CN 51-1346/O4 ISSN 1003-501X (印刷版) ISSN 2094-4019 (网络版)

光电工程
Opto-Electronic Engineering

中文核心期刊 中国科技核心期刊
Scopus CSCD

基于Swin-AK Transformer的智能手机拍摄图像质量评价方法

侯国鹏，董武，陆利坤，周子镱，马倩，柏振，郑晟辉

引用本文:

侯国鹏，董武，陆利坤，等. 基于Swin-AK Transformer的智能手机拍摄图像质量评价方法[J]. 光电工程，2025，**52**(1): 240264.

Hou G P, Dong W, Lu L K, et al. Smartphone image quality assessment method based on Swin-AK Transformer[J]. *Opto-Electron Eng*, 2025, **52**(1): 240264.

https://doi.org/10.12086/oee.2025.240264

收稿日期: 2024-11-11; 修改日期: 2024-12-23; 录用日期: 2024-12-23

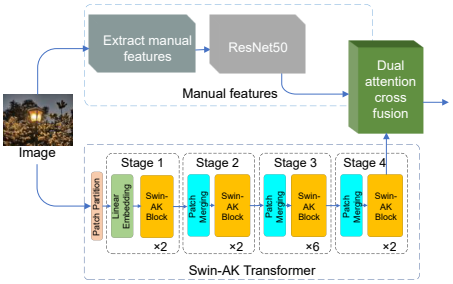
Opto-Electronic Engineering
光 电 工 程

Article

2025 年, 第 52 卷, 第 1 期

DOI: 10.12086/oee.2025.240264 CSTR: 32245.14.oee.2025.240264

基于 Swin-AK Transformer 的智能手机拍摄图像质量评价方法



侯国鹏，董　武*，陆利坤，周子镱，
马　倩，柏　振，郑晟辉

北京印刷学院高端印刷装备信号与信息处理北京市重点实验室，
北京 102600

摘要：本文提出了一种基于双交叉注意力融合的 Swin-AK Transformer (Swin Transformer based on alterable kernel convolution) 和手工特征相结合的智能手机拍摄图像质量评价方法。首先，提取了影响图像质量的手工特征，这些特征可以捕捉到图像中细微的视觉变化；其次，提出了 Swin-AK Transformer，增强了模型对局部信息的提取和处理能力。此外，本文设计了双交叉注意力融合模块，结合空间注意力和通道注意力机制，融合了手工特征与深度特征，实现了更加精确的图像质量预测。实验结果表明，在 SPAQ 和 LIVE-C 数据集上，皮尔森线性相关系数分别达到 0.932 和 0.885，斯皮尔曼等级排序相关系数分别达到 0.929 和 0.858。上述结果证明了本文提出的方法能够有效地预测智能手机拍摄图像的质量。

关键词：图像质量评价；智能手机拍摄图像；Swin Transformer；手工特征；空间注意力；通道注意力

中图分类号：TP391.4　　**文献标志码：**A

侯国鹏，董武，陆利坤，等. 基于 Swin-AK Transformer 的智能手机拍摄图像质量评价方法 [J]. 光电工程，2025，**52**(1): 240264

Hou G P, Dong W, Lu L K, et al. Smartphone image quality assessment method based on Swin-AK Transformer[J]. *Opto-Electron Eng*, 2025, **52**(1): 240264

Smartphone image quality assessment method based on Swin-AK Transformer

Hou Guopeng, Dong Wu*, Lu Likun, Zhou Ziyi, Ma Qian, Bai Zhen, Zheng Shenghui

Beijing Key Laboratory of Signal and Information Processing for High-end Printing Equipment, Beijing Institute of Graphic Communication, Beijing 102600, China

Abstract: This paper proposes a smartphone image quality assessment method that combines the Swin-AK Transformer based on alterable kernel convolution and manual features based on dual attention cross-fusion. Firstly, manual features that affected image quality were extracted. These features could capture subtle visual changes in images. Secondly, the Swin-AK Transformer was presented and it could improve the extraction and

收稿日期：2024-11-11；修回日期：2024-12-23；录用日期：2024-12-23

基金项目：北京市数字教育研究重点课题 (BDEC2022619027)；北京市高等教育学会 2023 年立项面上课题 (MS2023168)；北京印刷学院校级科研项目 (Ec202303, Ea202301, E6202405)；北京印刷学院学科建设和研究生教育专项 (21090323009, 21090224002, 21090124013)；北京市教育委员会出版学新兴交叉学科平台建设-数字喷墨印刷技术及多功能轮转胶印机关键技术研发平台项目 (04190123001/003)；北京邮电大学网络与交换技术全国重点实验室开放课题资助项目 (SKLNST-2023-1-12)；北京印刷学院“人工智能+”课程建设项目

*通信作者：董武，dongwu@bigc.edu.cn。

processing of local information. In addition, a dual attention cross-fusion module was designed, integrating spatial attention and channel attention mechanisms to fuse manual features with deep features. Experimental results show that the Pearson correlation coefficients on the SPAQ and LIVE-C datasets reached 0.932 and 0.885, respectively, while the Spearman rank-order correlation coefficients reached 0.929 and 0.858, respectively. These results demonstrate that the proposed method in this paper can effectively predict the quality of smartphone images.

Keywords: image quality assessment; smartphone image; Swin Transformer; manual features; spatial attention; channel attention

1 引　言

随着智能手机摄像技术的快速发展，人们越来越习惯使用智能手机进行拍摄，产生了大量的图像。这些图像广泛应用在社交媒体、电子商务等领域中，准确评价智能手机拍摄图像的质量变得尤为重要。然而，由于智能手机拍摄条件的多样性和随机性，准确地评价图像的质量成为一个难点。

图像质量评价 (Image quality assessment, IQA) 分为两大类：主观评价和客观评价。主观评价主要依赖于人们观察图像后给出的评分，虽然接近用户的真实感受，但此方法耗时且成本高昂，不易于在实际应用中广泛采用[1]；相反，客观评价使用算法模拟人眼视觉系统 (human visual system, HVS) 的特点，并对图像的质量进行评价，这种方法效率高，极大地提高了图像质量评价的实用性和广泛应用。

早期的图像质量评价方法大多为通用型方法，虽然在一般场景中表现较好，但在面对智能手机拍摄的复杂图像失真时存在局限性，尤其是在处理复杂非线性失真时效果较差。Ke 等[2] 提出了一种多尺度的图像质量评价方法，虽然能够处理不同分辨率和长宽比的失真图像，避免了传统卷积神经网络输入固定图像尺寸导致的效果下降问题。Varga 等[3] 采用了一种基于全局和局部图像统计特征的无参考通用型图像质量评价方法，该方法模拟了人类视觉系统感知图像质量的特点，并将图像的全局统计特征和局部统计特征组合在一起。Jain 等[4] 提出了一种无参考通用型图像质量评价方法，此方法结合了自然场景的统计特征和卷积神经网络来预测图像质量的分数。Shao 等[5] 针对图像内容失真的特点，提出了一种新的无参考通用型图像质量评价方法，该方法使用内容感知和失真推理网络有效预测了合成和真实失真图像的质量分数。Zhao 等[6] 提出了一种面向无参考图像质量评价的质量感知预训练模型，旨在克服标记数据稀缺的问题来提升性能。对于智能手机拍摄图像，通用型图像质量评价方法的效果普遍较差，原因在于智能手机图像的拍摄环境多样化，以及不同品牌和型号摄像头的技术差异。这使得用户难以高效挑选高质量的图像，因此开发专门针对智能手机图像特点和失真类型的质量评价方法显得尤为重要。

近几年，越来越多的研究专注于智能手机拍摄图像质量评价，Fang 等[7] 首次开展了智能手机摄影感知质量评价的研究，并创建了一个包含 11125 张由 66 款智能手机拍摄的图像数据集 (smartphone photography attribute and quality, SPAQ)。Yuan 等[8] 提出了一种专门针对智能手机摄影图像的质量评价方法，此方法结合了美学和人眼视觉感知的特性，从曝光、噪声、颜色和纹理这四个方面对图像的质量进行评价。该方法只提取了手工特征，没有使用深度学习，从而限制了此方法的适应性和泛化能力。Zhou 等[9] 提出了一种多指标智能手机图像质量评价模型，根据人眼视觉感知的特点，使用两种图像裁剪方法获取输入图像，并使用回归分析来预测颜色、纹理、噪声和曝光的分数。Huang 等[10] 提出了一种多任务深度卷积神经网络模型，用于智能手机拍摄图像的无参考质量评价。在此模型中，把场景类型的检测作为辅助任务，但这一任务的准确性并不稳定，场景类型的分类错误可能会影响整体的质量评价。Yao 等[11] 提出了一种基于残差块的卷积神经网络 (convolutional neural network, CNN) 来进行智能手机拍摄图像的无参考质量评价，该模型生成显著性区域并选取图像的特定子区域作为输入信息，这种处理方式会忽略全局的图像信息，从而导致质量评价结果不具有全局性。以上方法研究了智能手机拍摄图像相关属性的评价，并没有进行图像整体内容的质量评价，而本文研究了智能手机拍摄图像整体内容的质量评价。

综上所述，传统手工特征方法[12] 在应对智能手机图像中的复杂失真时表现较差，难以捕捉图像的细

微特征；现有的基于深度学习的图像质量评价方法也未能完全适应智能手机图像的多样性。因此，开发能够全面反映智能手机图像整体内容和细节特征的评价方法成为了一个亟待解决的问题。本文针对智能手机拍摄的图像常常受到噪声、压缩伪影以及复杂光照条件的影响，而导致图像出现质量问题，提出了基于 Swin-AK Transformer (Swin Transformer based on alterable kernel convolution) 和手工特征进行双交叉融合的智能手机拍摄图像客观质量评价方法，用于评价智能手机拍摄图像的质量。此模型首先提取影响智能手机拍摄图像质量的 4 种手工特征。然后，使用 Swin-AK Transformer 提取图像的深度特征，其中 AKconv 的动态卷积核调整机制学习图像的局部特性 (如噪声分布、光照不均匀和边缘强度等)，动态调整卷积核的权重，更好地适应噪声和弱光环境下的细节表现。此外，AKconv 的多尺度调整能力有效捕获了智能手机图像中模糊和细节丢失的问题，确保从全局结构到局部细节的全面表征能力。接着，利用双交叉注意力融合模块 (dual attention cross fusion, DACF) 将上述两种特征进行融合，DACF 模块使用通道注意力和空间注意力机制，针对智能手机拍摄图像的特点，更有效地融合手工特征和深度特征，弥补单一特征表征能力的不足。通道注意力机制能够分析智能手机图像中纹理和光照动态，可以平衡手工特征的全局统计信息与深度特征的局部细节。空间注意力机制精准定位智能手机图像中的关键区域，如细节纹理和边缘区域，同时抑制伪影和噪声的干扰。通道注意力与空间注意力协同作用，能够充分捕捉智能手机图像的全局与局部特性。最后，得到智能手机拍摄图像的质量分数。

2 本文方法

本文提出的基于 Swin-AK Transformer 的智能手机拍摄图像质量评价方法的网络架构主要分为三部分，如图 1 所示。在图 1 中，第一部分先提取手工特征，然后再将这些特征送入到 ResNet50 去提取更加复杂的特征信息；第二部分使用 Swin-AK Transformer 提取深度特征；第三个部分使用双交叉注意力融合模块来融合前两个部分的特征。

手工特征提取网络部分包括多种空间特征和颜色特征的提取：亮度特征、彩色特征、锐度特征和运动模糊特征。这些特征不仅提供了关于图像内容的信息，而且还涵盖了与图像质量相关的信息。将上述提取的手工特征送入 ResNet50。

2.1 手工特征提取

在本文中，手工特征提取部分符合人眼视觉感知的特点，它是提高图像质量预测精确度的重要组成部分。手工特征的提取如图 2 所示。本文使用多种图像处理技术从原始图像中提取关键信息，这些信息包括彩色、亮度、锐度以及与智能手机拍摄图像质量紧密相关的运动模糊。此外，这些特征的选择和处理方式都考虑了人眼视觉感知的特性，以确保评价结果符合人眼对图像质量的真实感受。这些图像特征输入到 ResNet50 中，进行更深入的分析和处理。

2.1.1 彩色特征的提取

当 RGB 空间转换到 HSV 空间时，需要将红、绿、蓝三种彩色通道的数值转换为色调 (hue, H)、饱和度 (saturation, S)、亮度[13] (value, V) 这三个数值[14]。HSV 模型更贴近人眼对彩色的感知特点，使得彩色的处理和分析更加直观和有效。使用 R、G、B 分别表示红、

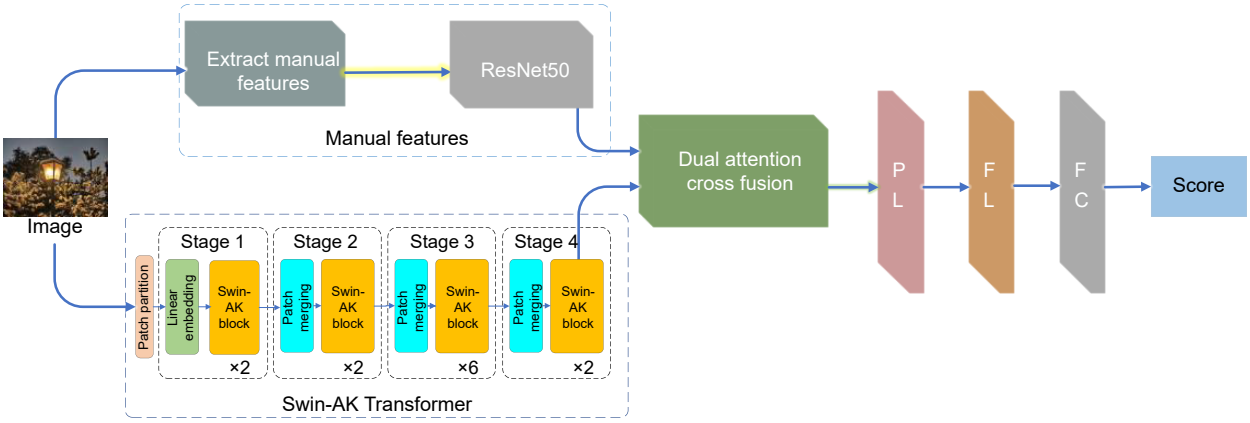


图 1 本文方法的整体结构图

Fig. 1 Overall structure diagram of the proposed method

绿、蓝通道的数值 (通常归一化到 0 到 1 之间)。在进行转换时，首先计算 R、G、B 这三个通道的最大值 (Max) 和最小值 (Min)，然后计算色调，它的计算相对复杂，需要根据最大值来确定。色调的计算分为三种情况，如式 (1) 所示：

$$H=\begin{cases}60^\circ\times\left(\dfrac{G-B}{Max-Min}\bmod 6\right), & Max=R\\ 60^\circ\times\left(\dfrac{B-R}{Max-Min}+2\right), & Max=G\\ 60^\circ\times\left(\dfrac{R-G}{Max-Min}+4\right), & Max=B\end{cases}. \tag{1}$$

如果 $Max=Min$，色调定义为 0。饱和度的计算与最大值有关联，如果最大值为 0 (图像完全为黑色)，饱和度定义为 0；否则，饱和度的计算如式 (2) 所示：

$$S=\frac{Max-Min}{Max}. \tag{2}$$

亮度等于 R、G、B 的最大值。

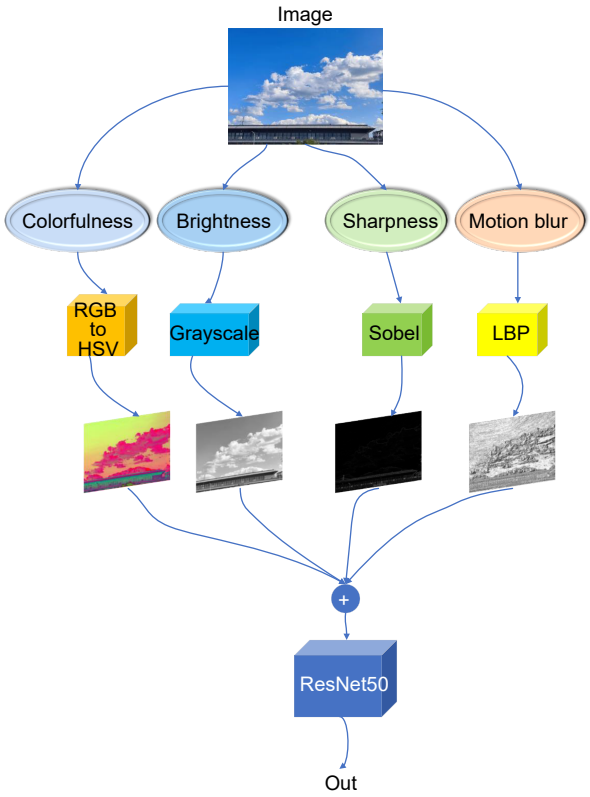


图 2　手工特征提取的示意图

Fig. 2　Diagram of manual feature extraction

2.1.2 亮度特征的提取

为了研究亮度对图像质量的影响，需要进行灰度值的转换。灰度值是所有彩色成分的加权平均，每个像素点的亮度信息用 0 到 255 之间的数值表示。这种方法考虑了人眼对不同彩色敏感度的差异，并对不同的彩色使用不同的权重，使用的公式[15] 如式 (3) 所示：

$$G_{\mathrm{r}}=0.299\times R+0.587\times G+0.114\times B, \tag{3}$$

其中：G_{r}表示灰度，R、G、B 分别代表红色、绿色和蓝色通道的数值。式 (3) 可以确保图像在转换为灰度图时，能够保持人眼视觉的亮度感知特点，同时反映了人眼视觉对不同彩色的敏感度。由于人眼对绿色的敏感度最高，对蓝色的敏感度最低，所以式 (3) 中绿色通道的权重最大，蓝色通道的权重最小。式 (3) 广泛应用于图像处理和计算机视觉领域，使用这个公式获得的灰度图像，在视觉上更接近原始彩色图像的亮度特点。

2.1.3 锐度特征的提取

图像锐度受多个因素的影响，例如对焦的准确性和摄影的质量等。边缘是人眼视觉系统处理图像的关键元素，因为它们帮助大脑解释物体的形状、大小和空间位置。锐度较高的图像具有更加明显的边缘，这种边缘能够帮助人眼视觉系统更快地识别和处理图像的内容。

Sobel 算子使用图像在水平方向和垂直方向上的梯度幅度值来计算边缘的强度值。它利用两个尺寸为 3×3的卷积核分别对图像进行水平方向和垂直方向的卷积操作，计算出每个像素的梯度幅度值[16]，从而得到每个像素的边缘强度值，如式 (4) 所示：

$$G=\sqrt{G_x^2+G_y^2}, \tag{4}$$

其中：G 表示边缘的强度值，G_x和G_y分别代表水平方向和垂直方向上的梯度幅度值。使用这种方式得到的边缘强度图揭示了图像中所有边缘成分的位置及其强度。

2.1.4 运动模糊特征的提取

局部二值模式 (Local binary patterns, LBP) 专注于图像的局部结构，而运动模糊对图像的局部结构有着显著的影响。使用 LBP 分析图像局部结构的改变，能够帮助研究人员了解模糊对图像局部特征产生的影响。LBP 操作首先定义一个以某个像素为中心的圆形邻域，圆周上均匀分布有 P 个邻域像素，其半径为 R_0。然后，将每个邻域像素的灰度值与中心像素的灰度值进行比较。如果邻域像素的灰度值大于或等于中心像素的灰度值，则该邻域像素的比较结果为 1；否则为 0。每个中心像素都会生成一个 P 位的二进制数。最后，把这个二进制数转换成十进制数，作为该中心像素的 LBP 值[17]。LBP 特征的计算如式 (5) 所示：

$$LBP_{P,R_o}=\sum_{p=0}^{P-1}s(g_p-g_c)2^p\,, \tag{5}$$

其中：LBP_{P,R_o}表示得到的 LBP 值；g_c表示中心像素的灰度值；g_p表示第 p 个邻域像素的灰度值；P 表示邻域内的像素数量；R_o 表示圆形邻域的半径；$s(x)$表示一个阶跃函数，如式 (6) 所示：

$$s(g_p-g_c)=\begin{cases}1\,, & g_p\geqslant g_c\\ 0\,, & g_p<g_c\end{cases}\,. \tag{6}$$

2.1.5 ResNet50

本文在使用 ResNet50 提取特征时，只使用其卷积层和池化层，没有使用最后的全连接层和输出层，ResNet50 的网络结构如图 3 所示。本文在对智能手机拍摄的图像进行质量评价时，手工提取的特征与智能手机拍摄图像的质量之间存在非线性映射关系。ResNet50 中的深层神经网络可以学习这种非线性映射关系，而简单地把手工特征进行组合无法做到这一点。ResNet50 的加入可以作为一座桥梁，将手工提取的低级特征转化为抽象程度更高的特征。ResNet50 得到的深度学习特征可以进一步提取手工特征中的信息，从而提供更加丰富的图像质量信息，这些信息包含了手工特征没有包含的图像质量因素。从特征结构的角度来看，手工特征从原始图像直接提取，而深度学习模型可以提供从不同网络层提取的特征。经过 ResNet50 处理后的特征可以更好地与 Swin-AK Transformer 中提取的特征进行结合，从而增强最终质量特征的描述能力。

2.2 Swin-AK Transformer

Swin Transformer[18] 使用了滑动窗口自注意力机制，此机制利用不重叠的局部窗口自注意力操作有效地降低了全局自注意力机制的计算复杂度，从而能够在资源受限的情况下处理更大的图像，并保持较高的计算效率。它侧重于全局信息的处理，而忽略了局部特征产生的影响。然而，由于滑动窗口具有固定的尺寸，Swin Transformer 在处理局部特征和复杂的纹理时存在不足，特别是针对图像中的细致结构[19]，其固定窗口无法适应不同尺度的特征。为了解决这个问题，本文提出了 Swin-AK Transformer，在 Swin Transformer 中使用可动态改变卷积核尺寸的卷积操作。AKConv 根据图像的局部内容自适应地调整卷积核的尺寸，使模型能够更灵活地处理不同尺度的图像特征。这一改进显著增强了网络对复杂局部细节的捕捉能力，尤其是在处理具有精细的纹理和高度复杂的图像中表现出更优异的性能。Swin-AK Transformer 不仅保留了 Swin Transformer 在全局上下文建模中的优势，还显著提升了对细节特征的理解能力，在图像质量评价任务中表现出更高的精确性和鲁棒性。

在 Swin Transformer 中，首先将图像划分为互不重叠的多个窗口，如图 4 所示；然后，在每个窗口内独立计算自注意力，从而显著降低了计算复杂度。

在图 4 中，黄色框表示用于计算自注意力的滑动窗口，灰色框表示一个图像块。第 N 层将图像划分为不重叠的局部窗口 (黄色边框)，在每个窗口内计算自注意力以提取局部特征，把这种操作得到的自注意力

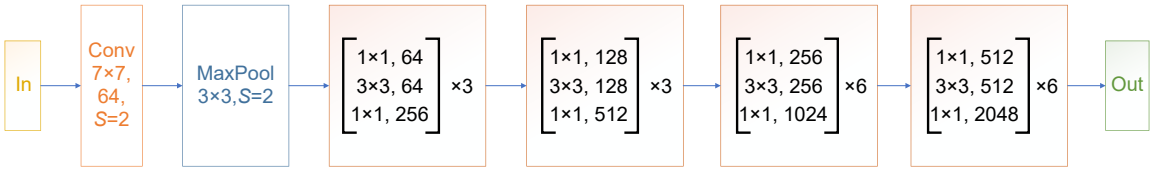


图 3 ResNet50 网络结构图

Fig. 3 ResNet50 architecture diagram

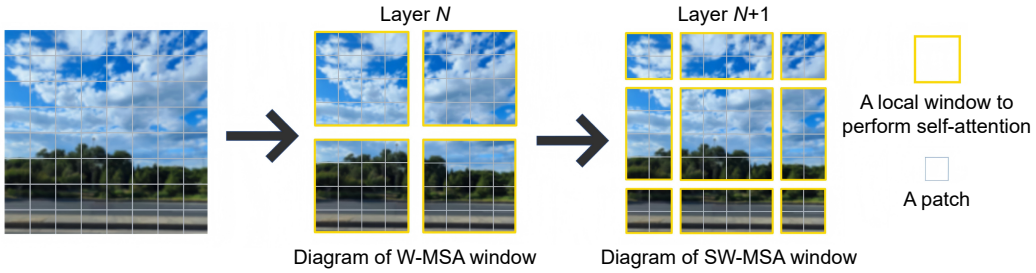


图 4 Swin Transformer 的滑动窗口操作示意图

Fig. 4 Diagram of the sliding window operation in Swin Transformer

称为窗口多头自注意力(window multihead self-attention, W-MSA)。第 N+1 层进行滑动窗口操作(即窗口边界的移动)实现跨窗口连接，逐步整合全局信息，并在重新分配的窗口内计算自注意力，把这种操作得到的自注意力称为滑动窗口多头自注意力(shifted window multi head self attention, SW-MSA)。交替进行注意力的计算和滑动窗口操作，模型在保持计算复杂度的同时，能够逐步融合局部与全局信息，从而适应高分辨率图像的处理需求。每个窗口进行展平操作，用于后续的多头自注意力机制，如式 (7) 所示。

$$X_i \in R^{M^2\times C} \to X_i^{\mathrm{p}} \in R^{M^2\times C}, \tag{7}$$

其中：X_i为尚未展平的第 i 个窗口；维度为$R^{M^2\times C}$；窗口的空间尺寸为M^2；C 为通道的数量；X_i^{p}为展平后的第 i 个窗口。对展平后的窗口X_i^{p}，使用线性变换得到查询向量、键向量和值向量，如式 (8) 所示。

$$\begin{cases} \boldsymbol{Q} = X_i^{\mathrm{p}}\boldsymbol{W}_{\mathrm{Q}} \\ \boldsymbol{K} = X_i^{\mathrm{p}}\boldsymbol{W}_{\mathrm{K}} \\ \boldsymbol{V} = X_i^{\mathrm{p}}\boldsymbol{W}_{\mathrm{V}} \end{cases}, \tag{8}$$

其中：$\boldsymbol{Q}$ 表示查询向量；$\boldsymbol{K}$ 表示键向量；$\boldsymbol{V}$ 表示值向量；$\boldsymbol{W}_{\mathrm{Q}}$、$\boldsymbol{W}_{\mathrm{K}}$、$\boldsymbol{W}_{\mathrm{V}}$表示权重矩阵。

接着，计算查询向量和键向量之间的点积，并使用 Softmax 函数进行归一化操作。最后，将多个头的结果进行拼接，并使用线性变换得到最终的输出结果。

本文提出的 Swin-AK Transformer 的主要结构如图 5 所示，它包括图像分块 (patch partition)、线性嵌入 (linear embedding) 以及多个包含滑动窗口自注意力机制的 Swin-AK 块 (Swin-AK block)。在每个阶段 (Stage 1 至 Stage 4) 之间，使用图像块合并 (patch merging) 操作减少特征图的分辨率，同时增加通道的数量。

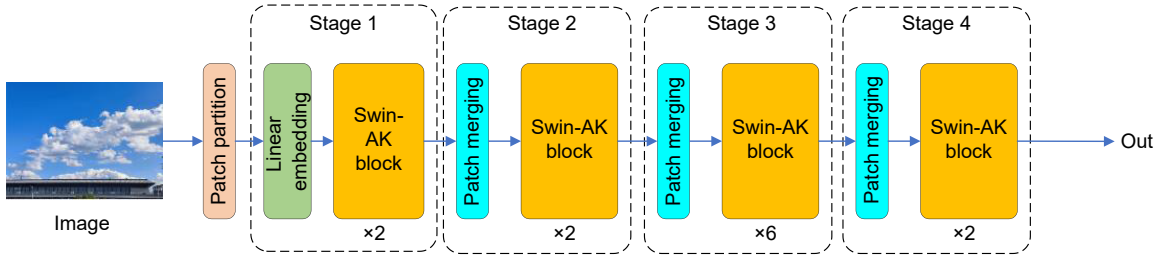


图 5 Swin-AK Transformer 结构图

Fig. 5 Swin-AK Transformer architecture diagram

Swin-AK blocks 的结构如图 6 所示。首先，输入的特征经过层归一化 (layer normalizaiton, LN)。然后，使用 AKConv 和 W-MSA 并行处理。AKConv 进行动态卷积操作，W-MSA 则使用划分窗口计算局部注意力机制，两者输出的结果融合在一起，融合后的特征使用 LN 和多层感知器 (multilayer perceptron, MLP) 进行处理。最后，使用 SW-MSA 代替 W-MSA 去重复前面的操作。和 W-MSA 相比，SW-MSA 使用了滑动窗口机制，所以它能够更好地提高模型的全局感知能力和特征提取能力。该结构使用 AK-Conv 和 W-MSA 的并行工作，增强了特征提取的多样性和鲁棒性，有效提高了图像特征的表示能力。

AKConv 是 Swin-AK blocks 中的关键组成部分，它显著增强了特征的理解能力和细节捕捉能力，它的结构如图 7 所示。AKConv 使用局部卷积操作，并结合偏移卷积来调整采样点位置，从而更有效地捕捉图

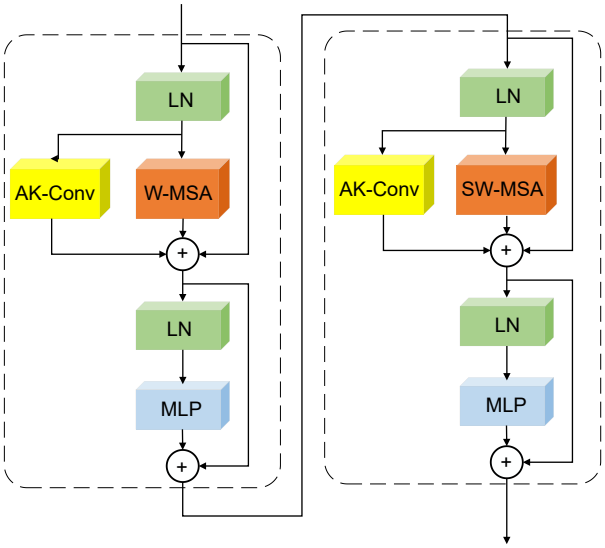


图 6 Swin-AK blocks 的结构图

Fig. 6 Swin-AK blocks architecture diagram

像中的细节和纹理信息。AKConv 模块首先使用一个卷积层计算出偏移量，并利用偏移量对输入特征进行重新采样；然后，使用双线性插值把来自四个相邻点的特征值结合起来；接着，使用卷积层对重新采样后的特征图进行卷积操作；最后使用批归一化操作和 Sigmoid 门控线性单元 (sigmoid gated linear unit, SiLU) 激活函数进一步提升特征的表达能力。AKConv 能够更好地提取复杂的图像特征，从而提高了本文方法对图像质量的评价能力。

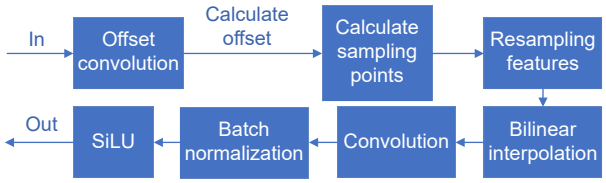


图 7 AKConv 的结构图

Fig. 7 AKConv architecture diagram

为实现跨窗口的信息交互，Swin-AK Transformer 在相邻层中对窗口的位置进行偏移操作，将窗口在水平方向和垂直方向上分别移动$\left\lfloor \frac{M}{2} \right\rfloor$个单位 (⌊⌋表示向下取整符号，取不大于$\frac{M}{2}$的最大整数)，其中窗口的尺寸为$M \times M$，如式 (9) 所示。

$$X_{\text{shift}} = \text{shift}\left[X, \left(-\left\lfloor \frac{M}{2} \right\rfloor, -\left\lfloor \frac{M}{2} \right\rfloor\right)\right], \tag{9}$$

其中：X 表示输入特征图；X_{shift} 表示滑动后的特征图；shift[·] 表示窗口的滑动操作。

对滑动后的特征图重新划分窗口，并使用多头自注意力机制，如式 (10) 所示。

$$X_{\text{shift}} \to \{X_{\text{shift},i}\}_{i=1}^{H}, \tag{10}$$

其中：$X_{\text{shift},i}$ 表示第 i 个滑动后的窗口；H 表示窗口的数量。

将滑动窗口后的输出结果和原窗口的输出结果进行合并，形成最终的特征图。

2.3 双交叉注意力特征融合网络

本文提出的双交叉注意力特征融合网络 (dual attention cross fusion, DACF) 使用了通道注意力机制和空间注意力机制，将手工提取的特征和 Swin-AK Transformer 提取的特征融合，显著提升了模型对关键信息的捕获能力。

在 DACF 模块中，通道注意力机制专注于优化特征图的不同通道响应。该机制使用自适应地分配各通道的权重，抑制不重要的特征，增强对关键通道的关注，从而确保模型能够从多个通道中提取具有较强判别力的特征。与此同时，空间注意力机制则侧重于识别图像中的重要区域。该机制通过提取局部空间信息，优化特征图在空间维度上的分布。它能够有效过滤掉背景噪声，同时保留关键的结构和细节信息。

DACF 模块的创新在于将上述两种注意力机制有机结合，使得融合后的特征不仅在通道维度上得到优化，还在空间维度上进行了精细调整，从而实现了多维度的特征感知。这种融合方式有效提高了模型对图像整体和局部细节的敏感度，确保模型能够兼顾全局结构和局部信息。使用这种双重注意力机制，DACF 模块能够更好地捕捉图像中的复杂信息，提升特征表达的鲁棒性和准确性。

此外，DACF 模块的自适应调整能力使其能够根据图像的内容和特性灵活处理不同类型的图像。该机制在多种数据集上的高效表现表明，适应智能手机拍摄图像中多样化的失真。这种灵活性和自适应性显著提高了本文方法在智能手机图像质量评价中的泛化能力。

双交叉注意力模块的结构如图 8 所示。其输入的特征由手工提取的特征与 Swin-AK Transformer 提取的特征组成。首先，通道注意力模块分别处理这两部分特征，通过 Softmax 函数进行初步融合，之后将融合后的特征输入空间注意力模块进行进一步优化。最终，优化后的特征与原始输入特征再次融合，从而生成更加全面的特征表示。

通道注意力模块的结构如图 9 所示。在通道注意力模块中，首先分别使用最大池化操作和平均池化操作提取该模块输入信号的全局信息；然后，使用 MLP 进行处理，生成通道权重，用于重新加权输入特征，从而突出重要的通道特征；最后，把加权后的两个特征组合在一起，作为通道注意力模块的输出结果。

空间注意力模块的结构如图 10 所示。空间注意力模块首先使用平均池化操作和最大池化操作；然后使用 CNN 提取特征的空间信息；最后，把 CNN 的结果作为空间注意力模块的输出结果。

DACF 模块对图像特征的通道和空间维度的双重优化，保证了特征融合过程的全面性。结合通道和空间的双重注意力机制，不仅增强了模型对全局信息的把握，也提升了细节特征的捕捉能力。这种多层次的特征融合策略更加符合人眼的主观感知特性，确保了智能手机拍摄图像质量评价的高准确性和鲁棒性。

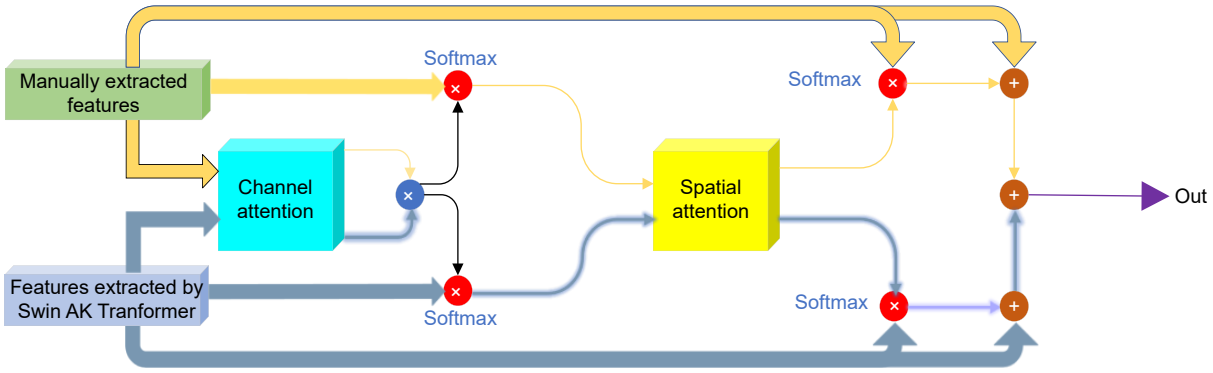


图 8 双交叉注意力特征融合模块的结构示意图

Fig. 8 Structure diagram of the dual attention cross fusion module

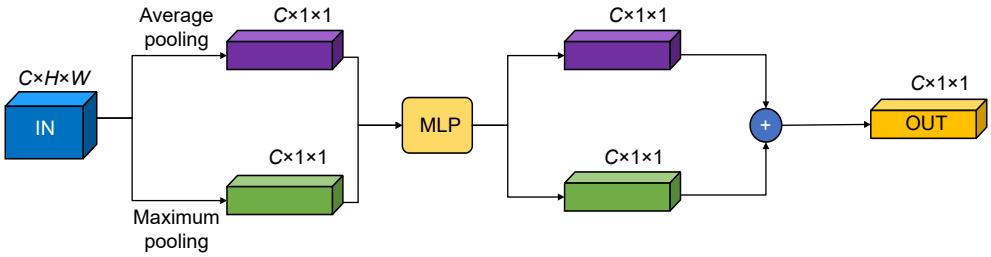


图 9 通道注意力模块的结构示意图

Fig. 9 Channel attention network structure diagram

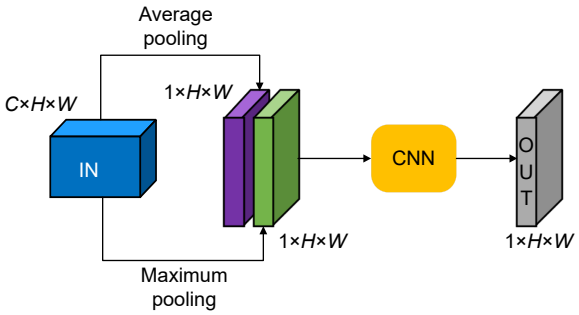


图 10 空间注意力模块的结构示意图

Fig. 10 Structure diagram of the spatial attention module

3 实验及结果分析

3.1 数据集及评价指标

本文采用 LIVE-C 数据集和 SPAQ 数据集进行实验。

LIVE-C[20] (LIVE challenge) 数据集包含 1170 个图像，各种移动摄像设备在不同条件下进行拍摄而得到了这些图像。这些图像在获取过程中受到了多种失真类型的影响。LIVE-C 数据集中的图像由超过 8100 名参与者在严格监控的众包研究中进行评分。

SPAQ[7] (Smartphone photography attribute and quality) 数据集是一个专门用于评价智能手机摄影质量的大规模图像数据集。该数据集包含了 11125 张由 66 款智能手机拍摄的照片，每张照片包含主观评分、亮度评分、锐度评分和色彩度评分等。SPAQ 数据集的主观评分实验由超过 600 名受试者参与。

在 SPAQ 数据集中，图像的三种属性评分与整体主观质量评分之间的散点图如图 11 所示。通过分析 SPAQ 数据集的亮度评分、锐度评分和色彩度评分，可以发现这三种图像属性评分均与图像的主观质量评分存在一定的正相关关系。亮度评分较高的图像往往具有较高的质量评分，锐度评分与质量评分之间的相关性最为显著，色彩度评分也表现出明显的正相关趋势。然而，并不是所有数据点都严格遵循这一关系，这表明还有其它因素影响图像质量评分。此外，对亮度、锐度和色彩度的评分是为了更好地模拟人眼的视觉感知，从而提供更全面的图像质量评价。

本文采用的评价指标是斯皮尔曼等级排序相关系数 (Spearman rank order correlation coefficient, SROCC) 和皮尔森线性相关系数 (Pearson linear correlation coefficient, PLCC)。*SROCC* 是一种衡量两个向量排序一致性的非参数统计度量，它反映了图像的质量预测分数与主观质量评分之间的排序相关性，如式 (11) 所示。

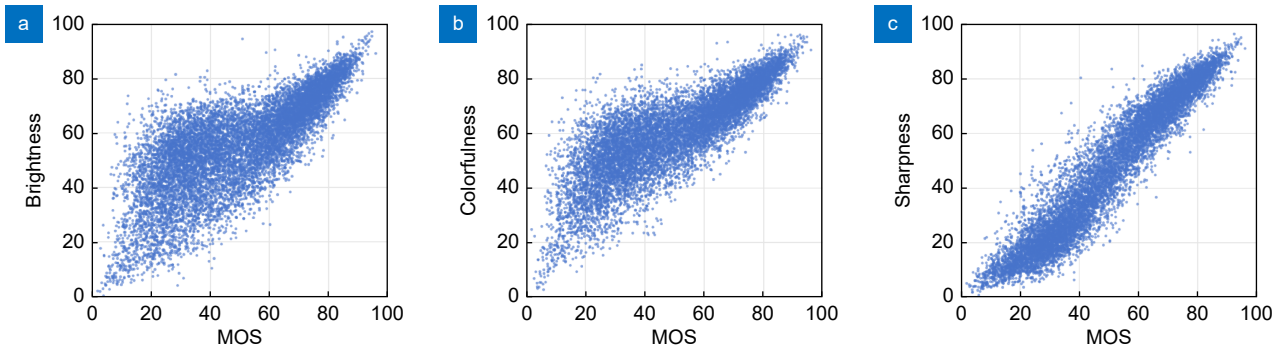


图 11 SPAQ 数据集中图像属性评分与整体主观质量评分之间的散点图。(a) 亮度; (b) 色彩度; (c) 锐度

Fig. 11 Scatter plot of image attribute scores versus overall subjective quality scores in the SPAQ. (a) Brightness; (b) Colorfulness; (c) Sharpness

$$SROCC = 1 - \frac{6\sum_{i=1}^{N_t}(\boldsymbol{x}_i - \boldsymbol{y}_i)^2}{N_t(N_t^2 - 1)}, \tag{11}$$

式中：$\boldsymbol{x}_i$、$\boldsymbol{y}_i$分别表示主观评价向量、客观评价向量分别按相同顺序 (由小到大或由大到小) 排序后，对于第 i 个成绩在各自序列中的序号；N_t 表示图像的数量。

PLCC 是一种衡量两个向量线性相关性的统计度量，它反映了图像的质量预测分数与主观质量评分之间的线性关系。同时使用这两个指标可以有效地评价图像质量评价模型的预测性能，如式 (12) 所示。

$$PLCC = \frac{\sum_{i=1}^{N_t}(s_i - \bar{s})(p_i - \bar{p})}{\sqrt{\sum_{i=1}^{N_t}(s_i - \bar{s})^2 \sum_{i=1}^{N_t}(p_i - \bar{p})^2}}, \tag{12}$$

式中：s_i、p_i分别表示第 i 幅图像的主观质量分数、客观质量分数；$\bar{s}$、$\bar{p}$分别表示s_i、p_i的平均值。

3.2 散点图分析

在本文的实验中，随机选取了 LIVE-C 和 SPAQ 两个数据集中的 80% 和 20% 作为训练集和测试集。实验使用 Python 3.11.7 语言，并使用 Pytorch2.2.1 深度学习框架进行编译和运行。CPU 为 Intel Core i5-6500，计算机内存的容量为 16.0 GB，GPU 为 NVIDIA GTX 4090Ti，其显存的容量为 24 GB。图 12 和图 13 分别展示了在 LIVE-C 和 SPAQ 数据集上本文方法的质量预测分数与主观质量评分之间的散点图。从这两张图可以看出，这两张图均显示出主观质量评分值 (mean opinion score, MOS) 与本文方法的质量预测值具有正相关的关系。

在 LIVE-C 数据集中，散点数量较少，散点的分布较为稀疏。而在 SPAQ 数据集中，散点的数量多，散点分的布更为密集，显示出更为明显的线性趋势。这表明在 SPAQ 数据集上，本文方法的预测性能更加稳定和准确。

3.3 对比实验

为了验证本文方法的性能，将本文方法与 16 种无参考质量评价方法进行比较，其中 6 种方法是基于手工特征提取的方法，它们分别是 BLINDS-II[21]、DIIVINE[22]、BRISQUE[23]、CORNIA[24]、IL-NIQE[25] 和 HOSA[26]；9 种方法是基于深度学习的方法，它们分别是 DIQaM-NR[27]、WaDIQaM-NR[27]、TS-CNN[28]、

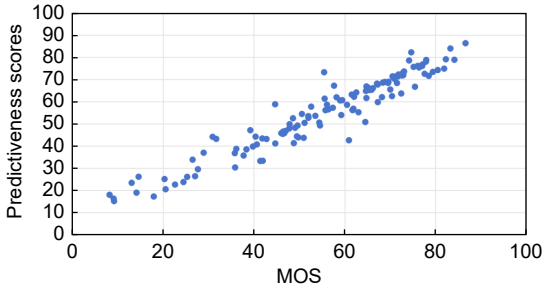


图 12 LIVE-C 数据集上的散点图

Fig. 12 Scatter plot on the LIVE-C dataset

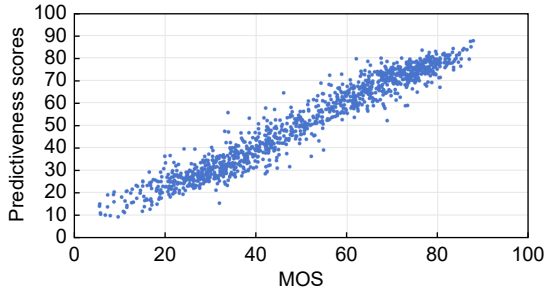


图 13 SPAQ 数据集上的散点图

Fig. 13 Scatter plot on the SPAQ dataset

TReS[29]、DB-CNN[30]、HyperIQA[31]、CaHDC[32]、MT-A[7]、MUSIQ[2]、DACNN[33]、Re-IQA[34]、DEIQT[35]和 LoDa[36]。在 SPAQ 和 LIVE-C 数据集上，本文方法与其它方法的对比分别如表 1 和表 2 所示。

表 1 给出了本文方法在 SPAQ 数据集上与其它方法的 SROCC 和 PLCC 对比结果。可以看到，本文方法在两个指标上均表现出色，PLCC 的值和 SROCC 的值分别为 0.932 和 0.929。这表明本文方法在预测图像质量方面具有更高的相关性和一致性，能够更准确地反映主观质量的评分。此外，本文的实验结果验证了本文提出的本文方法在智能手机拍摄图像质量评价任务中具有较好的稳健性和可靠性。

表 2 给出了本文方法在 LIVE-C 数据集上与其它方法的 SROCC 和 PLCC 对比结果。从表 2 中可以看出，本文方法同样在两个指标上均表现优异，PLCC 的值和 SROCC 的值分别为 0.885 和 0.865。这表明本文方法在不同数据集上具有较好的适应性和泛化能力，特别是与 BLINDS-II、DIIVINE 这两种传统的无参考质量评价方法相比，本文方法的 PLCC 和 SROCC 分别提高了 0.388 和 0.409、0.328 和 0.352，这表明本文提出的本文方法在智能手机拍摄图像质量评价任务中具有显著优势。

表 1　本文方法在 SPAQ 数据集上与其它方法的对比

Table 1　Comparison of the proposed method with other methods on the SPAQ dataset

Methods	PLCC	SROCC
BLINDS-II[21]	0.539	0.478
DIIVINE[22]	0.603	0.596
BRISQUE[23]	0.817	0.828
CORNIA[24]	0.724	0.709
IL-NIQE[25]	0.704	0.695
HOSA[26]	0.824	0.817
DIQaM-NR[27]	0.836	0.824
WaDIQaM-NR[27]	0.843	0.821
TS-CNN[28]	0.811	0.801
TReS[29]	0.911	0.902
DB-CNN[30]	0.913	0.909
HyperIQA[31]	0.919	0.916
CaHDC[32]	0.841	0.833
ResNet50[7]	0.909	0.908
MT-A[7]	0.916	0.916
MUSIQ[2]	0.921	0.917
DACNN[33]	0.921	0.915
Re-IQA[34]	0.925	0.918
DEIQT[35]	0.923	0.919
LoDa[36]	0.928	0.925
Ours	**0.932**	**0.929**

3.4 消融实验

为了评价本文方法中的各个模块对此模型整体性能产生的影响，本文在 SPAQ 数据集上进行了一系列的消融实验，实验结果如表 3 所示。

从表 3 可以看出，首先，基础的 Swin Transformer 模型在 SPAQ 数据集上的 PLCC 和 SROCC 分别为

表 2　本文方法在 LIVE-C 数据集上与其它方法的对比

Table 2　Comparison of the proposed method with other methods on the LIVE-C dataset

Methods	PLCC	SROCC
BLINDS-II[21]	0.497	0.456
DIIVINE[22]	0.557	0.513
BRISQUE[23]	0.637	0.616
CORNIA[24]	0.659	0.617
IL-NIQE[25]	0.516	0.539
HOSA[26]	0.691	0.674
DIQaM-NR[27]	0.645	0.633
WaDIQaM-NR[27]	0.692	0.669
TS-CNN[28]	0.667	0.655
TReS[29]	0.877	0.846
DB-CNN[30]	0.859	0.852
HyperIQA[31]	0.870	0.855
CaHDC[32]	0.738	0.734
MUSIQ[2]	0.875	0.862
DACNN[33]	0.882	0.861
Re-IQA[34]	0.854	0.84
Ours	**0.885**	**0.865**

表 3　消融实验的结果

Table 3　Results of the ablation experiment

Model	PLCC	SROCC
Swin Transformer	0.921	0.918
Swin-AK Transformer	0.923	0.920
Manual features+Swin Transformer	0.924	0.922
Manual features+Swin-AK Transformer	0.929	0.925
Ours	**0.932**	**0.929**

0.921 和 0.918；然后，Swin-AK Transformer 模型的 PLCC 和 SROCC 分别提高到了 0.923 和 0.920，Swin Transformer 和 Swin-AK Transformer 的注意力热力图如图 14 所示。从图 14 可以看出，和 Swin Transformer 模型相比，Swin-AK Transformer 模型显著增强了对局部特征的捕捉能力。

其次，将 Swin Transformer 模型和提取手工特征融合后，PLCC 和 SROCC 的值提高到 0.924 和 0.922；接着，当进一步将 Swin-AK Transformer 融合手工特征后，PLCC 和 SROCC 分别增至 0.929 和 0.925，这反映出手工特征对提升模型预测精度的重要性；最后，使用双交叉注意力特征融合网络将 Swin-AK Transformer 和手工特征融合后，本文方法在 PLCC 和 SROCC 上都表现出了最佳的性能，这表明本文提出的方法具有更好地细节捕捉能力，能够更好地符合人眼视觉感知的特点。

3.5 图像分析实验

为了进一步验证本文方法的性能，本文在 SPAQ 数据集的测试集中按照 MOS 值从小到大的顺序选取了 8 张图像。对于这些图像，本文方法的质量预测值、MUSIQ[2] 方法的质量预测值与 MOS 值如图 15 所示。从图 15 可以看出，本文方法的预测值相比于 MUSIQ[2] 方法的预测值能够更准确地反映图像的主观质量评分。无论是白天拍摄的城市景观、夜景、室内的食物照片，还是模糊的图像，本文方法的质量预测值均接近实际的 MOS 值。本文方法的均方误差 (mean squared error,

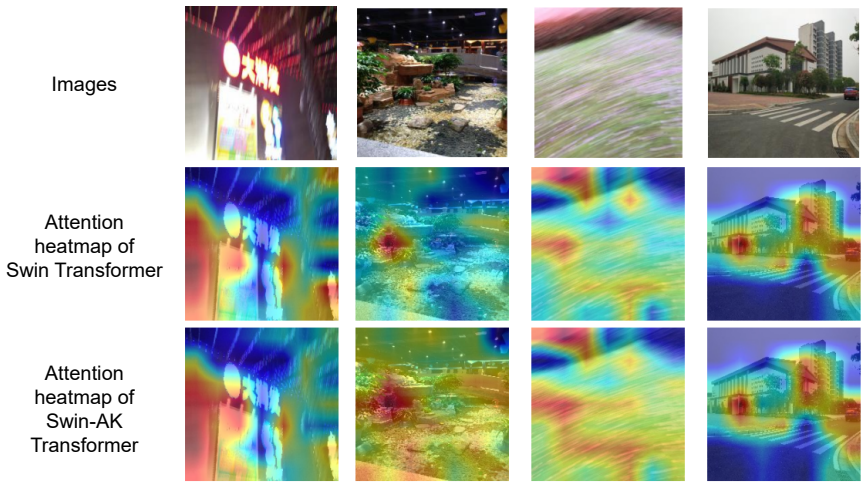


图 14 Swin Transformer 和 Swin-AK Transformer 的注意力热力图对比

Fig. 14 Comparison of attention heatmaps between Swin Transformer and Swin-AK Transformer

Images				
MOS	15.3	28.25	37.75	40
MUSIQ	25.04	18.89	40.26	16.55
Ours	16.82	29.17	42.83	33.53

Images				
MOS	57	63	86.33	90.38
MUSIQ	52.38	43.62	72.95	68.02
Ours	53.76	58.10	84.28	88.75

图 15 SPAQ 数据集中图像的 MOS 值与本文方法的质量预测值

Fig. 15 MOS values of images in the SPAQ dataset and the quality prediction values of the proposed method

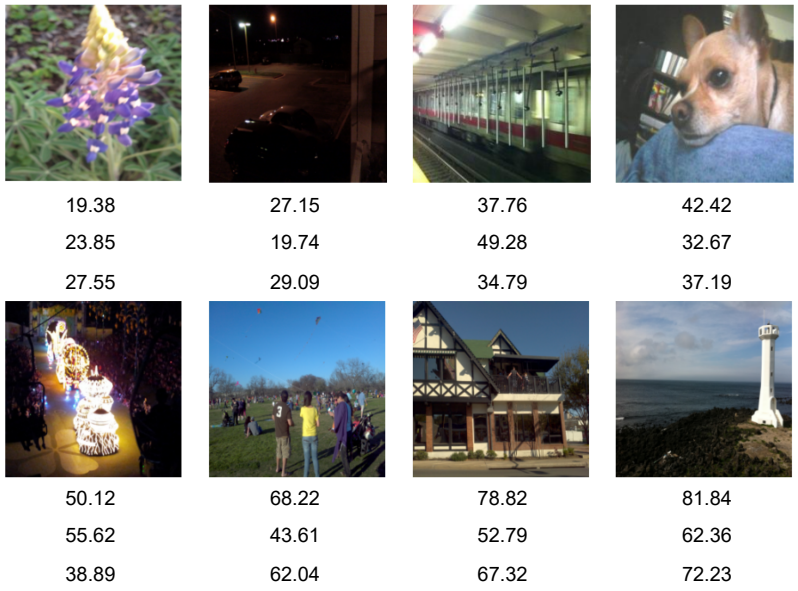


图 16 LIVE-C 数据集中图像的 MOS 值与本文方法的质量预测值

Fig. 16 MOS values of images in the LIVE-C dataset and the quality prediction values of the proposed method

MSE) 值为 14.02，远低于 MUSIQ[2] 方法的 MSE 值 226.83。这些结果表明，本文方法在不同类型图像上的表现更加稳定，能够较好地反映人类视觉对图像质量的感知特点。同时，本文方法对高质量和低质量图像的预测均表现出较高的精度，进一步验证了其在图像质量评估任务中的优越性能。

在 LIVE-C 数据集的测试集中，按照 MOS 值从小到大的顺序选取了 8 张图像，这些图像的 MOS 值、本文方法的质量预测值和 MUSIQ[2] 方法的质量预测值如图 16 所示。通过对比这些图像，可以看出本文方法在各种不同拍摄场景下的预测效果。从夜间的灯光展示、白天的户外活动，到风景和植物的拍摄，本文方法的质量预测值非常接近于实际的 MOS 值。即便在光照变化较大的场景下，例如夜景和逆光，本文方法依然能够提供接近实际评分的预测值。对于这 8 张图像，本文方法的 MSE 的值为 61.95，明显低于 MUSIQ[2] 方法的 MSE 值 249.45。但是，由于 LIVE-C 数据集的图像数据数量较少，导致预测值与实际值之间的差异略大于 SPAQ 数据集。总体而言，实验结果验证了本文方法在多样化拍摄条件下具有较好的鲁棒性，并表明其在不同环境下进行图像质量评价具有较好的可靠性和稳定性。

4 结 论

本文针对目前已有的智能手机拍摄图像无参考质量评价算法存在的不足，提出了一种基于手工特征和 Swin-AK Transformer 双交叉注意力融合网络的智能手机拍摄图像质量评价方法。首先，本文根据人眼视觉感知的特性，提取影响图像质量的手工特征。为了学习手工特征和图像质量之间的非线性映射，本文在手工特征提取后加入了 ResNet50，将手工提取的低级特征转化为更抽象的高级特征；然后，使用 Swin-AK Transformer 中的自注意力机制捕获图像的局部特征，增强了对局部空间信息的处理能力；最后，设计了双交叉注意力融合模块，把上述得到的两种特征进行融合，确保在融合过程中不仅关注图像的全局信息，还能精确地捕捉到图像中的细节变化。在 SPAQ 和 LIVE-C 数据集上的实验结果表明本文提出的方法能够精准地预测智能手机拍摄图像的质量，并符合人眼视觉感知的特点。

课题的后续研究将致力于智能手机拍摄图像的质量增强，使智能手机拍摄图像的质量接近甚至达到单反相机拍摄图像的质量。这包括提升图像的清晰度、色彩还原、细节表现和在各种拍摄条件下 (如低光、高动态范围等) 的整体质量。这不仅能够为普通用户提供更高质量的拍摄体验，也为专业摄影和图像处理应用提供了更广泛的研究方向。

参考文献

[1] Yan J B, Fang Y M, Liu X L. The review of distortion-related image quality assessment[J]. *J Image Graphics*, 2022, **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[J]. 中国图象图形学报, 2022, **27** (5): 1430−1466.

[2] Ke J J, Wang Q F, Wang Y L, et al. MUSIQ: multi-scale image quality transformer[C]//*2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021: 5128–5137. https://doi.org/10.1109/ICCV48922.2021.00510.

[3] Varga D. No-reference image quality assessment using the statistics of global and local image features[J]. *Electronics*, 2023, **12** (7): 1615.

[4] Jain P, Shikkenawis G, Mitra S K. Natural scene statistics and CNN based parallel network for image quality assessment[C]//*2021 IEEE International Conference on Image Processing (ICIP)*, 2021: 1394–1398. https://doi.org/10.1109/ICIP42928.2021.9506404.

[5] Shao X, Liu M Q, Li Z H, et al. CPDINet: blind image quality assessment via a content perception and distortion inference network[J]. *IET Image Processing*, 2022, **16** (7): 1973−1987.

[6] Zhao K, Yuan K, Sun M, et al. Quality-aware pretrained models for blind image quality assessment[C]//*Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023: 22302–22313. https://doi.org/10.1109/CVPR52729.2023.02136.

[7] Fang Y M, Zhu H W, Zeng Y, et al. Perceptual quality assessment of smartphone photography[C]//*Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020: 3674–3683. https://doi.org/10.1109/CVPR42600.2020.00373.

[8] Yuan Z F, Qi Y, Hu M H, et al. Opinion-unaware no-reference image quality assessment of smartphone camera images based on aesthetics and human perception[C]//*2020 IEEE International Conference on Multimedia & Expo Workshops*, 2020: 1–6. https://doi.org/10.1109/ICMEW46912.2020.9106048.

[9] Zhou Y W, Wang Y L, Kong Y Y, et al. Multi-Indicator image quality assessment of smartphone camera based on human subjective behavior and perception[C]//*2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2020: 1–6. https://doi.org/10.1109/ICMEW46912.2020.9105971.

[10] Huang C H, Wu J L. Multi-task deep CNN model for no-reference image quality assessment on smartphone camera photos[Z]. arXiv: 2008.11961, 2020. https://arxiv.org/abs/2008.11961.

[11] Yao C, Lu Y R, Liu H, et al. Convolutional neural networks based on residual block for no-reference image quality assessment of smartphone camera images[C]//*Proceedings of the 2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2020: 1–6. https://doi.org/10.1109/ICMEW46912.2020.9106034.

[12] Wang B, Bai Y Q, Zhu Z J, et al. No-reference light field image quality assessment based on joint spatial-angular information[J]. *Opto-Electron Eng*, 2024, **51** (9): 240139.
王斌, 白永强, 朱仲杰, 等. 联合空角信息的无参考光场图像质量评价[J]. 光电工程, 2024, **51** (9): 240139.

[13] 陈松, 温宇鑫, 安浩铭. 基于多尺度双边滤波 Retinex 的非均匀光照散斑图像矫正[J/OL]. 激光技术, 1-16 [2025-01-18]. http://kns.cnki.net/kcms/detail/51.1125.TN.20240116.1129.004.html.

[14] Liu J, Tang J L, Lin B, et al. Rust spot image recognition of coatings based on HSV and shape feature[J]. *China Surf Eng*, 2023, **36** (4): 217−228.
刘佳, 唐鋆磊, 林冰, 等. 基于 HSV (色相-饱和度-明度) 与形状特征的涂层锈点图像识别[J]. 中国表面工程, 2023, **36** (4): 217−228.

[15] Liu Q G, Liu P, Wang Y H, et al. Semi-parametric decolorization with Laplacian-based perceptual quality metric[J]. *IEEE Trans Circuits Syst Video Technol*, 2017, **27** (9): 1856−1868.

[16] Ma C H, Hu W H, Zhong H C, et al. SAR image structure optimization method using Sobel operator fusion[J]. *J Detect Control*, 2024, **46** (2): 119−124.
马常昊, 胡文惠, 钟海超, 等. 融合 Sobel 算子的 SAR 图像结构优化方法[J]. 探测与控制学报, 2024, **46** (2): 119−124.

[17] Luo X Y, Hu Z, Tang W C, et al. Species identification of ore particles combined with Fourier and LBP descriptors[J]. *Transducer Microsyst Technol*, 2023, **42** (11): 147−150.
罗小燕, 胡振, 汤文聪, 等. 傅里叶和 LBP 描述子相结合的矿石颗粒种类识别[J]. 传感器与微系统, 2023, **42** (11): 147−150.

[18] Liu Z, Lin Y T, Cao Y, et al. Swin Transformer: hierarchical vision transformer using shifted windows[C]//*Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021: 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986.

[19] Wang G P, Li X, Jia X F, et al. STransMNet: a stereo matching method with Swin Transformer fusion[J]. *Opto-Electron Eng*, 2023, **50** (4): 220246.
王高平, 李珣, 贾雪芳, 等. 融合 Swin Transformer 的立体匹配方法 STransMNet[J]. 光电工程, 2023, **50** (4): 220246.

[20] Ghadiyaram D, Bovik A C. Massive online crowdsourced study of subjective and objective picture quality[J]. *IEEE Trans Image Process*, 2016, **25** (1): 372−387.

[21] Saad M A, Bovik A C, Charrier C. Blind image quality assessment: a natural scene statistics approach in the DCT domain[J]. *IEEE Trans Image Process*, 2012, **21** (8): 3339−3352.

[22] Moorthy A K, Bovik A C. Blind image quality assessment: from natural scene statistics to perceptual quality[J]. *IEEE Trans Image Process*, 2011, **20** (12): 3350−3364.

[23] Mittal A, Moorthy A K, Bovik A C. No-reference image quality assessment in the spatial domain[J]. *IEEE Trans Image Process*, 2012, **21** (12): 4695−4708.

[24] Ye P, Kumar J, Kang L, et al. Unsupervised feature learning framework for no-reference image quality assessment[C]//*Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012: 1098–1105. https://doi.org/10.1109/CVPR.2012.6247789.

[25] Zhang L, Zhang L, Bovik A C. A feature-enriched completely blind image quality evaluator[J]. *IEEE Trans Image Process*, 2015, **24** (8): 2579−2591.

[26] Xu J T, Ye P, Li Q H, et al. Blind image quality assessment based on high order statistics aggregation[J]. *IEEE Trans Image Process*, 2016, **25** (9): 4444−4457.

[27] Bosse S, Maniry D, Müller K R, et al. Deep neural networks for no-reference and full-reference image quality assessment[J]. *IEEE Trans Image Process*, 2018, **27** (1): 206−219.

[28] Yan Q S, Gong D, Zhang Y N. Two-stream convolutional networks for blind image quality assessment[J]. *IEEE Trans Image Process*, 2019, **28** (5): 2200−2211.

[29] Golestaneh S A, Dadsetan S, Kitani K M. No-reference image quality assessment via transformers, relative ranking, and self-consistency[C]//*Proceedings of the 2022 IEEE/CVF Winter*

Conference on Applications of Computer Vision, 2022: 3989–3999. https://doi.org/10.1109/WACV51458.2022.00404.

[30] Zhang W X, Ma K D, Yan J, et al. Blind image quality assessment using a deep bilinear convolutional neural network[J]. *IEEE Trans Circuits Syst Video Technol*, 2020, **30** (1): 36−47.

[31] Su S L, Yan Q S, Zhu Y, et al. Blindly assess image quality in the wild guided by a self-adaptive hyper network[C]//*Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020: 3664–3673. https://doi.org/10.1109/CVPR42600.2020.00372.

[32] Wu J J, Ma J P, Liang F H, et al. End-to-end blind image quality prediction with cascaded deep neural network[J]. *IEEE Trans Image Process*, 2020, **29**: 7414−7426.

[33] Pan Z Q, Zhang H, Lei J J, et al. DACNN: Blind image quality assessment via a distortion-aware convolutional neural network[J]. *IEEE Trans Circuits Syst Video Technol*, 2022, **32** (11): 7518−7531.

[34] Saha A, Mishra S, Bovik A C. Re-IQA: Unsupervised learning for image quality assessment in the wild[C]//*Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023: 5846–5855. https://doi.org/10.1109/CVPR52729.2023.00566.

[35] Qin G Y, Hu R Z, Liu Y T, et al. Data-efficient image quality assessment with attention-panel decoder[C]//*Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*, 2023: 2091–2100. https://doi.org/10.1609/aaai.v37i2.25302.

[36] Xu K M, Liao L, Xiao J, et al. Boosting image quality assessment through efficient transformer adaptation with local feature enhancement[C]//*Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024: 2662–2672. https://doi.org/10.1109/CVPR52733.2024.00257.

作者简介

侯国鹏 (1999-)，男，硕士研究生，主要研究方向为智能手机拍摄图像质量评价。

E-mail：18910915164@163.com

【通信作者】董武 (1980-)，男，北京印刷学院信息工程学院副教授，硕士生导师，主要研究方向为深度学习、图像质量评价。

E-mail：dongwu@bigc.edu.cn

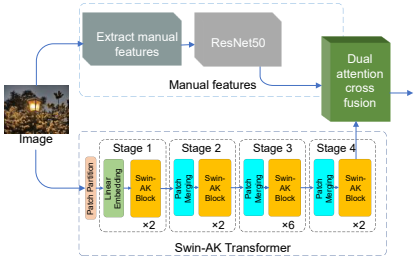
陆利坤 (1971-)，男，北京印刷学院信息工程学院副教授，硕士生导师，主要研究方向为喷墨数字印刷技术。

E-mail：lklu@bigc.edu.cn

扫描二维码，获取PDF全文

Smartphone image quality assessment method based on Swin-AK Transformer

Hou Guopeng, Dong Wu*, Lu Likun, Zhou Ziyi, Ma Qian, Bai Zhen, Zheng Shenghui



The overall structure diagram of the method

Overview: With the extensive use of smartphones, users' expectations for smartphone image quality have risen significantly. However, due to limitations in camera hardwares, smartphones often face constraints in light capture, especially in complex or low-light scenarios, which can lead to image quality degradation. Existing no-reference image quality assessment (IQA) algorithms frequently show limitations when handling smartphone-captured images, motivating the development of a more accurate quality evaluation method. This study proposes an approach based on manual features and a Swin-AK Transformer with dual cross-attention fusion, designed to assess smartphone image quality with greater precision. First, manual features affecting image quality are extracted, guided by the human visual system, enabling the capture of subtle visual variations such as color, contrast, and texture, which enhances the model's sensitivity to image quality. To further improve the discriminative power for image quality assessment, ResNet50 is introduced after manual feature extraction to establish a nonlinear mapping between manual features and image quality. This process transforms initial low-level features into more representative high-level features, allowing for a more comprehensive expression of image content. Subsequently, the study introduces the Swin-AK Transformer, which utilizes a self-attention mechanism to capture local image features, thereby enhancing the model's capability to recognize and process local information in smartphone images. This method effectively adapts to the unique characteristics of smartphone images, offering robust handling of intricate details. Additionally, a dual cross-attention fusion module is designed to integrate manual and deep features efficiently. The module combines spatial and channel attention mechanisms: spatial attention aids the model in focusing on key areas within the image, while channel attention optimizes feature representation by adjusting the weights of each channel. As a result, the fused features reflect both global image information and local detail variations, aligning well with the human visual system's natural perception of image quality. Experiments were conducted on two public datasets, SPAQ and LIVE-C, to evaluate the proposed model. The results demonstrate the model's superior performance in image quality prediction, achieving Pearson correlation coefficients of 0.932 and 0.885 and Spearman rank correlation coefficients of 0.929 and 0.858 on the SPAQ and LIVE-C datasets, respectively. These outcomes validate the proposed method's effectiveness in smartphone image quality assessment tasks, showcasing improved sensitivity to quality changes and excellent accuracy and robustness.

Hou G P, Dong W, Lu L K, et al. Smartphone image quality assessment method based on Swin-AK Transformer[J]. *Opto-Electron Eng*, 2025, **52**(1): 240264; DOI: 10.12086/oee.2025.240264

Foundation item: The Important Project of Digital Education Research of Beijing (BDEC2022619027), 2023 Project Proposal of Beijing Higher Education Association (MS2023168), the Research Project of Beijing Institute of Graphic Communication (Ec202303, Ea202301, E6202405), the Disciplinary Construction and Postgraduate Education Project of Beijing Institute of Graphic Communication (21090323009, 21090224002, 21090124013), Classification Development of Beijing Municipal Universities-Construction Project of Emerging Interdisciplinary Platform for Publishing at Beijing Institute of Graphic Communication-Key Technology Research and Development Platform for Digital Inkjet Printing Technology and Multifunctional Rotary Offset Press (04190123001/003), Open Foundation of State key Laboratory of Networking and Switching Technology (Beijing University of Posts and Telecommunications) (SKLNST-2023-1-12) , and the Project of the "Artificial Intelligence Plus" Course Construction of Beijing Institute of Graphic Communication

Beijing Key Laboratory of Signal and Information Processing for High-end Printing Equipment, Beijing Institute of Graphic Communication, Beijing 102600, China

* E-mail: dongwu@bigc.edu.cn