、fMSE、PSNR 和 SSIM 分别达到 0.908 4、58.372 8、50.603 4 和 0.998 4. 相比仅使用局部损失的模型，完整损失函数进一步降低了 MSE 和 fMSE，并提升了 PSNR，说明两种损失具有良好的互补作用. 其中，局部损失增强了模型对前景区域的约束能力，特征重建损失提高了扩散过程中特征传递的稳定性，二者共同促进了图像和谐化质量的提升.

同时,对各损失函数项的权重设置进行系统的消融实验,以分析不同损失约束在模型训练过程中的作用及其对生成结果的影响,实验结果如表5所示.表5中:均衡权重配置表示将全局损失L_{whole}、局部损失L_{local}和特征重建损失L_{fea}的权重均设置为1;优化权重配置表示将全局损失权重λ_{whole}设置为1,局部损失权重λ_{local}设置为500,特征重建损失权重λ_{fea}设置为0.03.

表 5　损失函数权重消融实验

Tab.5　Ablation study of loss-function weights

设置	MSE	fMSE	PSNR	SSIM
均衡权重配置	10.452 2	482.680 5	46.898 0	0.996 4
优化权重配置	**0.652 3**	**39.851 4**	**51.921 2**	**0.998 5**

由表5可知:采用均衡权重配置时,模型性能相对较低,其MSE、fMSE、PSNR和SSIM分别为10.452 2、482.680 5、46.898 0和0.996 4;当采用优化权重配置后,模型性能得到显著提升,MSE和fMSE分别降低至0.652 3和39.851 4,PSNR和SSIM分别提升至51.921 2和0.998 5.该结果表明,不同损失项在模型优化过程中承担的作用并不相同,简单设置相同权重难以充分发挥各损失项的约束作用.

具体而言,当各损失权重均设置为1时,由于特征重建损失L_{fea}在数值尺度上较大,其梯度贡献容易占据优化过程的主导地位,而用于约束前景区域生成质量的局部损失L_{local}所产生的梯度影响相对较弱,模型难以充分关注异常扣件区域的细节恢复.通过提高局部损失权重并降低特征重建损失权重,可以增强模型对前景区域的约束能力,同时保持重建分支对噪声预测过程的有效指导.因此,合理的损失函数权重配置对于平衡不同优化目标、提升图像和谐化性能具有重要作用.

3　结论

1)提出一种基于内容匹配注意力机制的双分支图像和谐化方法(DB-CMAM).针对扩散模型应用于图像和谐化任务时存在的"去噪目标"与"和谐化目标"相互耦合导致优化困难的问题,设计双分支网络结构,通过重建分支学习稳定的噪声预测能力,并利用特征约束引导和谐化分支训练,有效提升了模型的生成质量.

2)设计一种内容匹配注意力模块,通过挖掘背景区域中与前景异常扣件具有结构相似性的有效参考区域,增强前景与背景之间的特征交互能力,减少无关轨面区域对生成过程造成的干扰,使模型能够更加准确地恢复异常扣件区域的纹理、颜色和光照信息.

3)在自建铁路异常扣件图像数据集上的实验结果表明,DB-CMAM相比现有主流图像和谐化方法,在均方误差(MSE)和前景均方误差(fMSE)两个关键指标上分别降低了14.37%和8.91%,同时在PSNR和SSIM指标上取得最优结果,验证了所提出方法在生成质量和视觉一致性方面的有效性.

4)通过下游异常扣件检测实验进一步验证了生成图像的应用价值.实验结果表明,DB-CMAM生成的异常扣件图像能够在不同检测器中实现有效检出,说明所生成图像不仅具有较高的视觉真实性,同时能够满足铁路异常扣件检测任务对数据有效性的需求,为构建高质量铁路巡检数据集提供了可靠支持.

尽管所提出方法在铁路异常扣件图像和谐化任务中取得了较好的效果,但仍存在进一步优化空间.未来工作主要从以下几个方面展开:首先,将继续优化所提出的图像和谐化方法,探索更加有效的特征交互机制,以进一步提升复杂场景下生成图像的真实性和稳定性;然后,考虑将该方法推广应用于更多类型的铁路图像生成任务,为铁路智能巡检系统提供更加丰富的数据支持;最后,针对当前扩散模型推理过程较为耗时的问题,未来将研究模型蒸馏、采样加速等优化策略,在保证生成质量的同时降低模型推理成本,提高方法在实际铁路巡检场景中的应用效率.

参考文献(References):

[1] 武福, 蒋鹏民, 李忠学, 等. 基于轻量化YOLOv8s的轨道扣件状态检测方法. 北京交通大学学报, 2024, 48(5): 59－68.
Fu Wu, Pengmin Jiang, Zhongxue Li, et al. Lightweight detection method for track fastener status based on improved YOLOv8s. Journal of Beijing Jiaotong University, 2024, 48(5): 59－68.

[2] 陈修忻. 基于改进YOLOv5s算法的钢轨扣件状态检测方法. 城市轨道交通研究, 2024, 27(增1): 147－151.
Chen Xiuxin. Rail fastener status detection method based on improved YOLOv5s algorithm. Urban Mass Transit, 2024, 27(S1): 147－151.

[3] Zheng X Y, Zhang L F, Luo Y H, et al. A generation-based defect detection system for rail transit infrastructure[J]. High-speed Railway, 2026, 4(1): 1－9.

[4] Tsai Y H, Shen X H, Lin Z, et al. Deep image harmonization[C]//2017 IEEE Conference on Computer Vision and

Pattern Recognition (CVPR). Honolulu, 2017: 2799－2807.

[5] Cong W Y, Zhang J F, Niu L, et al. DoveNet: deep image harmonization via domain verification [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, 2020: 8391－8400.

[6] Ling J, Xue H, Song L, et al. Region-aware adaptive instance normalization for image harmonization [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, 2021:9357－9366.

[7] Cong W Y, Niu L, Zhang J F, et al. Bargainnet: background-guided domain translation for image harmonization[C]//2021 IEEE International Conference on Multimedia and Expo (ICME). Shenzhen, 2021: 1－6.

[8] Sofiiuk K, Popenova P, Konushin A. Foreground-aware semantic representations for image harmonization [C]// 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). Waikoloa, 2021: 1619－1628.

[9] Hang Y C, Xia B, Yang W M, et al. SCS-co: self-consistent style contrastive learning for image harmonization[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, 2022: 19678－19687.

[10] Guo Z H, GU Z R, Zheng B, et al. Transformer for image harmonization and beyond [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(11): 12960－12977.

[11] Guo Z H, Guo D S, Zheng H Y, et al. Image harmonization with transformer[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, 2021: 14850－14859.

[12] Cong W Y, Tao X H, Niu L, et al. High-resolution image harmonization via collaborative dual transformations [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, 2022: 18449－18458.

[13] Xue B, Ran S H, Chen Q, et al. DCCF: deep comprehensible color filter learning framework forHigh-resolution image harmonization [C]//Computer Vision-ECCV 2022. Cham, 2022: 300－316.

[14] Ke Z H, Sun C Y, Zhu L, et al. Harmonizer: learning toPerform white-box image andVideo harmonization [C]//Computer Vision-ECCV 2022. Cham: Springer, 2022: 690－706.

[15] Liang J T, Cun X D, Pun C M, et al. Spatial-separated curve rendering network forEfficient andHigh-resolution image harmonization [C]//Computer Vision-ECCV 2022. Cham: Springer, 2022: 334－349.

[16] Zhou P F, Feng F X, Wang X J. DiffHarmony: latent diffusion model meets image harmonization [C]//Proceedings of the 2024 International Conference on Multimedia Retrieval. Phuket Thailand. ACM, 2024: 1130－1134.

[17] Zhou P F, Feng F X, Liu G, et al. DiffHarmony++: enhancing image harmonization with harmony-VAE and inverse harmonization model [C]//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne VIC Australia. ACM, 2024: 10592－10601.

[18] Cao J Y, Hong Y, Niu L. Painterly image harmonization in dual domains[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(1): 268－276.

[19] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models [C]// Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS). Vancouver, 2020: 6840－6851.

[20] Song Y, Ermon S. Score-based generative modeling through stochastic differential equations [C]// Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS). Virtual, 2020: 1856－1866.

[21] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, 2022: 10674－10685.

[22] Ramesh A, Dhariwal P, Nichol A, et al. Hierarchical text-conditional image generation with CLIP latents [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, 2023: 10242－10252.

[23] Watanabe T, Chen W, Hedman P, et al. Dprecon: Diffusion prior for generalizable neural reconstruction [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, 2023: 2664－2674.

[24] Luo Z, Gustafsson F K, Zhao Z, et al. Image Restoration with Mean-Reverting Stochastic Differential Equations [C]// Proceedings of the 40th International Conference on Machine Learning (ICML). Hono lulu, 2023: 23045－23066.

[25] Luo W. DB-CMAM [EB/OL]. (2026-03-08)[2026-03-10]. https://github.com/luo-0728/DB-CMAM.