中图法分类号:TP391.41　　**文献标识码**:A　　**文章编号**:1006-8961(XXXX)XX-0001-15

论文引用格式:Wang Tonghan, Li Xianing, Jia Huizhen, Wu Jiasong. Diffusion prior multi-scale hierarchical adaptive no-reference image quality assessment[J/OL]. Journal of Image and Graphics, XXXX:1-15. DOI: 10.11834/jig.260455. (王同罕, 李夏宁, 贾惠珍, 伍家松. 扩散先验多尺度层级自适应无参考图像质量评价[J/OL]. 中国图象图形学报,XXXX:1-15. DOI: 10.11834/jig.260455.)[DOI:10.11834/jig.260455]

扩散先验多尺度层级自适应无参考图像质量评价

王同罕[1],李夏宁[1],贾惠珍[1*],伍家松[2]

1. 东华理工大学江西省核地学数据科学与系统工程技术研究中心,南昌　330013; 2. 东南大学计算机科学与工程学院,南京　211189

摘　要:**目的**　真实场景图像失真呈现多尺度与空间非均匀特性,现有无参考图像质量评价方法难以同步捕获全局语义与局部纹理,且多阶段特征融合策略固定,导致预测精度、主观一致性及泛化性能受限。**方法**　提出一种基于扩散先验的多尺度层级自适应无参考图像质量评价方法。以预训练扩散模型为骨干,首先通过变分自编码器(variational autoencoder,VAE)获取全局语义,并构建包含频域感知增强与跨尺度注意力融合的多级特征提取分支补偿高频细节损失;其次设计质量条件嵌入模块,将质量描述文本经文本适配器转换为条件嵌入以引导去噪;随后提出层级自适应质量解码器,通过层级自适应注意力融合机制动态聚合多阶段特征;最终经多层感知器回归质量分数。**结果**　在CLIVE、KonIQ-10k、LIVEFB和SPAQ四个数据集上的实验表明,所提方法的皮尔森线性相关系数(Pearson linear correlation coefficient,PLCC)和斯皮尔曼等级相关系数(Spearman rank-order correlation coefficient,SRCC)均为最优或次优:在KonIQ-10k上,PLCC达0.955、SRCC达0.946,较第2名提升0.63%和0.53%;在CLIVE上,PLCC达0.924、SRCC达0.907,较第2名提升1.43%和1.80%;在LIVEFB上,PLCC达0.706、SRCC达0.589,较第2名PLCC提升3.67%、SRCC提升1.90%。消融实验验证了多级特征提取分支及层级自适应融合机制的有效性,完整模型较基线PLCC和SRCC均有显著提升。泛化实验表明,本文方法在跨数据集迁移中平均性能表现良好。**结论**　本文方法通过全局语义与局部纹理的协同提取及内容驱动的自适应特征融合,有效提升了真实场景下质量预测的精度、主观一致性与泛化性能,为无参考图像质量评价提供了一种有效方案。本文所提方法的源代码及相关数据集已发布至https://github.com/ECUT-VQA/Diffusion-prior-multi-scale-hierarchical-adaptive-no-reference-image-quality-assessment/tree/master。

关键词:图像质量评价;扩散模型;真实失真;特征提取;层级融合

Diffusion prior multi-scale hierarchical adaptive no-reference image quality assessment

Wang Tonghan[1], Li Xianing[1], Jia Huizhen[1], Wu Jiasong[2]

1. Jiangxi Engineering Technology Research Center of Nuclear Geoscience Data Science and System, East China University of Technology, Nanchang 330013, China; 2. School of Computer Science and Engineering, Southeast University, Nanjing 211189, China

Abstract: **Objective** Real-world image distortions exhibit multi-scale and spatially non-uniform characteristics, which

收稿日期:2026-07-26;**修回日期**:2026-11-12

*通信作者:贾惠珍,女,副教授,主要研究方向为计算机视觉、模式识别。E-mail: jiahuizhen@ecut.edu.cn;伍家松,男,副教授,主要研究方向为模式识别、计算机视觉。E-mail: jswu@seu.edu.cn

基金项目:国家自然科学基金(62261001,62266001);江苏省创新支撑计划项目(BZ2023042);江西省研究生创新专项资金项目(YC2025-S544)

Supported by: National Natural Science Foundation of China (Grant No. 62261001, 62266001); the innovation project of Jiangsu Province under Grant BZ2023042; Jiangxi Provincial Graduate Student Innovation Special Fund Project(Grant No. YC2025-S544)

© 中国图象图形学报版权所有

pose significant challenges for no-reference image quality assessment (NR-IQA). Existing NR-IQA methods struggle to simultaneously capture global semantics and local textures due to inherent architectural limitations. Convolutional neural network (CNN) based methods suffer from restricted receptive fields, making them inadequate for modeling long-range dependencies and handling non-uniform distortions across scales. Transformer-based approaches, while capable of global semantic integration, lack sufficient sensitivity to fine-grained local distortions and incur high computational costs. Moreover, fixed multi-stage feature fusion strategies commonly adopted in existing methods cannot adapt to diverse image contents, leading to limited prediction accuracy, poor subjective consistency, and restricted cross-dataset generalization. To overcome these limitations, this study develops a novel NR-IQA method that leverages powerful visual priors of pre-trained diffusion models while establishing adaptive feature encoding and fusion mechanisms tailored to content characteristics. **Method** A multi-scale hierarchical adaptive NR-IQA method based on diffusion priors is proposed. The method employs a pre-trained text-to-image diffusion model (Stable Diffusion v1.5) as the backbone. The framework consists of three core modules: multi-scale feature extraction, quality conditional embedding, and hierarchical adaptive quality decoding. In the multi-scale feature extraction module, the input image is first encoded into latent space representations through a variational autoencoder (VAE) to obtain compact global semantic features. However, VAE encoding inevitably loses high-frequency textures, edge details, and local distortion information during downsampling. To compensate for this loss, a multi-level feature extraction branch (MFEB) is constructed in the pixel space, extracting multi-scale residual features aligned with the downsampling stages of the denoising U-Net. The MFEB comprises four cascaded adapter blocks with channel attention and spatial gating mechanisms. A frequency-aware enhancement module (FEM) is embedded after each adapter block, employing depthwise separable convolution with center-initialized kernels to isolate high-frequency residual components, effectively enhancing perception of blur and compression artifacts. Cross-scale attention fusion modules (CAFM) are introduced between adjacent adapter blocks to promote multi-scale information interaction through spatial and channel attention mechanisms. The extracted residual features are injected into the downsampling path of the U-Net as additive residuals. In the quality conditional embedding module, predefined quality description texts covering continuous quality levels are encoded by a frozen CLIP text encoder to obtain initial embeddings. A lightweight text adapter with residual connections is applied for adaptation. The adapted embeddings are replicated along the batch dimension and fed into cross-attention layers of the U-Net to guide conditional denoising in quality-aware feature space. In the hierarchical adaptive quality decoding module, feature maps from four stages of the U-Net upsampling path are unified to the same spatial resolution through bicubic interpolation and projected to 512-dimensional channel space. Squeeze-and-excitation (SE) modules refine the channels. A layer adaptive attention fusion (LAAF) mechanism dynamically allocates aggregation weights for the four stages based on global content descriptors. Global average pooling extracts descriptors from each stage, which are concatenated and processed by a lightweight multilayer perceptron to learn optimal hierarchical combinations. The generated weights are normalized by Softmax and applied for element-wise weighted fusion of refined stage features. The fused features undergo further convolution and channel refinement before quality score regression through a three-layer multilayer perceptron. **Result** Extensive experiments are conducted on four widely recognized real-world distortion datasets: CLIVE, KonIQ-10k, LIVEFB, and SPAQ The proposed method is compared against five state-of-the-art approaches: HyperIQA, SaTQA, LoDa, DP-IQA, and MDM-GFIQA. Evaluation metrics include Pearson linear correlation coefficient (PLCC) and Spearman rank-order correlation coefficient (SRCC), where higher absolute values indicate stronger correlation. Experimental results demonstrate optimal or suboptimal PLCC and SRCC across all datasets. On KonIQ-10k, PLCC reaches 0.955 and SRCC reaches 0.946, surpassing the second-best method (DP-IQA: PLCC 0.949, SRCC 0.941) by 0.63% and 0.53%. On CLIVE, PLCC reaches 0.924 and SRCC reaches 0.907, outperforming the second-best (DP-IQA: PLCC 0.911, SRCC 0.891) by 1.43% and 1.80%. On LIVEFB, PLCC reaches 0.706 and SRCC reaches 0.589, exceeding the second-best (DP-IQA: PLCC 0.681, SRCC 0.578) by 3.67% and 1.90%. On SPAQ, PLCC reaches 0.927 and SRCC reaches 0.924, comparable to the best-performing method LoDa (PLCC 0.928, SRCC 0.925) with gaps within 0.001 and superior to the remaining compared methods. Ablation experiments on KonIQ-10k and CLIVE verify synergistic effects of proposed modules. The baseline with frozen U-Net encoder achieves PLCC/SRCC of 0.949/0.941 on KonIQ-10k and 0.911/0.891 on CLIVE. Adding the frequency-aware enhancement (FA) module improves PLCC to 0.951

© 中国图象图形学报版权所有

and 0. 914 respectively. Adding FA with cross-scale attention (CSA) achieves 0. 953/0. 945 and 0. 921/0. 902. The complete model with layer adaptive attention fusion (LAAF) achieves best performance of 0. 955/0. 946 and 0. 924/0. 907. Cross-dataset generalization experiments among KonIQ-10k, CLIVE, and LIVEFB demonstrate favorable average cross-domain performance, with SRCC of 0. 781 for LIVEFB-to-KonIQ transfer and 0. 856 for KonIQ-to-CLIVE transfer. **Conclusion** This study proposes a multi-scale hierarchical adaptive NR-IQA method based on diffusion priors. By synergistically extracting global semantics through VAE latent encoding and compensating local high-frequency details through the multi-level feature extraction branch, and dynamically aggregating multi-stage features through the hierarchical adaptive attention fusion mechanism, the proposed method effectively improves prediction accuracy, subjective consistency, and cross-dataset generalization for real-world image quality assessment. Experimental results on multiple public datasets validate superiority over state-of-the-art approaches, and ablation studies confirm contributions of each core module. This work provides an effective solution for no-reference image quality assessment in real-world scenarios. The source code and associated datasets for the proposed method are publicly available athttps://github. com/ECUT-VQA/Diffusion-prior-multi-scale-hierarchical-adaptive-no-reference-image-quality-assessment/tree/master.

Key words: image quality assessment; diffusion model; authentic distortion; feature extraction; hierarchical fusion

0 引　言

随着移动互联网与智能终端的快速发展，视觉内容已成为信息传播的主要载体。图像质量评价(image quality assessment，IQA)作为衡量视觉感知体验的核心技术，在图像压缩、图像恢复、生成模型评估等任务中发挥着基础性作用(Zhang 等人，2025)。根据参考图像的可用性，IQA 方法通常分为全参考(full-reference，FR)、半参考(reduced- reference，RR)与无参考(no-reference，NR)三类(Zhai 和 Min，2020)。其中，无参考图像质量评价(no-reference image quality assessment，NR-IQA)无需依赖原始参考图像，仅通过待评价图像本身推断主观质量分数，贴合真实拍摄、社交平台传输等应用场景需求，成为当前 IQA 领域的研究热点(方玉明等人，2021；鄢杰斌 等人，2022；程茹秋 等人，2022)。

围绕 NR-IQA 任务，国内外学者开展了大量研究。早期方法基于自然场景统计理论，通过建模清晰自然图像的统计分布规律量化质量退化程度；随着深度学习的发展，基于卷积神经网络(convolutional neural networks，CNN)与 Transformer 的端到端方法显著提升了质量预测精度；近年来，大规模预训练模型凭借其从海量数据中学习的通用先验，为真实失真 NR-IQA 提供了新的解决思路。

现有研究表明，扩散模型先验在真实失真质量评价中具有独特优势。然而，真实场景图像的失真往往呈现多尺度、空间非均匀等复杂特性，对特征提取的充分性与融合的有效性提出了更高要求。一方面，VAE 编码在获得紧凑语义表示的同时会丢失高频细节(Rombach 等人，2022；Li 等人，2025)，需要补偿机制补充局部纹理与边缘信息；另一方面，去噪 U-Net 上采样过程输出的层级化特征蕴含丰富的多尺度质量线索，但如何根据内容特性自适应地聚合这些互补特征，是提升质量预测精度的关键。因此，充分挖掘扩散先验中的多尺度质量特征，并建立与内容特性相适应的特征编码与融合机制，对于真实场景下的 NR-IQA 具有重要意义。

据此，本文提出一种基于扩散模型的真实失真无参考图像质量评价方法。该方法以预训练文本到图像扩散模型为骨干，首先构建多尺度特征提取模块，通过 VAE 潜空间编码与多级特征提取分支的协同，在获取全局语义表示的同时补偿高频信息损失；其次设计质量条件嵌入模块，将预定义质量描述文本经文本适配器转换为条件嵌入，引导去噪 U-Net 进行条件化去噪；最后提出层级自适应质量解码模块，通过层级自适应注意力融合机制为不同阶段动态分配权重，实现内容驱动的选择性聚合与解码，经三层感知器回归图像质量分数。在多个真实失真数据集上的实验表明，所提方法能够有效提升质量预测精度与主观感知一致性，且泛化性表现良好，为真实场景下的无参考图像质量评价提供了一种有效的解决方案。

© 中国图象图形学报版权所有

1 相关工作

图像与视频质量评价研究历经多年发展，已形成较为完整的技术体系。程茹秋等人(2022)系统综述了图像与视频质量评价方法的研究进展，从主观评价与客观评价、全参考与无参考评价、图像评价与视频评价等维度对代表性方法进行了归纳与比较分析；Zhai和Min(2020)、方玉明等人(2021)与鄢杰斌等人(2022)也分别从感知质量评价、无参考质量评价及失真类型等角度对领域工作进行了梳理。本节按照技术演进脉络，从基于自然场景统计的方法、基于深度学习的方法与基于预训练先验的方法三个方面对NR-IQA相关工作进行梳理。

1.1 基于自然场景统计的无参考图像质量评价

早期NR-IQA方法主要基于自然场景统计(natural scene statistics，NSS)理论，通过建模清晰自然图像的统计分布规律，以分布偏移程度量化图像质量。Mittal等人(2012)在空间域提取统计特征构建了盲质量分析框架，Moorthy和Bovik(2011)则从变换域角度建立了从自然场景统计到感知质量的映射模型。这类方法在合成失真场景中效果良好，但依赖特定失真先验，对真实场景复杂失真适应性不足。

1.2 基于深度学习的无参考图像质量评价

基于深度学习的端到端特征学习能力为突破上述局限提供了新路径。Zhang等人(2018)设计双分支卷积神经网络建模图像内容与失真特征；Su等人(2020)引入自适应超网络架构动态调整质量预测参数；Zhu等人(2020)则从元学习角度出发提升模型的泛化能力。尽管CNN类方法大幅提升了质量评价性能，但其局部感受野特性导致模型难以捕捉长距离依赖关系，在应对非均匀失真、跨尺度失真时性能显著下降。

为弥补局部建模的短板，具有全局感知能力的Transformer架构被引入视觉质量评估任务。Dosovitskiy等人(2021)提出的视觉Transformer为全局结构建模提供新路径；在此基础上，Ke等人(2021)通过多尺度建模实现了分辨率自适应的质量评估；Yang等人(2022)设计多维注意力机制强化失真特征表达。然而，自注意力机制侧重全局语义整合，对局部细粒度失真感知能力不足，且计算复杂度较高，难以满足实际落地需求。

1.3 基于预训练先验的无参考图像质量评价

近年来，大规模预训练模型凭借其从海量数据中学习的通用先验，为真实失真NR-IQA提供了新的解决思路。视觉语言预训练模型展现出强大的零样本感知能力，Wang等人(2023)探索了对比语言-图像预训练(contrastive language-image pre-training，CLIP)在图像质量评价中的应用；Zhang等人(2023)则从视觉语言对应关系出发构建了多任务学习框架。但后续研究证实这类编码器对细粒度失真敏感性有限(Bianchi等人，2024)。另一方面，基于生成先验的方法展现出更强的表征潜力，Rombach等人(2022)提出的潜在扩散模型通过在潜在空间建模图像的逆向扩散过程获得高层语义与低层纹理的联合表征。基于此，Fu等人(2024)将预训练文本到图像扩散模型引入NR-IQA任务，探索了去噪网络层级化特征的有效性；Li等人(2025)进一步优化了感知先验的提取效率，通过特征去噪扩散模块增强了失真表征能力。

与上述直接利用预训练扩散先验的思路不同，Wang等人(2025)探索了扩散模型在NR-IQA中的另一种利用方式：基于自由能原理模拟人眼视觉系统的自修复机制，以IQA数据集中的参考图像为条件训练扩散恢复网络，通过迭代去噪采样生成失真图像的恢复结果及中间含噪图像，进而设计视觉补偿引导分支与视觉差异分析分支，将恢复图像作为"伪参考"进行差异化的质量评价，在多个合成失真数据集上取得了优异性能。该方法验证了扩散生成过程用于质量评价的可行性，但其扩散恢复网络需依赖参考图像进行训练，且依赖完整的迭代采样过程，计算开销较大；同时，受恢复网络对复杂真实失真恢复能力的限制，其在真实失真数据集上的性能提升相对有限。

综上，扩散模型在NR-IQA中的利用方式可分为两类：一类以恢复生成为中介，将扩散模型作为图像增强器提供伪参考信息；另一类以先验表征为核心，直接提取预训练去噪网络的层级化特征。需要指出的是，后一路线所称"扩散先验"，是指预训练扩散模型在其去噪网络参数中编码的视觉知识先验，而非完整的迭代扩散采样过程：已有研究表明(Chen等人，2025)，去噪扩散模型的表征能力主要来源于去噪器本身而非扩散采样过程，在单一时间步进行

© 中国图象图形学报版权所有

前向去噪即可充分激活先验表征(Fu等人,2024)。相比前者,该路线无需额外训练生成模型、无需迭代采样,且能充分利用大规模预训练模型从海量真实图像中学习的通用先验,更契合真实失真场景的需求。本文遵循后一技术路线,并针对其中的VAE高频信息损失与多尺度特征固定融合两方面问题展开研究。

2 模型设计

2.1 模型概述

本文所提质量评价流程如图1所示,该方法以预训练Stable Diffusion的去噪U-Net为特征提取骨干。首先,输入图像经VAE编码压缩为潜空间表示;在像素空间,通过多级特征提取分支提取多尺度残差特征并注入U-Net下采样路径,补偿VAE编码丢失的高频细节。与此同时,预定义的质量描述文本经文本适配器转换为条件嵌入,引导去噪U-Net进行条件化去噪。随后,从U-Net上采样路径四个阶段提取层级特征映射,送入多层级自适应质量解码器,通过层级自适应注意力融合机制为不同阶段动态分配权重,实现内容驱动的选择性聚合与解码。最后,融合特征经三层MLP回归图像质量分数。

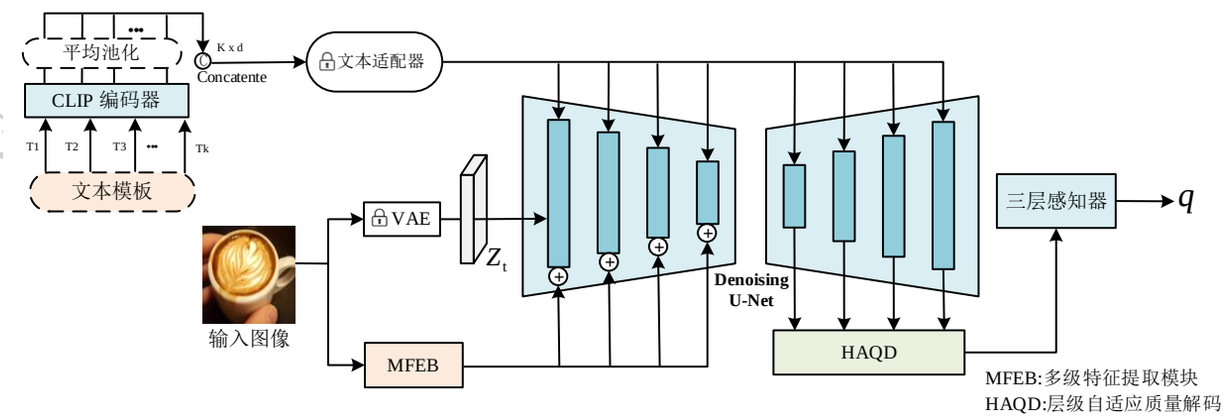


图1 质量评价流程图

Fig. 1 Flowchart of quality assessment

2.2 多尺度特征提取

在本文提出的评价方法中,为实现输入图像的多尺度特征表示,构建多尺度图像特征提取模块。该模块包含两个互补分支:鉴于Stable Diffusion的去噪U-Net需以潜变量为输入,输入图像首先经预训练VAE压缩为潜空间表示,提取全局语义特征以适配扩散模型的计算域;然而,VAE的下采样编码不可避免地导致高频纹理、边缘细节及局部失真信息的损失,仅依靠潜空间单尺度特征难以满足质量评估对局部细节的需求。因此,在像素空间同步构建多级特征提取分支(multi-level feature extraction branch,MFEB),提取与U-Net下采样阶段对齐的多尺度残差特征并注入下采样路径,从而补充VAE编码丢失的高频细节。通过潜空间全局特征与像素空间多尺度局部特征的协同,形成从粗到细的层级化特征输入,为后续条件化去噪提供完整的特征基础。

2.2.1 VAE潜空间编码

输入图像$\boldsymbol{x}$先经预训练VAE编码器$\mathcal{E}(\cdot)$压缩为低维潜变量表示,该过程可表示为:

$$z = \mathcal{E}(\boldsymbol{x}) \cdot s \tag{1}$$

式中,s为缩放因子,z为VAE编码得到的潜变量。VAE通过多层下采样将高维图像映射至潜空间,获得紧凑的全局语义表示,作为去噪U-Net的主输入。

2.2.2 多级特征提取

为弥补VAE编码过程中的信息损失,引入多级特征提取分支(MFEB)。与VAE在潜空间操作不同,MFEB直接在原始像素空间对输入图像进行多尺度特征提取,通过构建与U-Net下采样阶段相对应的多个特征提取层级,捕获VAE编码过程中丢失的细粒度纹理与局部结构信息。

MFEB的整体结构如图2所示。输入图像首先经像素反重排(pixel unshuffle)操作进行空间维度的

© 中国图象图形学报版权所有

下采样与通道维度的扩展，下采样因子为8。该操作是Shi等人(2016)提出的亚像素卷积中像素重排(pixel shuffle)的逆变换，可在不损失信息的前提下将空间信息转换至通道维度；由于该操作直接作用于原始像素图像而非经上采样重建的特征图，故处理流程中无需先执行像素重排。经像素反重排后的特征在空间分辨率上与VAE潜变量严格对齐，为与U-Net下采样各阶段特征的逐层对齐注入奠定基础。随后，特征经初始卷积层映射至特征空间；在此基础上构建四级级联的适配器模块(adapter block)，沿空间维度逐层下采样，依次输出通道数为320、640、1280、1280的多尺度特征，其空间分辨率与通道维度均与去噪U-Net下采样各阶段严格对齐。

适配块内部以条件化残差块(res blocks)作为基础特征提取单元，采用双卷积结构，通过3×3卷积提取局部特征，经ReLU激活后由1×1卷积进行通道投影；并引入挤压激励(squeeze-and- excitation，SE)模块(Hu等人，2018)与空间门控机制(Woo等人，2018)，分别对通道重要性与空间显著性进行自适应调制，最终通过跳跃连接与输入特征相加形成残差输出。

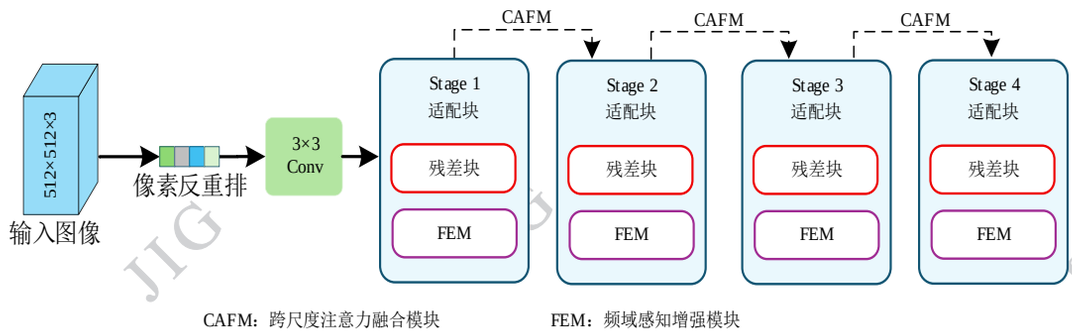


图2　多级特征提取分支

Fig. 2　Multi-level feature extraction branch

在此基础上，为进一步增强对频域失真的感知能力，本文在每个适配块的残差输出后嵌入频域感知增强模块(frequency-aware enhancement module，FEM)，其结构如图3所示。该模块的核心滤波操作记为频域感知卷积(frequency-aware convolution，FreqConv)，其采用深度可分离卷积对特征进行滤波，卷积核通过中心初始化策略构建，即中心位置权重置1、其余位置置0，实现恒等映射。输入特征$\boldsymbol{h}$经该滤波器处理后，减去其空间全局均值以隔离高频残差分量，得到频域响应$\mathcal{F}(\boldsymbol{h})$；随后经可学习增益参数$\alpha$调制，以残差形式注入主路径，实现对模糊、压缩等引入的频域损失的有效增强。该过程可表示为：

$$\boldsymbol{h}' = \boldsymbol{h} + \alpha \cdot \mathcal{F}(\boldsymbol{h}) \tag{2}$$

式中，$\boldsymbol{h}$为残差块输出特征，$\mathcal{F}(\cdot)$表示频域感知滤波，α为可学习增益系数。

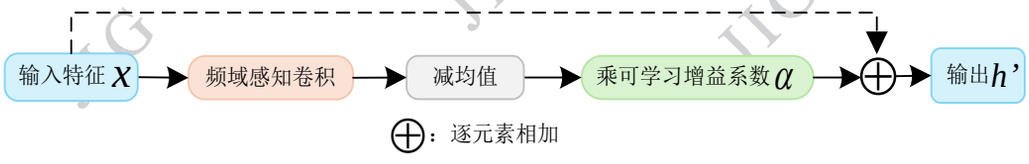


图3　频域感知增强模块

Fig. 3　Frequency-aware enhancement module

此外，为打破各级特征独立处理的局限性并促进多尺度特征间的信息交互，本文采用跨尺度注意力机制(Chen等人，2024)，在相邻适配块之间引入跨尺度注意力融合模块(cross-scale attention fusion module，CAFM)，其结构如图4所示。对于第i个适配块，前级输出特征$\boldsymbol{f}_{i-1}$首先经1×1卷积投影至当前阶段通道维度，并通过双线性插值进行空间分辨率对齐，得到对齐特征$\hat{\boldsymbol{f}}_{i-1}$；当前阶段残差输出$\boldsymbol{h}_i'$则并

© 中国图象图形学报版权所有

行送入双分支注意力网络，分别生成空间注意力图与通道权重向量。其中，空间注意力分支通过卷积网络生成单通道空间注意力图$\boldsymbol{A}_{\mathrm{s}} \in \mathbf{R}^{1\times \mathrm{H}_i\times \mathrm{W}_i}$，对前级对齐特征$\hat{\boldsymbol{f}}_{i-1}$进行空间选择性增强；通道注意力分支通过全局平均池化与多层感知器生成通道权重向量$\boldsymbol{w}_{\mathrm{c}} \in \mathbf{R}^{1\times 1\times C_i}$，对当前特征$\boldsymbol{h}_i'$进行通道重要性校准。前级特征经空间注意力调制后与当前特征经通道权重调制的结果逐元素相加，得到跨尺度融合信息，再经可学习增益参数β调制，以残差形式与当前阶段特征$\boldsymbol{h}_i'$相加，实现跨层级的语义互补。该过程可表示为：

$$\boldsymbol{f}_i = \boldsymbol{h}_i' + \beta \cdot \left(\hat{\boldsymbol{f}}_{i-1} \odot \boldsymbol{A}_{\mathrm{s}} + \boldsymbol{h}_i' \odot \boldsymbol{w}_{\mathrm{c}}\right) \tag{3}$$

式中，β为可学习增益系数，$\odot$表示逐元素乘法。首阶段不存在前级特征，该模块不参与计算。

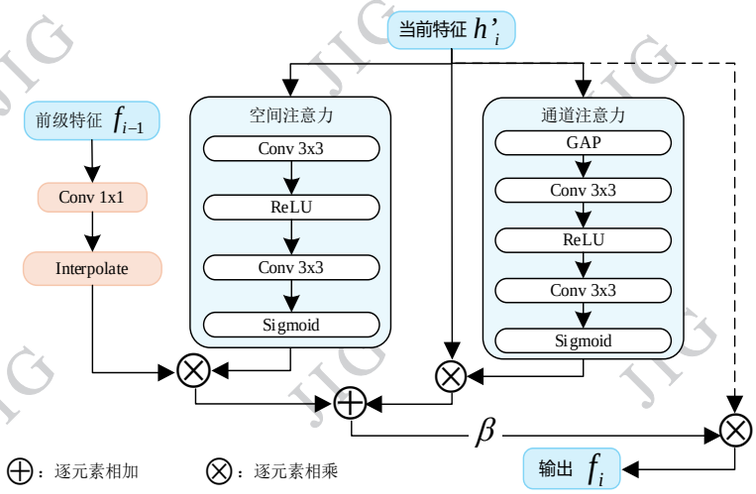


图4　跨尺度注意力融合模块

Fig. 4　Cross-scale attention fusion module

2.2.3　特征注入

在特征注入阶段，MFEB 提取的多尺度残差特征以附加残差形式注入去噪 U-Net 的下采样路径。对于 U-Net 第个下采样块的输出，注入后的特征表示为：

$$\boldsymbol{h}_i' = \boldsymbol{h}_i + \boldsymbol{r}_i \tag{4}$$

式中，$\boldsymbol{h}_i$为第i个下采样块的输出特征，$\boldsymbol{r}_i$为注入的残差特征。这种旁路残差注入机制使 U-Net 在去噪过程中能够同时利用 VAE 提供的全局语义压缩表示与 MFEB 补充的局部高频细节，从而在潜空间语义推理与像素空间纹理保留之间建立互补。通过 VAE 编码与 MFEB 补偿的协同作用，该模块为后续条件化去噪与质量解码提供了完整的输入表示。

2.3　文本条件嵌入

在本文提出的评价方法中，为使去噪 U-Net 能够感知图像质量语义并引导其向质量相关的特征空间进行条件化推理，设计质量条件嵌入模块，将预定义的质量描述文本转换为条件嵌入并注入 U-Net 的交叉注意力机制。具体地，采用质量描述文本模板（Fu 等人，2024）"a photo of a {scene} with {distortion type} distortion, which is of {quality level} quality" 构建条件文本，其中场景类别涵盖 animal、cityscape、human、indoor、landscape、night、plant、still life 等 11 类典型场景，失真类型涵盖 realistic blur、realistic noise、overexposure、underexposure、JPEG compression 等 35 种常见失真，质量等级划分为 bad、poor、fair、good、perfect 五级，三者共组合生成 1925 条质量描述文本。例如，"a photo of an animal with realistic blur distortion, which is of bad quality" 即描述一幅含真实模糊失真的低质量动物图像。采用恒定条件嵌入策略：全部描述文本同时经预训练 CLIP 文本编码器（Radford 等人，2021）编码，各句的令牌嵌入经平均池化得到句子级嵌入，全部句子嵌入拼接构成统一的条件嵌入，使同一条件嵌入适用于所有输入图像，无需针对单幅图像定制文本内容。该文本编码器在训练过程中保持冻结，以利用其在大规模图文数据上预训练的语义表征能力。获得初始文本嵌入后，引入文本适配器（text adapter）（Gao 等人，

© 中国图象图形学报版权所有

2024)对其进行轻量适配。

其中,text adapter 由一个两层全连接网络构成,以残差形式对初始嵌入进行调制,该过程可表示为:

$$\boldsymbol{t} = \boldsymbol{t}_0 + \gamma \cdot \mathrm{MLP}(\boldsymbol{t}_0) \tag{5}$$

式中,$\boldsymbol{t}_0 \in \mathbf{R}^{1\times D}$为 CLIP 文本编码器输出的初始质量文本嵌入,$D$为嵌入维度,MLP(·)为两层全连接网络,中间层采用高斯误差线性单元(Gaussian error linear unit,GELU)激活函数;γ为可学习调制参数,用于控制适配残差的注入强度;$\boldsymbol{t} \in \mathbf{R}^{1\times D}$为适配后的质量条件嵌入。随后,将适配后的条件嵌入沿批次维度复制扩展,与图像批次对齐,作为去噪 U-Net 交叉注意力层的条件输入,引导 U-Net 在质量感知空间中进行条件化去噪。通过将质量语义显式编码为条件嵌入,该模块使扩散先验能够关注与图像质量相关的语义特征,为后续质量解码提供质量感知的特征基础。

2.4 层级自适应质量解码

经 U-Net 条件化去噪后,其上采样路径四个阶段依次输出层级特征映射。由于各阶段特征在分辨率、语义粒度及通道维度上存在显著差异,若采用固定维度的直接拼接进行融合,难以根据输入内容自适应地调节各阶段对质量评估的动态贡献,导致多尺度质量线索的利用效率受限。现有工作表明,在真实失真质量评价中,对不同图像内容统一分配各尺度权重易引入误导性背景信息,而依据全局内容动态调节多尺度特征的聚合比例,能够有效提升关键质量信息的利用效率(Chen 等人,2025)。因此,为充分利用多层级特征的互补信息,本文设计层级自适应质量解码器(hierarchical adaptive quality decoder,HAQD),通过内容驱动的选择性聚合与特征解码,实现从多尺度扩散先验到质量分数的映射,解码器整体结构如图 5 所示,图中两处重复标记的含义分别为:层级特征处理支路的"×4"表示 U-Net 上采样路径四个阶段的特征映射各自独立地经过相同的并行处理单元;回归支路的"×3"表示由 1×1 卷积与激活函数构成的降维单元级联重复 3 次,对应质量分数回归阶段的三层逐层降维操作。

层级自适应质量解码器首先将四阶段特征映射通过双三次插值统一至相同空间分辨率 64×64,并分别经 3×3 卷积投影至 512 维统一通道空间,再依次通过 SE 模块进行通道精炼,筛选各阶段中对质量敏感的通道响应,得到精炼后的层级特征$\boldsymbol{x}_i \in \mathbf{R}^{512\times 64\times 64}$,其中$\boldsymbol{x}_1$对应 U-Net 最浅层上采样输出,$\boldsymbol{x}_4$对应最深层输出。

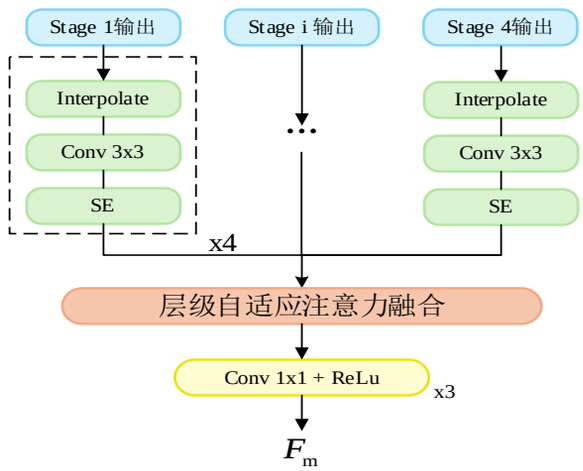


图 5 层级自适应质量解码器

Fig. 5 Hierarchical adaptive quality decoder

为实现内容驱动的层级选择性聚合,本文在解码器中引入层级自适应注意力融合(layer adaptive attention fusion,LAAF)机制。该机制的核心思想是基于全局内容描述符为四个阶段动态分配聚合权重,具体细节如图 6 所示。

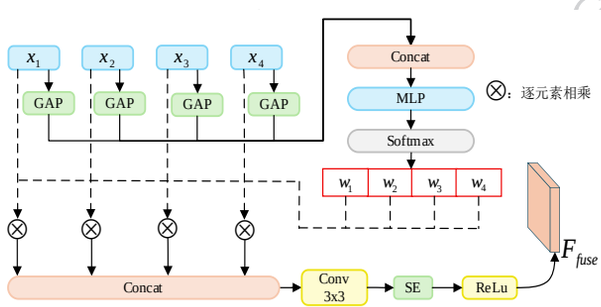


图 6 层级自适应注意力融合

Fig. 6 Layer adaptive attention fusion

具体而言,首先对各阶段特征$\boldsymbol{x}_i$执行全局平均池化(global average pooling,GAP),将空间维度压缩为 1,提取编码该阶段全局统计响应强度的内容描述符$\boldsymbol{g}_i \in \mathbf{R}^{512}$。该描述符包含了对应阶段通道激活的均值分布与整体能量信息,能够反映该尺度下特征对当前图像内容的响应强度。随后,将四阶段描述符沿通道维度拼接,形成联合全局表示$\boldsymbol{G} \in \mathbf{R}^{2048}$。进而,将联合全局表示$\boldsymbol{G}$输入轻量多层感知器,通过两级线性映射学习从全局内容到最优层级组合的决

© 中国图象图形学报版权所有

策函数，输出4维层级权重向量，并经Softmax函数归一化，确保权重和为1，实现层级选择，该过程可表示为：

$$\boldsymbol{w} = Softmax\left(\mathrm{MLP}(\boldsymbol{G})\right) \in \mathbf{R}^4, \sum_{i=1}^{4} w_i = 1 \tag{6}$$

式中，w_i为生成的权重，基于该权重对四阶段精炼特征进行逐元素加权，该过程可表示为：

$$\boldsymbol{x}_i = w_i \cdot \boldsymbol{x}_i \tag{7}$$

将加权后的层级特征沿通道维度拼接，通过3×3卷积进行深度融合，进一步经SE模块执行通道精炼，最后通过ReLU激活得到聚合质量特征，得到聚合质量特征$\boldsymbol{F}_{\mathrm{fuse}}$。

通过上述设计，多层级自适应质量解码器以多层级自适应加权替代传统固定拼接融合，使模型能够根据全局内容动态调节多尺度特征的聚合比例，在提高关键质量信息利用效率的同时，以极小的参数量开销实现了端到端优化，有效提升了质量预测精度与主客观一致性。

2.5 质量分数回归

融合后的特征$\boldsymbol{F}_{\mathrm{fuse}}$依次经三层1×1卷积逐层降维并激活，得到质量感知特征图$\boldsymbol{F}_{\mathrm{m}} \in \mathbf{R}^{8\times 64\times 64}$；将其展平为向量后，通过三层感知器回归图像质量分数，该过程可表示为：

$$q = R\left(\mathrm{Vec}\left(\boldsymbol{F}_{\mathrm{m}}\right)\right) \tag{8}$$

式中，Vec(·)表示由三层感知器完成的空间展平操作，$R(\cdot)$为三层全连接映射。

3 实　验

3.1 数据集

本文实验在以下4个真实失真数据集上展开。

1）CLIVE数据库（Ghadiyaram和Bovik，2015）。共包含1162幅真实场景图像，这些图像由用户通过移动设备在真实环境中自由拍摄采集，涵盖了多种复杂真实失真类型，包括噪声、模糊、颜色失真、过曝/欠曝、压缩伪影等，每幅图像均配有主观质量评分。

2）KonIQ-10k数据库（Hosu等人，2020）。共包含10073幅真实场景图像，涵盖7种场景类别和多种真实失真类型，包括传感器噪声、运动模糊、对焦模糊、压缩伪影、亮度失真等，每幅图像均通过众包方式收集了至少120个主观评分。

3）LIVEFB数据库（Ying等人，2020）。共包含约40000幅真实场景图像，这些图像从社交平台用户上传的真实照片中采样获得，涵盖了极其丰富的场景内容和多样化的真实失真类型，包括压缩伪影、噪声、模糊、曝光异常、颜色失真等，每幅图像均配有主观质量评分。

4）SPAQ数据库（Fang等人，2020）。共包含11125幅智能手机拍摄图像，这些图像由多种不同型号的智能手机在不同拍摄条件下采集，涵盖了丰富的场景类型和真实失真类型，包括噪声、模糊、曝光异常、颜色失真、压缩伪影等，每幅图像均配有主观质量评分。

3.2 实验设置

实验基于PyTorch框架实现，硬件平台为NVIDIA A100 GPU。采用Stable Diffusion v1.5作为预训练扩散主干，冻结其U-Net编码器参数，已有线性探测研究表明（Fu等人，2024），其冻结骨干的先验表征对质量评价任务已足够充分，本文方法的设计重点在于特征的补偿与自适应融合；且v1.5原生支持512×512分辨率，与本文输入设置相匹配，模型规模适中、适配器生态成熟；本文核心模块与骨干网络解耦，可直接迁移至更高版本的扩散模型，骨干升级的探索将列为未来工作。输入图像统一缩放至512×512，不进行分块处理。采用Adam优化器，训练批次大小设定为12，初始学习率为1×10^{-5}，训练轮次设置为15个epoch。学习率调度采用MultiStepLR策略，衰减因子为0.2，衰减节点依据各数据集规模设定。各数据集按8:2的比例随机划分为训练集与测试集，划分过程重复5次，报告5次实验结果的中位数，与该领域主流实验协议保持一致。以上超参均沿用对比基准的官方配置，保证与基线在完全一致的训练条件下公平对比。

3.3 评价指标

本文采用2种图像质量评价研究中常用的评价指标，即PLCC与SRCC。这两个指标的范围从-1到1，绝对值越高表示相关性越强。PLCC是指预测值和真实值的线性相关性，而SRCC衡量预测结果与真实值之间的单调关系。

3.4 性能比较

为了评估本文模型性能的有效性和优势，将提出的模型与目前主流的无参考图像质量评价方法进

© 中国图象图形学报版权所有

行综合性能比较，其中包括2种基于传统深度学习的模型：HyperIQA（hyper network-based blind image quality assessment）（Su等人，2020）和SaTQA（supervised contrastive learning transformer for image quality assessment）（Shi等人，2024），1种基于视觉语言预训练的模型：LoDa（local feature enhancement for image quality assessment）（Xu等人，2024），以及2种基于扩散先验或退化信息建模的模型：DP-IQA（diffusion prior for blind image quality assessment）（Fu等人，2024）和MDM-GFIQA（multi-scale dynamic modulation and degradation information for image quality assessment）（Zhao等人，2026）。不同方法的性能比较见表1。

本文方法与对比基准DP-IQA基于其官方代码在完全相同的实验协议下训练与测试，保证了与对比性能最强方法之间比较的严格公平。表1中其余对比方法的结果引自各方法原论文，这些方法均采用8∶2的实验比例进行数据划分。因此，在相同数据集与评价指标下，表1的对比结果具有可比性。

实验结果表明，本文方法在CLIVE、KonIQ-10k和LIVEFB数据集上均取得最优性能，在SPAQ数据集上取得具有竞争力的结果。相比基于分类先验与端到端学习的传统深度方法，本文方法在各数据集上均有明显优势，这说明扩散模型在海量真实图像上的生成式预训练使其先验同时涵盖高层语义与低层纹理，与质量评价对内容与失真双维度的感知需求天然契合；相比基于视觉语言先验的方法，本文方法在KonIQ-10k上PLCC较LoDa提升1.2%，原因在于CLIP图像编码器对细粒度失真不敏感（Bianchi等人，2024），而去噪先验对模糊、噪声、压缩等细粒度失真具有更强的敏感性；相比同为扩散先验方法的DP-IQA，本文方法在KonIQ-10k、CLIVE和LIVEFB上较次优结果的PLCC分别相对提升0.63%、1.43%和3.67%，SRCC分别相对提升0.53%、1.80%和1.90%，增益主要来自两方面：多级特征提取分支对VAE编码高频信息损失的补偿，以及层级自适应注意力融合机制对多阶段特征的内容自适应聚合。实验结果表明，本文提出的方法是有效的，在CLIVE、KonIQ-10k和LIVEFB数据集上表现稳定。在SPAQ数据集上，本文方法取得与最优方法LoDa相当的结果，且优于其余对比方法，展现出具有竞争力的性能。LoDa通过高效Transformer自适应与局部特征增强充分利用了大规模预训练模型的先验知识，与本文方法利用的扩散先验在规模与来源上相近，因此两者在SPAQ数据集上性能接近。

表1　与其他模型的对比实验结果

Table 1　Comparison experiment results with other models

	KONIQ-10K		CLIVE		SPAQ		LIVEFB	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
HyperIQA	0.917	0.906	0.882	0.859	0.915	0.911	0.602	0.544
SaTQA	0.941	0.930	0.903	0.877	—	—	0.676	0.582
LoDa	0.944	0.932	0.899	0.876	**0.928**	**0.925**	0.679	0.578
DP-IQA	0.949	0.941	0.911	0.891	0.924	0.921	0.681	0.578
MDM-GFIQA	0.942	0.930	0.908	0.887	—	—	0.653	0.571
本文方法	**0.955**	**0.946**	**0.924**	**0.907**	0.927	0.924	**0.706**	**0.589**

注：粗体表示最优结果，“—”表示该方法未在数据集上训练测试。

3.5　可视化对比实验

为直观验证本文方法对图像内容的选择性感知能力，在CLIVE、KonIQ-10k、LIVEFB和SPAQ四个数据集上开展定性对比实验。本文选取DP-IQA作为对比基准（Baseline），该方法与本文采用相同的扩散先验骨干，且在多个数据集上取得了仅次于本文方法的优异性能，便于在特征响应可视化层面进行公平对比。从各数据集中随机选取一张样本，对比结果如图7所示。每幅子图从左至右依次为原始图像及GT、Baseline预测分数及显著性叠加图、本文方法预测分数及显著性叠加图。两方法的显著性图均由质量解码器输出的多通道特征图经通道平均、归一

© 中国图象图形学报版权所有

化及双线性插值上采样得到，暖色区域表示质量感知特征响应较强的区域。

从实验结果可以看出，在所示样例中，本文方法的显著性图较Baseline能够更聚焦于语义显著区域与纹理复杂结构，对应预测分数与主观评分的偏差更小，与表1的定量结果趋势一致且相互印证。

3.6 消融实验

为验证所提模块的协同效应，本文选取KonIQ-10k与CLIVE作为消融实验数据集：前者为大规模真实失真数据集，能够充分验证模块在数据充足场景下的增益；后者为小规模真实失真数据集，可检验模块在有限样本下的稳定性。两个不同规模的数据集具有互补的验证价值，实验结果如表2所示。在固定Stable Diffusion v1.5中U-Net编码器的基础上，对MFEB的内部子模块FA、CSA及质量解码器中的LAAF机制进行组合消融。结果显示，MFEB的引入为模型带来了性能增益，包含的两个子模块呈递进式优化。进一步引入LAAF后，模型性能进一步增强，在KonIQ-10k与CLIVE上分别取得0.955/0.946与0.924/0.907的良好性能。

3.7 泛化性实验

为验证模型的泛化能力，本文在三个真实失真数据集KonIQ-10K，CLIVE与LIVEFB间进行了跨数据集实验来评估模型在不同规模数据集间的泛化能力。实验结果如表3所示，对比指标选用的是SRCC。LoDa在LIVEFB至CLIVE的特定迁移路径上表现突出，这源于其对局部纹理的强归纳偏置与CLIVE数据集的失真特性更为契合；MDM-GFIQA基于显式退化信息建模与多尺度动态调制，在有限样本下训练得到的退化特征表示与KonIQ-10K的失真统计分布具有更高的兼容性。但从整体来看，本文方法在平均跨域性能上表现良好，其原因主要在于扩散骨干在海量多样化图文数据上预训练所得的先验覆盖多种场景与失真模式，不依赖特定数据集的统计分布，因而迁移至未见数据集时性能衰减较小；同时，训练中冻结U-Net参数、仅优化轻量化的适配分支与质量解码器，避免了模型在小规模数据上的过拟合，最大限度保留了预训练先验的泛化表征；此外，层级自适应注意力融合机制依据输入图像的全局内容动态分配各层级权重，使模型关注与内容相关的质量线索，而非记忆特定数据集的失真模式。上述设计让本文方法具有良好的跨域泛化能力。

3.8 复杂度对比

为评估模型效率，本节从参数量、计算量与单张图像推理时间三个维度对比本文方法与基线DP-IQA，所有指标在同一NVIDIA A100 GPU上测得，结果如表4所示。通过输入固定尺寸为512×512的随机张量，批量大小为1，去除前10次预热后取100次前向推理的平均值来计算单张图像推理时间。选择DP-IQA作为对比对象，是因为本文方法以其为直接基线，两者共享完全相同的扩散先验骨干，且差异可明确归因于本文所提模块。

由表4可见，本文方法在性能提升的同时，总参数量与可训练参数量均较基线减少16.1M，可训练参数压缩约12.9%，这主要源于质量解码器的轻量化设计，也有助于缓解小规模数据集上的过拟合风险。两方法的浮点运算数（floating point operations，FLOPs）基本持平，单张推理时间相差较小，原因在于前向计算量由共享的冻结U-Net骨干主导，而本文新增模块均为轻量设计，其引入的开销可忽略。

4 结 论

本文针对真实场景无参考图像质量评价中，失真多尺度耦合、空间分布非均匀且失真类型多样等复杂特性带来的质量评价难题，以及现有方法难以在有限标注数据下同时实现全局语义理解与局部细节保持的局限，提出一种基于扩散先验的质量评价方法。该方法充分利用预训练扩散模型的视觉先验知识，通过图像编码与多尺度补偿、质量条件嵌入及层级自适应质量解码三个模块的协同设计，实现了从全局语义到局部细节的多层级质量特征提取与自适应融合。在多个公开真实失真数据集上的实验结果表明，本方法的质量预测精度与主观感知一致性均优于现有主流方法。此外，消融实验验证了各核心模块的协同增益效果，跨数据集迁移测试则进一步证实了方法优异的跨域泛化能力。但当前方法仍存在一定局限：一方面，模型参数量与计算开销较大，难以直接部署于资源受限的设备，未来将探索知识蒸馏架构，在保持多尺度细节增强能力的前提下压缩模型计算开销；另一方面，本文实验所用的真实失真数据集中，失真多为多种类型耦合且缺乏显式的失真类型标注，难以按失真类型进行严格的分组

© 中国图象图形学报版权所有



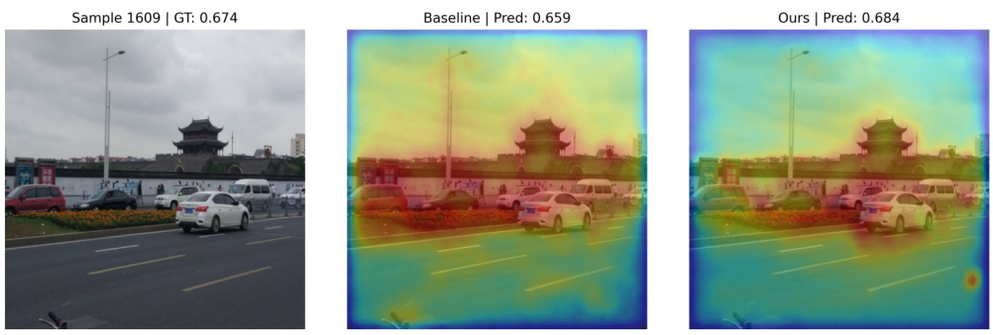
（a） CLIVE



（b） KonIQ-10k



（c） LIVEFB



（d） SPAQ

((a)CLIVE;(b)KonIQ-10k;(c)LIVEFB;(d)SPAQ)

图 7 四个数据集上的可视化对比

Fig. 7 Visual comparison on four datasets

© 中国图象图形学报版权所有

表2　消融实验

Table 2　Ablation experiment

FA	CSA	LAAF	KONIQ-10K		CLIVE	
			PLCC	SRCC	PLCC	SRCC
			0.949	0.941	0.911	0.891
√			0.951	0.941	0.914	0.895
√	√		0.953	0.945	0.921	0.902
		√	0.952	0.943	0.913	0.896
√	√	√	**0.955**	**0.946**	**0.924**	**0.907**

注："√"表示启用对应模块，粗体表示最优结果。

表3　泛化性能对比

Table 3　Generalization performance evaluation

Training on	LIVEFB		CLIVE	KonIQ
Testing on	KonIQ	CLIVE	KonIQ	CLIVE
HyperIQA	0.758	0.735	0.772	0.785
LoDa	0.763	**0.805**	0.745	0.811
DP-IQA	0.771	0.770	0.766	0.833
MDM-GFIQA	–	–	**0.799**	0.816
本文方法	**0.781**	0.793	0.796	**0.856**

注：粗体表示最优结果，"—"表示未开展该组实验。

表4　模型复杂度与推理效率对比

Table 4　Comparison of model complexity and inference efficiency

方法	Params/M	Trainable Params/M	FLOPs/G	推理时间/ms
DP-IQA	1191.3	125.1	1796.9	**96.2**
本文方法	**1175.2**	**109.0**	1796.8	97.0

注：粗体表示最优结果。

统计与细粒度分析，未来将结合失真类型标注，进一步细化模型对各类失真感知能力的评估。

参考文献(References)

Bianchi L, Carrara F, Messina N and Falchi F. 2024. Is CLIP the main roadblock for fine-grained open-world perception? [C]//2024 International Conference on Content-Based Multimedia Indexing (CBMI). Reykjavik, Iceland: IEEE: 1-8 [DOI: 10.1109/CBMI60961.2024.10678732]

Chen C, Mo J, Hou J, Wu H, Liao L, Sun W, et al. 2024. Topiq: a top-down approach from semantics to distortions for image quality assessment[J]. IEEE Transactions on Image Processing, 33: 2404-2418 [DOI: 10.1109/TIP.2024.3376512]

Chen J, Qin F and Lu F. 2025. CSPP-IQA: a multi-scale spatial pyramid pooling-based approach for blind image quality assessment[J]. Neural Computing and Applications, 37(11): 7499-7510

Chen X, Liu Z, Xie S, et al. Deconstructing denoising diffusion models for self-supervised learning [C]// International Conference on Learning Representations. 2025, 2025: 55458-55472.

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. 2021. An image is worth 16x16 words: transformers for image recognition at scale [C]// 9th International Conference on Learning Representations (ICLR). Virtual, Austria: OpenReview [EB/OL]. [2021-05-03]. https:// openreview. net/forum? id= YicbFdNTTy

Fang Y, Zhu H, Zeng Y, Ma K and Wang Z. 2020. Perceptual quality assessment of smartphone photography [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, USA: IEEE: 3677-3686 [DOI: 10.1109/CVPR42600.2020.00374]

Fang Y M, Sui X J, Yan J B, Liu X L and Huang L P. 2021. Progress in no-reference image quality assessment. Journal of Image and Graphics, 26(2): 265-286 (方玉明, 眭相杰, 鄢杰斌, 刘学林, 黄丽萍. 无参考图像质量评价研究进展. 中国图象图形学报, 2021, 26(2): 265-286) [DOI: 10.11834/jig.200274]

Fu H, Wang Y, Yang W, Kot A C and Wen B. 2024. DP-IQA: utilizing diffusion prior for blind image quality assessment in the wild [EB/OL].[2024-05-31]. https://arxiv.org/abs/2405.19996

Gao P, Geng S, Zhang R, Ma T, Fang R, Zhang Y, et al. 2024. CLIP-adapter: better vision-language models with feature adapters [J]. International Journal of Computer Vision, 132 (2): 581-595 [DOI: 10.1007/s11263-023-01891-x]

Ghadiyaram D and Bovik A C. 2015. Massive online crowdsourced study of subjective and objective picture quality [J]. IEEE Transactions on Image Processing, 25(1): 372-387 [DOI: 10.1109/TIP.2015.2499326]

Hosu V, Lin H, Szirányi T and Saupe D. 2020. KonIQ-10k: an ecologically valid database for deep learning of blind image quality assessment[J]. IEEE Transactions on Image Processing, 29: 4041-4056 [DOI: 10.1109/TIP.2019.2955243]

Hu J, Shen L and Sun G. 2018. Squeeze-and-excitation networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7132-7141 [DOI: 10.1109/CVPR.2018.00745]

Ke J, Wang Q, Wang Y, Milanfar P and Yang F. 2021. MUSIQ: multi-scale image quality transformer[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Virtual, Canada: IEEE: 5148-5157 [DOI: 10.1109/ICCV48922.2021.00510]

Li X, Zhang Y, Shen Y, Li K, Hu R, Zheng X, et al. 2025. Feature

© 中国图象图形学报版权所有

denoising diffusion model for blind image quality assessment [C]// Proceedings of the AAAI Conference on Artificial Intelligence. Philadelphia, USA, 39(5): 5004-5012 [DOI: 10.1609/aaai.v39i5.32530]

Mittal A, Soundararajan R and Bovik A C. 2012. Making a "completely blind" image quality analyzer[J]. IEEE Signal Processing Letters, 20(3): 209-212 [DOI: 10.1109/LSP.2012.2227726]

Moorthy A K and Bovik A C. 2011. Blind image quality assessment: from natural scene statistics to perceptual quality[J]. IEEE Transactions on Image Processing, 20(12): 3350-3364 [DOI: 10.1109/TIP.2011.2140650]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. 2021. Learning transferable visual models from natural language supervision[C]//Proceedings of the 38th International Conference on Machine Learning. Virtual, USA: PMLR, 139: 8748-8763

Rombach R, Blattmann A, Lorenz D, Esser P and Ommer B. 2022. High-resolution image synthesis with latent diffusion models [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 10684-10695 [DOI: 10.1109/CVPR52688.2022.01042]

Cheng R Q, Yu Y, Shi D Z and Cai W. 2022. The critical review of image and video quality assessment methods. Journal of Image and Graphics, 27(5): 1410-1429 (程茹秋, 余烨, 石岱宗, 蔡文. 图像与视频质量评价综述. 中国图象图形学报, 2022, 27(5): 1410-1429) [DOI: 10.11834/jig.210314]

Shi J, Gao P and Qin J. 2024. Transformer-based no-reference image quality assessment via supervised contrastive learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada, 38(5): 4829-4837

Shi W, Caballero J, Huszár F, Totz J, Aitken A P, Bishop R, et al. 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 1874-1883 [DOI: 10.1109/CVPR.2016.207]

Su S, Yan Q, Zhu Y, Zhang C, Ge X, Sun J, et al. 2020. Blindly assess image quality in the wild guided by a self-adaptive hyper network [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, USA: IEEE: 3667-3676 [DOI: 10.1109/CVPR42600.2020.00373]

Wang J, Chan K C K and Loy C C. 2023. Exploring CLIP for assessing the look and feel of images[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, USA, 37(2): 2555-2563 [DOI: 10.1609/aaai.v37i2.25354]

Wang Z, Hu B, Zhang M, et al. Diffusion model-based visual compensation guidance and visual difference analysis for no-reference image quality assessment[J]. IEEE Transactions on Image Processing, 2025, 34: 263-278.

Woo S, Park J, Lee J Y and So I S. 2018. CBAM: convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision. Munich, Germany: Springer: 3-19 [DOI: 10.1007/978-3-030-01234-2_1]

Xu K, Liao L, Xiao J, Chen C, Wu H, Yan Q, et al. 2024. Boosting image quality assessment through efficient transformer adaptation with local feature enhancement[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 2662-2672 [DOI: 10.1109/ CVPR52733.2024.00255]

Yan J B, Fang Y M and Liu X L. 2022. The review of distortion-related image quality assessment. Journal of Image and Graphics, 27(5): 1430-1466 (鄢杰斌, 方玉明, 刘学林. 图像质量评价研究综述——从失真的角度. 中国图象图形学报, 2022, 27(5): 1430-1466) [DOI: 10.11834/jig.210527]

Yang S, Wu T, Shi S, Lao S, Gong Y, Cao M, et al. 2022. MANIQA: multi-dimension attention network for no-reference image quality assessment[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 1191-1200 [DOI: 10.1109/CVPR52688. 2022.00124]

Ying Z, Niu H, Gupta P, Mahajan D, Ghadiyaram D and Bovik A C. 2020. From patches to pictures (PaQ-2-PiQ): mapping the perceptual space of picture quality [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, USA: IEEE: 3575-3585 [DOI: 10.1109/ CVPR42600.2020.00363]

Zhai G and Min X. 2020. Perceptual image quality assessment: a survey [J]. Science China Information Sciences, 63(11): 211301 [DOI: 10.1007/s11432-020-2898-8]

Zhang R, Tian K, Zhang Z, Li X, Wang X and Zheng X. 2025. FDG-Diff: frequency-domain-guided diffusion framework for compressed hazy image restoration[EB/OL].[2025-01-22]. https://arxiv.org/abs/2501.12832

Zhang W, Ma K, Yan J, Deng D and Wang Z. 2018. Blind image quality assessment using a deep bilinear convolutional neural network [J]. IEEE Transactions on Circuits and Systems for Video Technology, 30(1): 36-47 [DOI: 10.1109/TCSVT.2018. 2889897]

Zhang W, Zhai G, Wei Y, Yang X and Ma K. 2023. Blind image quality assessment via vision-language correspondence: a multitask learning perspective [C]//Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 14071-14081 [DOI: 10.1109/CVPR 52729.2023.01350]

Zhang Z, Zhou Y, Li C, Dong X, Liu X and Chen C. 2025. Quality assessment in the era of large models: a survey[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 21(7): 1-31 [DOI: 10.1145/3726303]

Zhao Y, Zhang Y and Xia T. 2026. No-reference image quality assessment based on multi-scale dynamic modulation and degradation

© 中国图象图形学报版权所有

information[J]. Displays, 91: 103207

Zhu H, Li L, Wu J, Dong R, Chen C and Lu Y. 2020. MetaIQA: deep meta-learning for no-reference image quality assessment [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, USA: IEEE: 14143- 14152 [DOI: 10.1109/CVPR42600.2020.01417]

© 中国图象图形学报版权所有