基于对比性视觉-文本模型的光场图像质量评估

王汉灵[1,2],柯 逍[1,3,4*],江澳鑫[1,3,4],郭文忠[1,3,4]

(1. 福州大学计算机与大数据学院,福建福州 350116;
2. 中国地震局工程力学研究所地震工程与工程振动重点实验室,黑龙江哈尔滨 150080;
3. 福建省网络计算与智能信息处理重点实验室,福建福州 350116;4. 大数据智能教育部工程研究中心,福建福州 350116)

摘 要: 光场图像作为一种能够捕获场景每个位置光线信息的图像类型,在电子成像、医学影像和虚拟现实等领域具有广泛的应用前景. 光场图像质量评估(Light Field Image Quality Assessment,LFIQA)旨在衡量此类图像的质量,但当前方法面临视觉效果与文本模态间异构性的重要挑战. 为解决上述问题,本文提出了一种基于文本-视觉的多模态光场图像质量评估模型. 具体来说,在视觉模态方面,我们设计了多任务模型,结合边缘自动阈值算法有效丰富了光场图像的关键表示特征. 在文本模态方面,基于输入噪声特征与预测噪声特征的对比,准确识别光场图像的噪声类别,并验证了噪声预测对优化视觉表示的重要性. 基于上述研究,进一步提出了一种优化的通用噪声文本配置方法,并结合边缘增强策略,显著提升了基线模型在光场图像质量评估中的准确性和泛化能力. 此外,通过消融实验,评估了各组件对整体模型性能的贡献,验证了本文方法的有效性和稳健性. 实验结果表明,该方法不仅在公开数据集Win5-LID和NBU-LF1.0的实验中表现出色,还在融合数据集中展示出优秀的实验结果,与现有最优算法相比,本文所提方法在两个数据库中的性能分别提升了2%和6%. 本文提出的噪声验证策略和配置方法不仅为图像质量评估中的噪声预测任务提供了有价值的参考,也可用于其它噪声预测类型的辅助任务.

关键词: 图像质量评估;光场图像;视觉-文本模型;多任务模式;噪声预测;图像增强

基金项目: 国家重点研发计划(No.2021YFB3600503);国家自然科学基金(No.61972097,No.U21A20472);福建省科技重大专项(No.2021HZ022007);福建省自然科学基金(No.2021J01612)

中图分类号: TP183 **文献标识码:** A **文章编号:** 0372-2112(2024)10-3562-16

电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20240533

Quality Assessment of Light Field Images Based on Contrastive Visual-Textual Model

WANG Han-ling[1,2], KE Xiao[1,3,4*], JIANG Ao-xin[1,3,4], GUO Wen-zhong[1,3,4]

(1. *College of Computer and Data Science*, *Fuzhou University*, *Fuzhou*, *Fujian* 350116, *China*;
2. *Key Laboratory of Earthquake Engineering and Engineering Vibration*, *Institute of Engineering Mechanics*, *China Earthquake Administration*, *Harbin*, *Heilongjiang* 150080, *China*;
3. *Fujian Provincial Key Laboratory of Networking Computing and Intelligent Information Processing*, *Fuzhou University*, *Fuzhou*, *Fujian* 350116, *China*;
4. *Engineering Research Center of Big Data Intelligence*, *Ministry of Education*, *Fuzhou*, *Fujian* 350116, *China*)

Abstract: Light field imaging, as an image type capable of capturing light information from every position in a scene, holds broad application prospects in fields such as electronic imaging, medical imaging, and virtual reality. Light field image quality assessment (LFIQA) aims to measure the quality of such images, yet current methods confront significant challenges arising from the heterogeneity between visual effects and textual modalities. To address these issues, this paper proposes a multi-modal light field image quality assessment model grounded in text-vision integration. Specifically, for the visual modality, we devise a multi-task model that effectively enriches the crucial representational features of light field images by incorporating an edge auto-thresholding algorithm. On the textual side, we accurately identify noise categories in light field images based on the comparison between input noise features and predicted noise features, thereby validating the importance of noise prediction in optimizing visual representations. Building upon these findings, we further introduce an

收稿日期:2024-06-05;修回日期:2024-07-23;责任编辑:覃怀银
*通讯作者:柯逍

optimized universal noise text configuration approach combined with an edge enhancement strategy, which notably enhances the accuracy and generalization capabilities of the baseline model in LFIQA. Additionally, ablation experiments are conducted to assess the contribution of each component to the overall model performance, thereby verifying the effectiveness and robustness of our proposed method. Experimental results demonstrate that our approach not only excels in tests on public datasets like Win5-LID and NBU-LF1.0 but also shows remarkable outcomes in fused datasets. Compared to the state-of-the-art algorithms, our method achieves performance improvements of 2% and 6% respectively on the two databases. The noise verification strategy and configuration method presented in this paper not only provide valuable insights for light field noise prediction tasks but can also be applied as auxiliary tools for other noise prediction types.

Key words: image quality assessment; light field images; visual-textual model; multi-task mode; noise prediction; image enhancement

Foundation Item(s): National Key Research and Development Plan of China (No. 2021YFB3600503); National Natural Science Foundation of China (No. 61972097, No. U21A20472); Major Science and Technology Project of Fujian Province (No.2021HZ022007); Natural Science Foundation of Fujian Province (No.2021J01612)

1 引言

光场图像[1]是一种特殊的图像类型,具有额外的深度和方向信息,能够捕捉场景中每个位置的完整光线信息. 这种独特特性使得光场图像在虚拟现实和增强现实等领域具有广泛的应用前景. 随着人工智能生成内容(Artificial Intelligence Generated Content, AIGC[2])的兴起,通过合成或系统生成的光场图像已经被广泛运用到日常生活中. 相较于传统的3D显示技术,光场图像技术不仅显著突破了眼镜等物理设备的限制实现了裸眼3D显示,为用户带来了更为自然和真实的视觉体验. 更重要的是,光场图像技术还具备捕捉精确深度信息和环境感知的能力,这一特性在机器人、自动驾驶等领域尤为关键,极大地提升了这些系统的决策与导航能力. 在工业领域,光场图像可用于检测和测量复杂物体的三维形态,提高检测精度和效率. 此外,光场技术在军事领域的应用也具有重要意义. 无人机、卫星成像、侦察和监视[3~6]等任务中,高精度的光场图像能够极大地提升战场态势感知能力,为国家安全提供更为坚实的保障. 因此,作为新一代图像处理和显示技术的前沿,光场图像技术的研发和掌握有助于推动科技创新,促进相关产业的升级换代.

评估图像质量的任务被称为图像质量评估[7](Image Quality Assessment, IQA). 关于IQA方法的分类,根据使用的参考图像数量,一般将IQA任务分为三类[8~13]:全参考图像质量评估(Full Reference IQA, FR-IQA)通常需要使用完整的参考图像进行评估;部分参考图像质量评估(Reduce Reference IQA, RR-IQA)和无参考图像质量评估(No Reference IQA, NR-IQA)通常使用部分参考图像或者不使用参考图像. 考虑到生产和日常生活的实际需要,获取参考图像往往涉及不可控的主观因素. 因此,现有的IQA研究更倾向于重视NR-IQA方法. 根据数字图像类型分类时[14~16],通常分为二维图像质量评估(2D-IQA),三维/立体图像质量评估(3D-IQA)以及一些新型图像的质量评估.

与评估其他类型数字图像的现有方法相比,这些方法通常考虑图像的二维或三维视觉特征,因此在应用于二维或三维/立体图像时表现出色. 然而,由于缺乏对光场图像特有属性的分析,直接将这些方法转移到光场图像质量评估领域(Light Field Image Quality Assessment, LFIQA)是具有挑战性的[17~19]. 二维和三维图像的质量评估方法主要依赖于图像的分辨率、对比度、边缘清晰度等视觉特征,而光场图像包含了更多的光线信息,如光线的方向、角度等. 光场图像通过捕捉到场景中每一点光线在不同方向上的传播信息,从而提供更加丰富的视角和深度特征. 这使得光场图像不仅在空间分辨率上有所不同,而且在光线场的分布和方向性上也表现出独特的特征. 因此,传统的二维/三维图像质量评估方法无法充分捕捉和分析这些独特的光线属性,导致在评估光场图像质量时效果不佳.

近期的光场图像质量评估研究已然意识到了这个问题,研究者们在考虑光场图像本质属性上设计、训练光场图像评估模型. 但是这也带来了泛化性下降的问题. 具体说来,这些方法往往是针对特定的光场图像格式和特征进行优化,缺乏对其他格式和特征的适应性. 因此,尽管这些模型在一定范围内表现良好,但在面对不同格式和特性的光场图像时,往往表现出不足和局限.

综上所述,二维和三维图像质量评估方法在迁移到光场图像质量评估时面临诸多挑战,主要原因在于它们无法充分捕捉光场图像的独特属性,且现有的光场图像评估模型缺乏泛化能力和适应性. 为此,本文提出了一种基于文本-视觉模型的全新方法来评估光场图像的质量分数. 通过利用CLIP[20,21]模型的泛化能力和视觉-文本转换能力,并结合增强策略,该方法在单一数据集和合并数据集的实验中表现出色.

2 相关工作

2.1 光场图像概述

近年来,由于光场图像[1]能够捕捉和表示光、物质和空间之间的复杂关系,这一领域引起了极大的关注. 光场[22,23]的概念最早由Adelson提出,他建立了一个描述图像中包含的视觉信息的理论框架. 自那时起,已经开发了许多技术来捕捉和处理光场数据,包括全光学相机、全息成像系统和基于多视图立体的方法[24,25]. 这些方法旨在记录或重建场景的详细4D信息,包括其3D几何形状、颜色和深度. 光场图像采集已被应用于计算机视觉、图形学和摄影等各个领域. 随着光场相机的可用性不断提高,科研人员对光场图像在虚拟现实(Virtual Reaity,VR[26,27])、增强现实(Augmented Reality,AR[27,28])和混合现实(Mixed Reality,MR[29])等应用的兴趣也日益浓厚,这进一步推动了对光场成像的研究.

从本质上来说,在光场研究初期,光场表示函数包含大量参数,使得该函数中参数的采集和处理成为一项艰巨的挑战. 随着光场图像生成设备的发展,采样光通常被视为单色、时不变且辐射稳定的[30,31]. 因此,与完整光度函数相关的复杂性已逐渐简化. 在当代光场理论中,光场模型通常被定义为四维函数,表示为$L(u,s,v,t)$,如式(1)所示. 其中u和v表示角坐标,s和t表示空间坐标. 对于给定的光场场景,可以使用特定的参数精确确定唯一的子孔径图像,每个子孔径图像从特定的视角捕获场景信息.

$$\text{plenoptic} \rightarrow L(u,s,v,t) \tag{1}$$

在光场图像成像的一般模式下,通过改变式(1)中的参数可以生成不同的子孔径图像[32,33](SAI). SAI的集合表现出高度的相似性,促使研究的焦点发生转移. 这些研究越来越强调对双眼竞争行为的探索,旨在捕捉差异区域内不同的SAI并消除冗余. 在此类研究中常用的方法中,Tucker分解脱颖而出. 通过Tucker分解将最初的高维LFI转化为低维图像用于后续处理成为一种成熟的思路[34~36]. 值得注意的是,最近的研究工作引入了一种创新方法,涉及从特定位置(例如中心区域、对角线区域等)提取的子孔径图像. 采用这种方法是为了涵盖LFI中封装的大部分信息.

2.2 光场图像质量评估

评估光场图像的质量是确保其在各种应用中有效使用的关键步骤. 近年来,随着光场图像技术的发展,光场图像质量评估成为了研究的热点. 传统的图像质量度量方法[38~40],如峰值信噪比(Peak Signal-to-Noise Ratio,PSNR)和结构相似性指数度量(Structural Similarity,SSIM),可以应用于光场内的各个视图或颜色通道. 然而,这些度量并不能完全捕捉光场的多个视图和维度之间的复杂关系. 人们针对光场图像的特性和应用需求,提出了各种不同的客观图像质量评估算法,旨在有效地、准确地评估光场图像[41]的质量. 这些算法结合了图像的空间、角度、深度等多维信息,利用先进的数学模型[42]和机器学习技术[43,44],能够全面地分析光场图像的视觉感知质量,为光场图像的采集、传输、处理和显示等各个环节提供了有力的评估工具. 一些全参考LFIQA方法(例如MDFM[45]和MP-PSNR[46])已被开发出来,通过表征原始图像和测试图像之间的差异来量化图像丢失的程度;LF-IQM[41]使用参考图像将其特征与光场图像的质量得分相关联. 然而,全参考和部分参考光场图像质量评估方法(FR/RR-LFIQA)的实际应用受到原始图像获取困难的限制. 因此,无参考光场图像质量评估方法(NR-LFIQA)对于在没有参考图像的情况下评估图像质量是必要的[47].

为了解决无参考图像下光场图像质量的评估问题,研究者们提出了多种NR-LFIQA方法. 这些方法包括BELIF[48]、DeeBLiF[49]和VBLFI[50]等. BELIF[48]通过表示光场图像为高阶张量并使用张量结构变化来评估其质量;DeeBLiF[49]则采用双流卷积神经网络模型,结合角度和空间信息来预测光场图像的质量分数;而VBLFI[50]则通过减少光场图像的冗余信息,并利用其深度线索的可视化来描述其深度和结构信息.

不可否认,上述所提及的NR-LFIQA方法均展现出了卓越的创新能力,它们不仅设计了新颖独到的评估策略,还显著提高了模型在预测光场图像质量序列时的精确度. 然而,这些方法的普遍局限性在于其泛化能力的不足. 具体而言,这些方法的卓越性能往往高度依赖于用于训练的具体光场图像数据集及其衍生数据,限制了它们在处理多样化、非标准或未知来源光场图像时的适用性. 在实际应用中,光场图像的多样性和不可预测性构成了严峻挑战,因为很难确保所有输入图像都严格符合训练数据的范式. 因此,当面临来自不同采集设备、不同场景条件或不同标签体系的光场图像时,现有方法可能会遭遇显著的预测偏差. 此外,当前的光场图像质量评估算法多聚焦于图像本身的视觉特征,忽略了多模态信息融合的重要性,特别是文本模态与图像模态之间的互补性. 在现实世界应用中,图像质量评价往往不仅仅是一个纯粹的视觉任务,它还可能受到用户评论、图像描述等文本信息的深刻影响. 这些文本信息可能蕴含了关于图像内容、拍摄意图、主观感受等多方面的线索,对于全面、准确地评估图像质量至关重要. 因此,为了克服现有方法的局限性并提升NR-LFIQA的泛化能力和综合评估能力,亟须一种新颖的方法论.

3　基于对比性视觉-文本模型的光场图像质量评估

3.1　模型概述

本文所提出网络的系统图如图1所示,该网络的基线模式源自CLIP以及张等人[51]所提的网络,CLIP模型通过图像编码器将图像转换为特征向量,通过文本编码器将文本转换为特征向量.然后,通过对比学习方法,CLIP模型最大化了相关图像和文本对的相似度,同时最小化了不相关图像和文本对的相似度,从而在一个共同的嵌入空间中实现了图像和文本的有效关联.

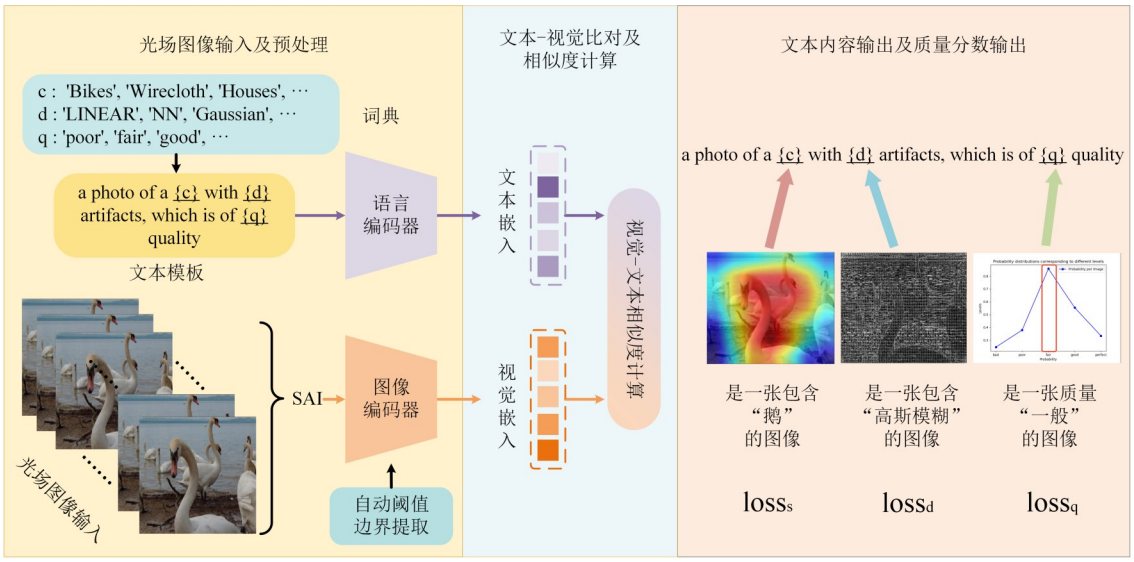


图1　基于CLIP模型的光场图像质量评估方法概念图

我们在基线模型与张等人[51]工作的启发下,针对光场图像的固有特性,对模型进行了本地化和微调,以应用于光场图像质量评估(LFIQA)领域.与先前工作不同的是,我们的研究主要聚焦于多任务设计过程的分析以及光场图像预处理方法对结果的影响分析.基线模型的输入是一组光场图像,输出包括这些光场图像的场景语义文本、噪声文本、光场图像质量文本以及该组光场图像的质量分数.最后,通过映射和联合概率分数对输入光场图像的质量进行估计.具体来说,图像特征向量$\boldsymbol{V}_{\text{vis}}$和文本特征向量$\boldsymbol{V}_{\text{text}}$首先被归一化,然后计算它们之间的点积,得到相似度分数.相似度公式为:

$$\text{Sim}(\boldsymbol{V}_{\text{vis}},\boldsymbol{V}_{\text{text}})=\frac{\boldsymbol{V}_{\text{vis}}\cdot\boldsymbol{V}_{\text{text}}}{\|\boldsymbol{V}_{\text{vis}}\|\cdot\|\boldsymbol{V}_{\text{text}}\|} \tag{2}$$

其中,$\boldsymbol{V}_{\text{vis}}\cdot\boldsymbol{V}_{\text{text}}$表示两个向量的点积,$\|\boldsymbol{V}_{\text{vis}}\|$和$\|\boldsymbol{V}_{\text{text}}\|$分别表示向量$\boldsymbol{V}_{\text{vis}}$和$\boldsymbol{V}_{\text{text}}$的模长.这个相似度分数越高,表示图像特征和文本特征相似度越高.在多模态光场图像质量评估中,这一过程至关重要,因为它直接衡量了图像和文本描述之间的相似性.

3.2　高能区域捕捉算法

在图像质量评估领域中,使用图像预处理手段处理待输入图片是一种常见方法.图像的边缘区域作为蕴含丰富信息与关键结构特征的部分,对于人类视觉感知及众多计算机视觉应用具有重要作用.通过精准识别并提取这些边缘及高能区域,我们能够定位图像中的核心特征与关键区域,这对于识别并量化图像中的噪声、失真等质量问题具有辅助效果.Canny边缘检测方法[52,53]是边缘检测领域中的经典算法,该算法通过多个步骤,包括高斯平滑和梯度计算,有效地提取图像中的边缘信息.其对噪声的抵抗能力和对真实边缘的准确检测,使其成为图像处理中首选的边缘检测方法之一.

为了进一步量化并评估边缘检测的效果,能量图谱[54]作为一种直观且强大的分析工具被广泛应用.作为表征图像能量分布的一种有效手段,能量图谱能够将图像内部的频谱特征与强度变化以视觉化的方式呈现出来,从而便于深入分析图像的结构特性与边缘检测算法的性能表现.通过细致分析能量图谱,研究人员能够直观地掌握边缘检测算法在图像边缘信息提取方面的能力,进而对算法的关键参数进行有针对性的优化与调整.基于上述背景,本节旨在以能量作为核心评估指标,深入探讨并论证Canny算法在作为LFIQA预处理步骤时的有效性.

3.2.1　基于自动阈值的Canny算法

在图像处理领域,边缘检测方法是提取图像中关键信息的关键预处理步骤之一.众多边缘检测算法中,

Canny算法因其独特的能量优化特性而备受青睐. 相较于Sobel算子[55]、尺度不变特征变换[56](SIFT)等方法，Canny算法在能量消耗与检测结果质量之间达到了更优的平衡，从而成为边缘检测领域的首选算法. 为了论证Canny算法的有效性，以Sobel算子(式3)为例，其利用图像中像素点之间的灰度差异来识别图像中的边缘. 梯度幅值G与方向θ的生成方式可以通过水平方向的梯度G_x和垂直方向的梯度G_y进行计算.

$$\begin{cases} G=\sqrt{G_x^2+G_y^2} \\ \theta=\arctan\left(\dfrac{G_y}{G_x}\right) \end{cases} \tag{3}$$

然而，Sobel算子在计算梯度时只考虑了每个像素点周围的局部信息，这种局部梯度计算可能会导致一些噪声或者局部变化引起的边缘被检测；同时，只考虑了局部信息的固有特质可能会导致边缘不够准确. Canny算法与其相比，包含了多个步骤，其中包括高斯滤波、梯度计算、非最大值抑制等. 其中，非最大值抑制阶段是关键的一步，确保只有局部梯度的极大值点才被保留，这样可以有效地抑制非边缘像素点的响应，从而使得边缘更加准确.

图2 不同边缘检测算法效果不同

鉴于光场图像数据集中图像内容和特征的丰富性，手动地为Canny边缘检测算法选择最佳参数被证明是不切实际的. 因此，基于图像中值的自动阈值处理机制变得势在必行. 这种算法策略的基本原理在于其对控制图像梯度分布统计特性的敏锐感知. 通过图像中值将Canny阈值选择相关的复杂性大大简化. 高阈值ξ_h和低阈值ξ_l的推导遵循式(4)，确保系统方法的有效性. 式(4)封装了图像梯度分布特征和参数调整之间复杂的相互作用，从而促进了鲁棒且自适应的阈值机制. 因此，所提出的算法不仅简化了阈值选择过程，还通过利用固有图像属性来提高其效率，其中ε为固定参数，v为单通道像素强度的中位数. 如表1所示，可以直观看出，与比一般的边界阈值算法相比，使用自动边界阈值算法能够产生更为优秀的结果.

$$\begin{cases} \xi_h=\max(0,(1-\varepsilon)\cdot v) \\ \xi_l=\min(0,(1+\varepsilon)\cdot v) \end{cases} \tag{4}$$

深入研究实验结果后发现，当运用相同的模型来处理图像，但采用不同的预处理方案——即宽阈值、窄阈值以及自动阈值，其所得的最终结果呈现出显著的不同. 综合表1以及图3，具体来说，不同的阈值设定直接影响了模型对图像特征的提取和识别能力，从而导致了性能上的波动. 这表明自动阈值的应用在处理光场图像时能够更有效地提取图像的特征信息，从而提高了模型的性能表现. 具体来说，窄阈值处理方法在某些情况下会遗漏重要的边缘细节，而宽阈值方法则可能引入过多的噪声. 相比之下，自动阈值算法能够根据图像内容自适应地调整阈值，从而在保持边缘细节的同时有效地抑制噪声. 这一特性使得自动阈值算法在各种不同场景中表现出色，提高了图像质量评估的可靠性. 因此，在实际应用中，为了获得更好的光场图像质量评估结果，建议优先选择自动阈值方案作为预处理方案.

表1 阈值算法结果

阈值算法	SROCC
宽阈值	0.887 5
窄阈值	0.851 4
自动阈值	**0.908 1**

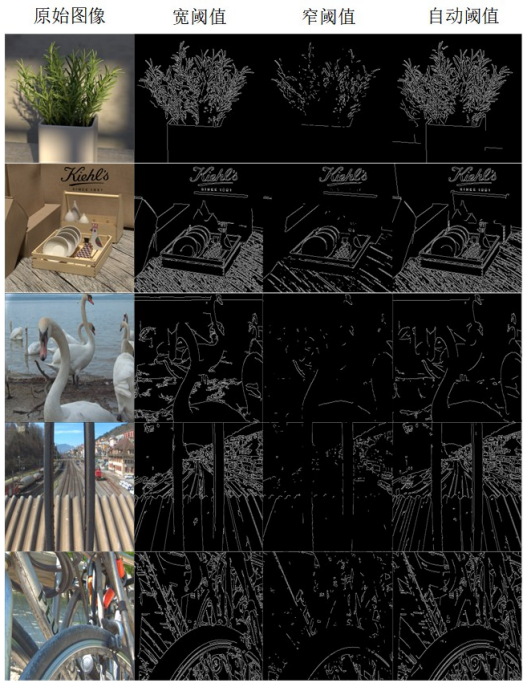


图3 宽、窄、自动阈值作用于同一个图像时的边缘效果

3.2.2 时间-尺度-小波能量图谱分析

在先前章节中分析了自动阈值算法的视觉特性，在此基础上，本节量化了自动阈值算法对光场图像边界的具体影响过程. 能量图谱分析是基于高能量区域捕捉的预处理方法的关键步骤之一. 小波变换作为一种多分辨率分析方法，可以将处理后的图像分解成不同尺度的子带，这些子带包含了图像在不同尺度上的频率信息，从

而形成能量图谱,这些图谱揭示了图像在不同尺度上的特征信息. 基于“时间-尺度-小波”能量图谱的分析方法是一种在信号处理和图像分析领域常用的方法,该方法用于分析信号在时间和尺度两个维度上的频谱特征. 由于小波变换具有等距效应,有以下变换:

$$\int |f(x,t)|^2 \mathrm{d}x = C_\varphi^{-1} \iint \left|\mathrm{WT}_f(a,b)\right|^2 \frac{1}{a^2} \mathrm{d}a\mathrm{d}b \tag{5}$$

其中,$\left|\mathrm{WT}_f(a,b)\right|$表示信号强度,$a$表示时间因子,$b$表示平移因子. 在式(5)的基础上,在时间因子$a$与平移因子$b$上分别进行加权积分,如式(6)所示.

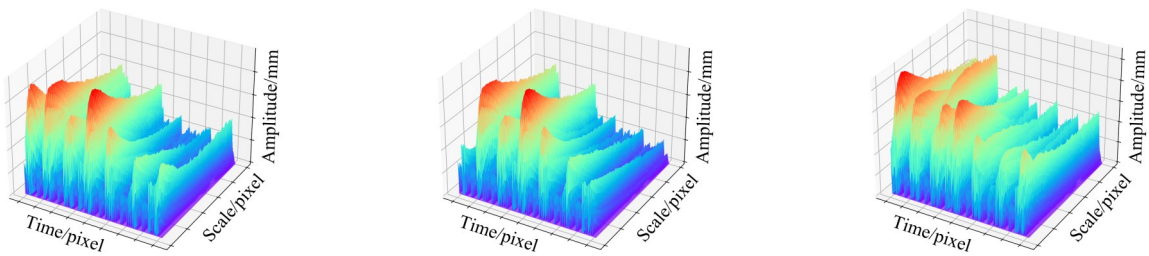
$$\begin{cases} E(a) = \int \left|\mathrm{WT}_f(a,b)\right|^2 \mathrm{d}b \\ E(b) = \int \dfrac{1}{C_\varphi a^2} \left|\mathrm{WT}_f(a,b)\right|^2 \mathrm{d}a \end{cases} \tag{6}$$

其中,$E(a)$小波尺度-能量图谱,反应信号的能量随着尺度变化的情况;$E(b)$表示小波时间-能量图谱,反应信号的小波能量沿时间轴分布的情况.

通过将时间-尺度-小波能量图谱应用于边缘自动阈值算法,我们能够更好地捕捉图像的多尺度特征并将其可视化,从而提高特征提取的准确性和鲁棒性. 图4展示了不同算法视角下的时间-小波-能量图谱,其中横轴表示时间,纵轴表示尺度,表面高度表示能量. 时间-尺度小波能量谱表示了信号在不同时间和尺度上的局部能量分布,能量谱的高度反映了信号在相应时间-尺度位置上的能量强度. 当使用自动阈值算法时,图像边界区域展示出的能量是最大的,这意味着在以CLIP为基线模型时,自动阈值算法下图像的边界区域更容易被模型检测到. 由于自动阈值算法在区分边缘与非边缘方面性能更优,使得边界上的能量分布更为明显,更有利于光场图像质量评估.

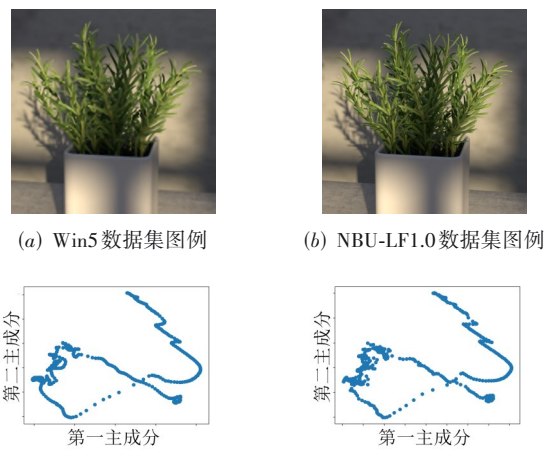
3.3　噪声预测系统评估光场图像质量

当前,光场图像评估模型的训练过程普遍受限于特定格式的光场图像集合或数据集,这种局限性使得训练出的模型仅能高效运作于该特定数据集的验证与测试环节. 一旦遭遇新的场景,或是面对标签体系不同但内容相似的数据时,模型的性能往往会遭受显著影响,展现出迁移性和泛化能力的不足. 以光场图像质量评估领域内广泛采用的Win5-LID和NBU-LF1.0数据集为例(如图5(a)和图5(b)所示),尽管两者在场景覆盖上有所重叠,但却具有不同的标签数值. 这一现象直接限制了模型在跨数据集应用时的发挥.



(a) 窄阈值能量图谱　　(b) 宽阈值能量图谱　　(c) 自动阈值能量图谱

图4　不同阈值算法的时间-小波-能量图谱



(a) Win5数据集图例　　(b) NBU-LF1.0数据集图例

(c) J_{noise}特征PCA表示　　(d) J_{ref}特征PCA表示

图5　不同数据集中包含相同场景的图片及特征

深入剖析这两个数据集,我们发现相同图像在Win5与NBU-LF1.0中的质量评价标签并不一致,这揭示了单一数据集训练的局限性. 参考二维与立体图像质量评估领域的研究成果,我们认识到图像的语义感知是评估图像质量的前期必要步骤,它为获取图像质量分数提供了有价值的线索. 然而,单纯依赖语义感知来确定图像质量分数存在局限性,因为它难以全面解释为何具有相似语义但来自不同数据集的图像会获得不同的质量评分,也无法合理解释为何在一个数据集上表现优异的模型在另一非同源数据集上表现不佳,即便两者在语义层面高度相似.

因此,我们得出结论:仅通过图像语义感知来解决光场图像质量评估问题,在特定数据集内部或许能取得良好效果,但在跨数据集或面对多样化数据时效果

可能下降. 为了提升评估模型的全面性和适应性,我们必须从更多维度出发,完善光场图像的评价机制.

3.3.1 噪声文本类型分析与处理

在实际应用中,光场图像(LFI)不可避免地受到噪声、失真及多种外部干扰因素的侵扰,这些挑战显著提升了图像处理和计算机视觉技术的复杂性与必要性. 鉴于这一背景,从噪声种类与失真类型的精细分析入手,对光场图像质量评估(LFIQA)进行全面考量显得尤为合理且迫切. 尽管利用噪声预测作为辅助任务以增强图像质量评估性能的策略已有所探索,但关于其内在机制与有效性的深入剖析尚显不足.

现有研究往往将“噪声预测对IQA任务具有正面影响”视为一种直观且不言而喻的结论,而缺乏坚实的理论基础与详尽的实验验证来支撑这一观点. 具体而言,缺乏系统性的数据分析与严谨的实验设计,仅凭单一的实验结果来断定噪声预测对LFIQA任务准确性的提升作用,不仅削弱了结论的说服力,也限制了其普遍适用性的验证. 鉴于此,本文致力于通过深入分析特征层面的相互作用与影响,从理论与实践两个维度上证明噪声预测在提升LFIQA任务准确性中的关键作用.

3.3.1.1 特征向量对比法

在噪声的预测分析任务中,首先要考虑对噪声的提取方法,同时要保证该方法的可行性. 借鉴一般性的程序设计思想,针对于任意子孔径图像J_{noise},所提出的噪声确认的方法可以概括为以下三步:

Step1 计算J_{noise}中包含噪声的光场图像中的实际噪声特征F_{noise};

Step2 计算视觉-文本模型预测的J_{noise}中包含噪声的光场图像中的实际噪声特征F'_{noise};

Step3 分析比较特征F_{noise}和特征F'_{noise},考虑特征F_{noise}和特征F'_{noise}的基本情况;

通过分析Step(3)的结果,如果训练完成的模型能够展现与F_{noise}特征类似的表征情况,则可以认为模型在训练时已经习得了分析光场图像特征的能力. 伴随着预测结果的提升,可以认为噪声预测能够提升LFIQA任务的准确性.

3.3.1.2 基于PCA的特征向量对比分析

基于3.3.1.1节的深入分析可以明确,针对特征F_{noise}与F'_{noise}的精准分离与有效提取,是整个处理流程中的关键环节. 具体而言,因为CLIP的多任务架构内嵌有利用噪声特征预测噪声的任务,F'_{noise}的分离与处理显得相对直接,可便捷地通过该任务模块的前置输出直接推导出噪声特征的表征. 然而,从复杂图像中单独剥离出F_{noise}则是一项极具挑战性的任务,当前技术手段难以确保全面而彻底地捕捉图像中的所有噪声成分. 此外,噪声分离技术的多样性进一步加剧了这一难题的复杂性. 不同分离方法的应用可能导致提取出的噪声特征存在差异.

在探讨图像质量评估(IQA)领域时,特别是在无参考图像质量评估(NR-IQA)的背景下,本文提出的方法巧妙地利用了IQA任务的内在特性,即现有数据集中普遍包含的原始参考图像J_{ref}. 这一设计初衷旨在促进全参考(FR-IQA)方法的研究与发展. 值得注意的是,尽管这些原始图像与受噪声干扰的图像通常由具有相同配置参数的相机捕获,确保了除噪声外其他成像条件的一致性,从而便于单独针对噪声因素进行深入分析,但在本研究的无参考框架下,J_{ref}并不直接参与CLIP模型驱动的光场图像质量评估过程. 本文所提出的方法仍然严格遵循NR-IQA的原则,即不依赖任何未受污染的原始图像作为参考,而是仅基于待评估的受噪声影响的光场图像进行质量预测. 这一设计选择不仅体现了研究的创新性,也凸显了其在现实应用场景中的实用性与鲁棒性.

当获取到原始图像J_{ref}与携带噪声的子孔径图像J_{noise}后,即可通过一般性的差异化方法分离原始噪声D,对于J_{noise}的任意位置噪声,其噪声特征向量生成方式符合式(7):

$$D_{x,y}=\left|J_{\text{ref}}(x,y,\partial)-J_{\text{noise}}(x,y,\partial)\right| \tag{7}$$

其中,(x,y)为矩阵中的位置坐标,∂为相机一般参数.

为了方便噪声特征的提取,使用标准化处理流程是一种常见的手段. 原始噪声矩阵数据D可以使用一般标准化流程作如式(8)处理,获得标准化的噪声特征$D_{\text{normalized}}$:

$$\begin{cases}\overline{D}=\dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\text{Array}_i(D)\\ D_{\text{std}}=\sqrt{\dfrac{1}{N}\displaystyle\sum_{i=1}^{N}(D_i-D)^2}\\ D_{\text{normalized}}=\dfrac{D-\overline{D}}{D_{\text{std}}}\end{cases} \tag{8}$$

其中,$\text{Array}_i(\cdot)$表示将第i行的特征进行序列化,N表示像素总数. $D_{\text{normalized}}$可以视为输入子孔径图像噪声F_{noise}的一种特征表达.

主成分分析(Principal Component Analysis,PCA)是一种常用的降维技术,用于发现数据中的主要特征或主要成分. 其主要目标是通过线性变换将原始数据投影到一个新的坐标系中,使得数据在新坐标系下的方差最大化. 对于包含同一语义的光场图像集合,可以从两个光场图像集合中的同一个位置中分离出J_{noise}和J_{ref}(如图5(c)和图5(d)所示,左边为J_{noise},质量低;右边为J_{ref},质量高),可以看到具有高质量的图像特征散点往往更为连续,而低质量的图像特征散点往往更为离散.

根据式(7)和式(8)绘制标准化处理后的噪声矩阵作为噪声的其中一种特征表示,噪声的特征往往是离散且无规律的. 图6表示随机选取部分噪声特征分布. 从J_{noise}中分离可视化的噪声后,为了说明该向量确实能够表示J_{noise}中的噪声,将这些点在J_{noise}和J_{ref}的特征(图6)上重新标注. 为了表述方便,定义一个阈值T,如果$D_{x,y}$大于T,则对应位置的点被标记为橙色,否则不进行标记,绘制的示意如图7所示. 观察可以发现所示的特征在J_{noise}中基本是均匀的,同时随着阈值的增大,被认为存在噪声的像素逐渐减少. 因此可以确认图6中显示的J_{noise}是噪声特征的表示方法之一.

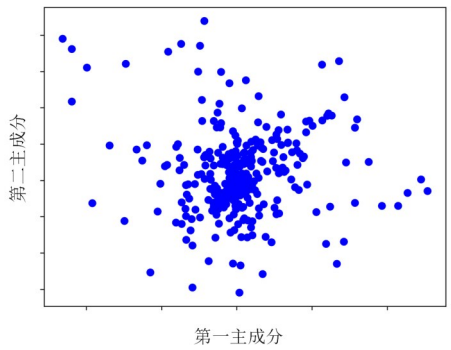


图6　J_{noise}中包含的噪声特征表示

深入分析图6所呈现的噪声特性后,我们观察到从CLIP模型预测结果中提炼出的噪声特征与实际噪声J_{noise}之间存在着显著的高度相似性. 值得注意的是,尽管CLIP模型内部实施的卷积变换导致预测数据的维度与实际噪声不完全匹配,这一差异并未削弱我们将这些输出视为有效预测噪声特征的合理性. 为了更直观地展示这种相似性,我们在图7中采用额外颜色编码来标记这些散点(蓝色表示参考图像,绿色表示携带噪声的图像,橙色表示噪声),并与图8中的噪声图像并置展示. 通过细致对比,我们发现这些散点间展现出了接近80%的高相似度,这一发现强烈支持了CLIP模型在噪声预测方面的准确性,其预测结果紧密贴近光场图像中实际展现的J_{noise}噪声模式.

此结果不仅彰显了CLIP模型在噪声处理领域的巨大潜力,也为探索和优化噪声处理算法图8开辟了新的视角和启示. 基于详尽的对比与分析,我们有充分的理由将CLIP模型的噪声预测功能视为一种高效且可靠的噪声预测工具. 展望未来,这一工具有望在实际应用中为光场图像的质量评估、噪声抑制等关键任务提供坚实的技术支撑和显著的性能提升.

3.3.2　基于噪声预测任务的文本聚类配置

在光场图像质量评估中,文本聚类步骤至关重要. 噪声类型繁多,若不加区分地处理,可能会导致模型内

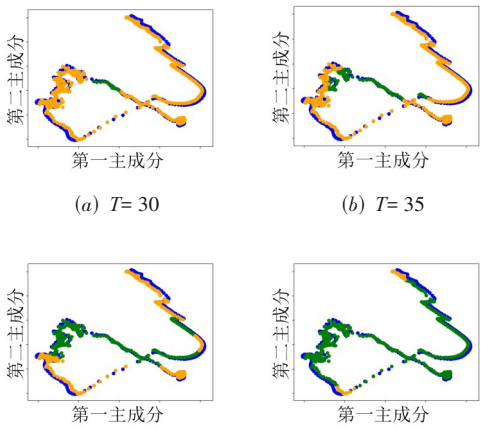


图7　阈值T不同时噪声散点的分布情况

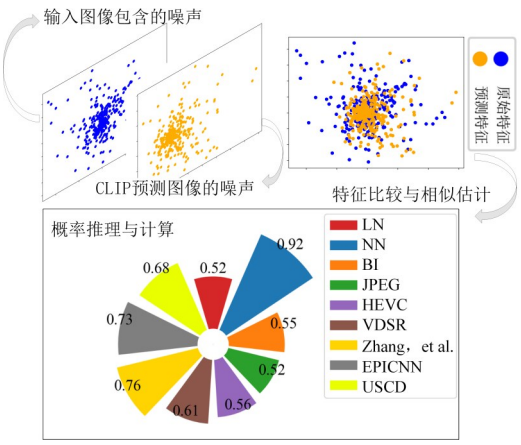


图8　CLIP的噪声预测流程

存需求的指数级膨胀. 因此,必须对噪声进行细致的分类以提高模型的预测精度. 粗糙的噪声分类方法不足以有效地训练模型,细致的颗粒度控制尤为重要. 光场数据集中噪声的分类更加精细,表2展示了光场图像数据库中不同类型的噪声.

为了在聚类性能、模型效率与内存占用之间寻求最佳平衡点,我们采用UMAP可视化技术来直观展示

表2　噪声分类与来源

噪声名称	所属类别	所属数据集
Linear	Reconstruction	Win5
NN	Reconstruction	Win5、NBU
JPEG2000	Compression	Win5
HEVC	Compression	Win5
VDSR	Reconstruction	NBU
Zhang et al.	Reconstruction	NBU
BI	Reconstruction	NBU
EPICNN	Reconstruction	Win5、NBU
USCD	Reconstruction	Win5

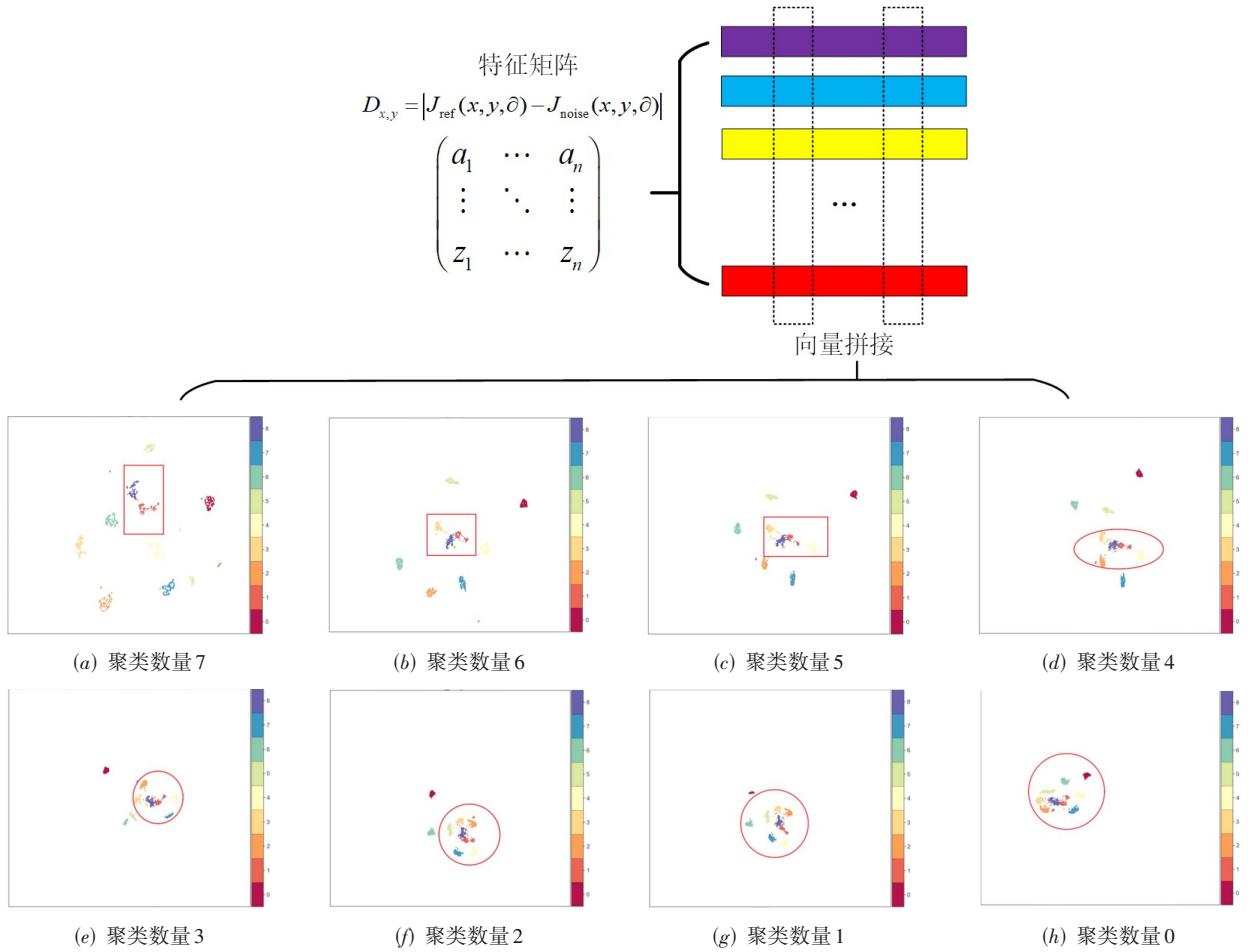
聚类效果. UMAP专为处理高维数据设计,相比其他降维算法,它有更高的处理效率,灵活参数设置允许用户根据数据集的特点进行定制,从而在各种应用场景下展现出卓越的适应性和可视化能力,将数据从高维空间映射至二维或三维空间.

为了量化评估文本聚类过程的成效,我们从数据集中提取了噪声特征F_{noise},并依据标准化流程对这些特征向量进行了细致标记. 为了提升方法的稳健性,我们创新性地采用了随机选取标签两端片段进行编码与连接的策略. 随后,通过设定不同的聚类阈值T,我们系统地评估了聚类效果,并将结果直观呈现于图9中.

经过综合考量与分析,当阈值T设定为120时,模型在聚类效果、模型性能和内存使用方面均达到了理想的平衡(如表3所示). 在此阈值下,我们识别出了三种不同类型的聚合噪声. 在阈值T为120的情境下,存在六种表现出高度相似性的噪声,因此预测它们较为困难;然而,这三种噪声在阈值T小于或等于120时,则显示出显著的区别,使得预测变得相对简单. 在这种情况下,我们使用"Other"来替代这些噪声类型. 根据标签标记,这三种噪声分别为JPEG2000、MLBM(Micro-Lens Based Matching)、VDSR(Very Deep Super-Resolution). 一般的,实际文本数量$\text{Cnt}_{\text{noise}}$应满足式(9):

$$\begin{aligned}\text{Cnt}_{\text{noise}} &= \text{Cnt}_{\text{allnoise}} - \text{Cnt}_{\text{similar}} + 1\\ &= \text{Cnt}_{\text{similar}} + 1\end{aligned} \tag{9}$$

其中,$\text{Cnt}_{\text{allnoise}}$表示数据集中含有噪声的种类个数,$\text{Cnt}_{\text{similar}}$表示聚类数量,$\text{Cnt}_{\text{type}}$表示输入到模型的噪声个数.



(a) 聚类数量7 (b) 聚类数量6 (c) 聚类数量5 (d) 聚类数量4

(e) 聚类数量3 (f) 聚类数量2 (g) 聚类数量1 (h) 聚类数量0

图9 表3UMAP视角下的噪声聚类效果不同聚类数量对应的内存应用情况和模型表现

3.4 损失函数设计

损失函数是机器学习和统计建模中用于评估模型预测误差的一个重要工具. 多任务损失函数在指导模型参数的优化,平衡多个任务权重中起着重要作用. 通过考虑了基线网络的工作与张等人[51]提出的损失函数,同时考虑到服务器性能与模型实际应用的基础上,本章将损失函数按照任务类别分为三类. 对于任意光场图像N,其中包含的子孔径图像记作x,$|N|$表示子孔径图像的数量. 质量任务损失函数、噪声任务损失函数以及场景任务损失函数分别记作$\overline{\text{loss}}_{\text{q}}(x)$,$\overline{\text{loss}}_{\text{d}}(x)$,$\overline{\text{loss}}_{\text{s}}(x)$,这三类损失函数经过线性变化后作为最终的损失函数.

3.4.1 质量任务损失函数

在光场图像质量评估中,对于任意光场图像的损

表3　不同聚类数量对应的内存应用情况和模型表现

τ_0	$\mathrm{Cnt}_{\mathrm{similar}}$	$\mathrm{Cnt}_{\mathrm{type}}$	SROCC	内存/GB
10	6	3	0.730 6	6.4
25	5	4	0.841 6	7.2
50	4	5	0.851 0	8.3
120	3	6	0.907 7	11.2
160	2	7	0.872 2	13.5
200	1	8	0.870 0	15.6

失函数为其对应子孔径图像的损失函数均值. $\overline{\mathrm{loss}}_{\mathrm{q}}(x)$ 的计算方式如式(10)所示. 其中 $\Phi(\cdot)$ 为标准正态累积分布函数,方差固定为1:

$$\overline{\mathrm{loss}}_{\mathrm{q}}(x)=\frac{1}{|N|}\sum_{x\in N}\Phi\left(\left|\hat{q}(x)-q(x)\right|\right) \tag{10}$$

3.4.2　噪声任务损失函数

通常情况下,光场图像与其对应的子孔径图像仅包含一种噪声,所以失真类型的识别任务可以被公式化为一个标准的多类分类问题. 使用多类保真度损失是多类分类问题常见的损失函数. $\overline{\mathrm{loss}}_{\mathrm{d}}(x)$ 的计算方式如式(11)所示. 其中 D 为噪声文本集合,d 为模型预测的噪声文本,如果预测结果中包含噪声 d,则 $p(d|x)=1$,否则为0.

$$\overline{\mathrm{loss}}_{\mathrm{d}}(x)=\frac{1}{|N|}\sum_{x\in N}\left[1-\sum_{d\in D}\sqrt{p(d|x)\hat{p}(d|x)}\right] \tag{11}$$

3.4.3　场景任务损失函数

一张图像常常融合了多个场景,从而引发了标签分配中的多分类挑战. 在图像处理,特别是复杂场景识别与内容深度分析领域,多标签分类显得尤为重要. 这类图像富含多样的场景元素,不仅提升了分析的难度,也对算法的精确度和泛化能力提出了更高要求. 在众多算法策略中,二元保真度算法因其在处理多标签、多类别数据上的卓越表现而备受青睐,成为了一种通用且有效的解决方案. $\overline{\mathrm{loss}}_{\mathrm{s}}(x)$ 的计算方式如式(12)所示:

$$\begin{cases}F(s,x)=1-\sqrt{p(s|x)\hat{p}(s|x)}\\ \quad-\sqrt{(1-p(s|x))(1-\hat{p}(s|x))}\\ \overline{\mathrm{loss}}_{\mathrm{s}}(x)=\dfrac{1}{|N||S|}\displaystyle\sum_{x\in N}\sum_{s\in S}F(s,x)\end{cases} \tag{12}$$

对于子孔径图像 x 的预测中包含场景 s,则 $p(s|x)=1$,否则为0;$\sum_{s\in S}p(s|x)=S$,其中 $1<S<|S|$,物理意义是表征 x 所属的目标类别的数量.

根据上述损失函数 $\overline{\mathrm{loss}}_{\mathrm{q}}(x)$,$\overline{\mathrm{loss}}_{\mathrm{d}}(x)$,$\overline{\mathrm{loss}}_{\mathrm{s}}(x)$ 的结果,在第 t 次训练迭代和第 m 个数据集中,对一个小批量 $\hat{B}_{\mathrm{t}}^{(\mathrm{m})}$ 进行采样,其损失函数如式(13)所示:

$$\mathrm{loss}(\hat{B},t,m)=\frac{1}{|\hat{B}|}\sum_{x\in\hat{B}}(\langle\lambda,\mathrm{loss}\rangle) \tag{13}$$

其中,loss为 $\overline{\mathrm{loss}}_{\mathrm{q}}(x)$,$\overline{\mathrm{loss}}_{\mathrm{d}}(\mathrm{x})$,$\overline{\mathrm{loss}}_{\mathrm{s}}(\mathrm{x})$ 组成的向量,λ 为权重损失,$\langle\cdot\rangle$ 为向量内积.

4　实验结果与分析

4.1　实验设置

在实验中使用的Win5-LID数据集,包含了10张原始光场图像,经历了两种干扰模式:重建伪影和压缩伪影,从而生成了220张失真的光场图像. 两种基于CNN的算法EPICNN和USCD,以及线性插值和最近邻插值被用于重建伪影. 失真类型分为五个级别,除了两个基于CNN的重建伪影算法之外,每个失真图像都被赋予了一个平均意见分数. NBU-LF1.0数据集侧重于角分辨率超分辨率重建,包含了14张光场图像,从而生成了210张相应的失真光场图像. 本次试验中随机抽取70%的样本作为训练集,10%为验证集,20%为测试集. 根据参考图像分割训练/验证/测试集,以确保内容独立性. 为了保证实验的一般性,本文重复此过程十次,并报告中值SROCC和PLCC结果作为预测结果. PLCC衡量预测质量分数与真实分数(MOS或DMOS)之间的线性关系. 它表明算法的预测与地面真实分数的吻合程度,接近1的高PLCC值表示强线性相关性,其计算方式遵循式(14);SROCC是两个变量之间统计依赖性的非参数度量. 与评估线性关系的PLCC不同,SROCC评估两组数据之间的单调关系. 其计算方式遵循式(15). 其中,d_i 表示序列 x 和 y 中每对数据排名之间的差异.

$$\begin{cases}\mathrm{PLCC}(x,y)=\dfrac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2(y_i-\bar{y})^2}}\\ \bar{x}=\dfrac{1}{N}\displaystyle\sum_{i=1}^{N}x_i\\ \bar{y}=\dfrac{1}{N}\displaystyle\sum_{i=1}^{N}y_i\end{cases} \tag{14}$$

$$\mathrm{SROCC}(x,y)=1-\frac{6\sum_{i=1}^{N}d_i^2}{N(N^2-1)} \tag{15}$$

模型方面,本文采用ViT-B/32作为视觉编码器,GPT-2作为文本编码器,基本大小为63 M参数. 通过使用AdamW最小化并采用 10^{-3} 的解耦权重衰减正则化进行模型训练. 初始学习率设置为 5×10^{-6},根据余弦退火规则进行调整. 实验过程中从子孔径图像中随机裁剪大小为符合输入规则的图像,而不改变它们的长宽比. 所有实验均在单个NVIDIA 2080Ti GPU上进行.

字典方面,本文使用了图中的三个预测任务,即场景预测、噪声预测和质量预测.其中表示质量的文本共有5中,分别为:“bad”,“poor”,“fair”,“good”以及“perfect”;本文总共使用了12种场景文本以及7种噪声文本.对于预测文本来说,本文总共有5×12×7=420个候选文本描述.

4.2 与先前方法比较

本文深入研究了CLIP模型的细致训练,采用了精心策划的噪声文本和场景文本内容的混合.通过细致的分析,如表4所示,其中最优结果使用粗体表示.表中结果强调了复杂设计的模型所展现出的卓越性能.值得注意的是,该模型在跨不同数据集的泛化方面表现出了非凡的能力,这是传统评估中以前忽视的一个关键属性.本文首次将混合数据库加入到模型的通用性整合评估中,提供对单体数据库指标以及混合数据库指标的报告,为论证模型的泛用性提供有力的佐证.

研究结果阐明了该模型在不同文本上下文中的卓越适应性,展示了其在复杂的文本环境中导航的能力.即使面对混合数据集,所提出的模型也展示了优秀的性能,强调了其从视觉模仿无缝过渡到文本理解的能力.与此同时,模型的计算效率也是我们重点关注的内容之一,使用推理时间短的模型将会显著提升实际应用中的效率与用户的实际体验,因此,我们将模型的计算时延一起报告在表4中.根据表4中的汇总(其中最优结果使用粗体表示),可以认为我们的模型在性能表现、混合数据集表现、计算时延三者中取得了一个较为优秀的平衡效果,可以认为该模型在更有效的同时更适合实际部署.

需要声明的是,虽然对比性视觉-文本模型在跨数据集和单一数据集中的表现相较于现有方法有显著提升,并且具有较低的计算时延,但在预测准确率和模型泛用性方面仍然存在差距,尤其是与二维/三维图像处理中的深度学习方法相比.这是一个艰巨的挑战,其主要限制在于光场图像的高维特性,使得特征提取过程比处理二维或三维图像更加复杂.

此外,作为一个正在发展中的领域,许多优秀的算法尚未完全开发和应用.尽管如此,需要承认的是,我们提出的方法相较于现有方法已经取得了显著的进展和突破.通过这些改进,我们在光场图像质量评估中实现了更高的性能,展现了该方法的潜力.

表4 与先前的方法比较结果

方法	Win5-LID		NBU-LF1.0		总体(overall)		时延/s
	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	
PSNR[57]	0.602 6	0.618 9	0.530 9	0.664 8	—	—	**0.818 8**
MDFM[58]	0.768 6	0.733 7	0.800 5	0.758 4	0.723 4	0.694 9	0.853 7
MP-PSNR[46]	0.733 5	0.676 6	0.710 9	0.655 0	0.679 6	0.703 4	3.4
BELIF[48]	0.602 1	0.519 5	0.708 4	0.606 9	0.759 2	0.720 0	207.682
VBLFI[50]	0.891 0	0.871 9	0.779 2	0.734 9	0.768 5	0.713 7	65.667
NSS-TD[59]	0.896 9	0.878 3	0.839 6	0.816 5	—	—	219.21
NR-LFQA[60]	0.882 9	0.901 5	0.849 9	0.816 9	0.762 1	0.713 4	183
DeeBLiF[49]	0.842 7	0.818 6	0.837 9	0.820 4	**0.891 6**	0.847 7	—
Xiang′s[34]	0.826 8	0.798 2	0.849 8	0.827 8	0.868 8	0.847 9	128.53
NSTSS[61]	0.725 8	0.672 8	0.773 7	0.737 7	—	—	—
VIDEVAL[62]	0.712 3	0.665 7	0.712 3	0.665 7	0.733 1	0.652 4	27.713
所提方法	**0.911 6**	**0.908 1**	**0.878 7**	**0.861 5**	0.890 3	**0.875 7**	48.367

表5 不同数据集中不同失真类型对视觉-文本模型的影响

方法	Win5-LID				NBU-LF1.0			
	真实场景		合成场景		真实场景		合成场景	
	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
BRISQUE[12]	—	0.591 7	—	0.549 3	—	0.517 6	—	0.506 0
NIQE[63]	0.603 9	0.580 5	0.557 3	0.507 2	0.563 1	0.409 4	0.579 7	0.508 9
MDFM[58]	0.779 6	0.756 0	0.729 0	0.719 7	0.835	0.809 9	0.833 3	0.827 2
Tensor-NLFQ[64]	0.892	0.884	0.925	0.912	0.849	0.842	0.850	0.843

4.3 不同标准划分下评估模型表现

一般来说,可以根据重建方法、场景的性质以及是否涉及合成场景等标准,系统地划分和分类光场图像数据集[65,66].这一过程有助于利用各种划分标准评估

所提出模型或方法对特定噪声或场景的敏感性. 例如,某些重建方法可能对高频噪声特别敏感,而某些场景特征如复杂纹理或快速运动也可能显著影响模型性能. 为了验证视觉-文本模型在不同划分情况下子数据集中的表现,本节在真实场景与合成场景的划分视角下报告 Win5-LID 和 NBU-LF1.0 的感知分数,如表5所示. Win5-LID 数据集主要包含真实场景,而 NBU-LF1.0 数据集则包含更多的合成场景. 通过这种划分,我们就可以评估模型在处理不同类型数据时的鲁棒性和准确性.

总的来说,视觉-文本模型在不同数据库中展示出出色的性能. 在处理包含复杂自然景观的真实场景时,模型能够相对准确捕捉图像细节和语义信息,而在合成场景中,模型也能有效识别合成特征和结构. 在不同光场图像数据集划分视角下,视觉-文本模型表现出良好的鲁棒性和一致性,证明了其在广泛应用场景中的潜力和实用性.

4.4　消融实验

4.4.1　多任务消融

表6详尽呈现了多任务消融实验的结果,其中最优结果使用粗体表示,明确揭示了不同类型任务对光场图像质量预测精准度所施加的差异化影响. 在这些多样化的任务配置中,质量预测任务作为核心组成部分,其必要性不言而喻. 具体而言,当同时融合了质量预测、场景语义理解以及噪声类型预测三项任务时,在 Win5-LID 与 NBU-LF1.0 两个基准数据集上均展现出了最优异的综合性能.

与之相对照的是,在基于质量预测任务之上,仅加入噪声预测或场景语义预测,由于缺少了对图像内容全面理解的互补优势,导致模型在捕捉关键特征或应对复杂场景变化时略显不足,进而使得整体性能出现了一定程度的下滑. 这一现象凸显了多任务学习中各任务间协同作用的重要性,以及全面考量图像特性对于提升光场图像质量预测准确性的关键作用.

表6　多任务消融实验

任务类型			Win5-LID		NBU-LF1.0	
质量	场景	噪声	PLCC	SRCC	PLCC	SRCC
√			0.560	0.534	0.572	0.595
√	√		0.846	0.832	0.837	0.830
√		√	0.810	0.808	0.773	0.766

4.4.2　文本配置消融

在本小节中,我们针对噪声文本与场景文本数量的优化问题,展开了一系列系统而深入的实验研究,旨在探索并确定能够最大化文本识别性能的最佳文本组合比例及相应的图像预处理策略. 鉴于光场图像质量评估领域对参数调整的极端敏感性,任何细微的变动都可能对实验结果产生显著影响,因此,全面且细致的实验设计显得尤为关键.

实验设计中,我们精心策划了多种噪声文本与场景文本数量的组合方案,以全面评估它们对系统识别效能的影响. 通过广泛采集并测试大量样本数据,我们观察到,当场景文本数量设定为12个,而噪声文本数量控制在7个时,系统展现出了最为理想的识别性能. 这一发现不仅确保了系统能够充分学习并利用场景文本中的有效信息,同时有效限制了噪声文本对识别过程可能造成的负面干扰,从而在两者之间达到了一个理想的平衡状态.

表7详细列出了本研究所进行的各项实验及其结果,其中最优结果使用粗体表示,*表示该 Canny 算法为自动阈值算法. 通过量化分析,进一步验证了上述结论的可靠性. 该结果不仅为后续的文本识别研究提供了

表7　噪声/文本组合消融实验

实验编号	学习率调整	场景文本	噪声文本	Canny	Win5-LID		NBU-LF1.0	
					PLCC	SROCC	PLCC	SROCC
1	否	9	6	否	0.751 7	0.616 5	0.774 0	0.614 5
2	是	9	6	否	0.810 0	0.808 1	0.772 5	0.765 6
3	是	9	6	是	0.846 1	0.832 2	0.837 2	0.829 9
4	是	9	7	是	0.871 4	0.881 0	0.852 5	0.838 3
5	是	12	5	是	0.859 6	0.841 6	0.833 5	0.830 1
6	是	12	6	是	0.831 8	0.851 0	0.831 0	0.831 9
7	是	12	7	是	0.893 1	0.907 7	0.861 2	0.845 0
8	是	12	8	是	0.877 1	0.872 2	0.855 5	0.834 0
9	是	12	7	是*	**0.911 6**	**0.908 1**	**0.878 7**	**0.861 5**
10	是	12	8	是*	0.900 7	0.887 1	0.861 6	0.858 0

宝贵的参考依据,也强调了在实际应用中,精准调控文本输入质量对于提升系统整体性能的重要性. 综上所述,本研究通过严谨的实验设计与深入的数据分析,为光场图像中文本识别的优化策略贡献了新的见解.

5 结论

在本文中,我们提出了一种基于对比性视觉-文本模型的光场图像质量评估方法. 通过结合视觉编码器和文本编码器,我们利用CLIP模型的泛化能力和视觉-文本转换能力,设计了一种优化的多模态模型,显著提高了光场图像质量评估的准确性和泛化能力. 具体来说,我们在视觉模态中采用了边缘自动阈值算法,从能量角度证明了其在图像质量评估中的有效性;在文本模态中,我们通过分析数据分布和多任务场景,验证了噪声预测任务的有效性,并提出了一种优化的噪声文本配置方法. 实验结果表明,本文提出的方法在公开数据集Win5-LID和NBU-LF1.0以及融合数据集上均表现优异. 此外,本文探讨了光场图像预处理方法对结果的影响,证明了高能区域捕捉在图像质量评估中的重要性. 我们采用的Canny边缘检测方法,通过能量图谱的可视化,成功地提取了图像中的关键特征,提高了模型的准确性和鲁棒性. 这一方法不仅适用于实时视频监控和智能安防系统中的图像质量评估,还在医疗影像处理、自动驾驶、虚拟现实以及数字内容创作中展现了广泛应用的潜力.

致谢 感谢福州大学计算机与大数据学院以及中国地震局工程力学研究所对本实验提供的硬件和软件支持.

参考文献

[1] WU G, MASIA B, JARABO A, et al. Light field image processing: An overview[J]. IEEE Journal of Selected Topics in Signal Processing, 2017, 11(7): 926-954.

[2] CAO Y, LI S, LIU Y, et al. A comprehensive survey of ai-generated content (AIGC): A history of generative ai from gan to chatgpt[EB/OL]. (2023-03-07)[2024-06-05]. https://arxiv.org/pdf/2303.04226.

[3] 林华. 无人机载太赫兹合成孔径雷达成像分析与仿真[J]. 信息与电子工程, 2010, 8(4): 373-377.
LIN H. Analysis and simulation of UAV terahertz wave synthetic aperture radar imaging[J]. Information and Electronic Engineering, 2010, 8(4): 373-377, 382. (in Chinese)

[4] 刘慧芳, 周骛, 蔡小舒, 等. 基于光场成像的三维粒子追踪测速技术[J]. 光学学报, 2020, 40(1): 0111014.
LIU H F, ZHOU W, CAI X S, et al. Three-dimensional particle tracking velocimetry based on light field imaging [J]. Acta Optica Sinica, 2020, 40(1): 0111014. (in Chinese)

[5] WANG Y, WANG L, LIANG Z, et al. NTIRE 2023 challenge on light field image super-resolution: Dataset, methods and results[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 1320-1335.

[6] WOOD D N, AZUMA D I, ALDINGER K, et al. Surface light fields for 3D photography[M]//Seminal Graphics Papers: Pushing the Boundaries, Volume 2. New York: ACM, 2023: 487-496.

[7] WANG Z, BOVIK A C. Modern Image Quality Assessment[D]. Kentfield: Morgan & Claypool Publishers, 2006.

[8] SHEIKH H R, SABIR M F, BOVIK A C. A statistical evaluation of recent full reference image quality assessment algorithms[J]. IEEE Transactions on Image Processing, 2006, 15(11): 3440-3451.

[9] BOSSE S, MANIRY D, MULLER K R, et al. Deep neural networks for no-reference and full-reference image quality assessment[J]. IEEE Transactions on Image Processing, 2017, 27(1): 206-219.

[10] LARSON E C, CHANDLER D M. Most apparent distortion: full-reference image quality assessment and the role of strategy[J]. Journal of Electronic Imaging, 2010, 19(1): 011006.

[11] TANG Z, ZHENG Y, GU K, et al. Full-reference image quality assessment by combining features in spatial and frequency domains[J]. IEEE Transactions on Broadcasting, 2018, 65(1): 138-151.

[12] MITTAL A, MOORTHY A K, BOVIK A C. No-reference image quality assessment in the spatial domain[J]. IEEE Transactions on Image Processing, 2012, 21(12): 4695-4708.

[13] KANG L, YE P, LI Y, et al. Convolutional neural networks for no-reference image quality assessment[C]// 2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2014: 1733-1740.

[14] WANG Y, YANG J, WANG L, et al. Light field image super-resolution using deformable convolution[J]. IEEE Transactions on Image Processing, 2020, 30: 1057-1071.

[15] CHEN M J, CORMACK L K, BOVIK A C. No-reference quality assessment of natural stereopairs[J]. IEEE Transactions on Image Processing, 2013, 22(9): 3379-3391.

[16] YANG J, WANG Y, LI B, et al. Quality assessment metric of stereo images considering cyclopean integration and visual saliency[J]. Information Sciences, 2016, 373: 251-268.

[17] HUANG H, ZENG H, TIAN Y, et al. Light field image quality assessment: An overview[C]//2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). Piscataway: IEEE, 2020: 348-353.

[18] TIAN Y, ZENG H, HOU J, et al. Light field image quality assessment via the light field coherence[J]. IEEE Transactions on Image Processing, 2020, 29: 7945-7956.

[19] TIAN Y, ZENG H, HOU J, et al. A light field image quality assessment model based on symmetry and depth features[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31(5): 2046-2050.

[20] HAFNER M, KATSANTONI M, KOSTER T, et al. CLIP and complementary methods[J]. Nature Reviews Methods Primers, 2021, 1(1): 1-23.

[21] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//2021 International Conference on Machine Learning. New York: PMLR, 2021: 8748-8763.

[22] CHAI J X, TONG X, CHAN S C, et al. Plenoptic sampling[C]//Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques. New York: ACM, 2000: 307-318.

[23] ADELSON E H, BERGEN J R. The plenoptic function and the elements of early vision[M]//Computational Models of Visual Processing. Cambridge: The MIT Press, 1991.

[24] SAHA N, IFTHEKHAR M S, LE N T, et al. Survey on optical camera communications: challenges and opportunities[J]. Iet Optoelectronics, 2015, 9(5): 172-183.

[25] HUNT J, GOLLUB J, DRISCOLL T, et al. Metamaterial microwave holographic imaging system[J]. Journal of the Optical Society of America A, 2014, 31(10): 2109-2119.

[26] PARK M J, KIM D J, LEE U, et al. A literature overview of virtual reality (VR) in treatment of psychiatric disorders: Recent advances and limitations[J]. Frontiers in Psychiatry, 2019, 10: 505.

[27] ARDINY H, KHANMIRZA E. The role of AR and VR technologies in education developments: opportunities and challenges[C]//2018 6th RSI International Conference on Robotics and Mechatronics. Piscataway: IEEE, 2018: 482-487.

[28] YUEN S C Y, YAOYUNEYONG G, JOHNSON E. Augmented reality: An overview and five directions for AR in education[J]. Journal of Educational Technology Development and Exchange (JETDE), 2011, 4(1): 11.

[29] SPEICHER M, HALL B D, NEBELING M. What is mixed reality?[C]//Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. New York: ACM, 2019: 1-15.

[30] 陈琦, 徐熙平, 姜肇国, 等. 基于光场相机的四维光场图像水印及质量评价[J]. 光学学报, 2018, 38(4): 0411003.
CHEN Q, XU X P, JIANG Z G, et al. Watermarking scheme for four dimensional light field imaging based on light field camera and its evaluation[J]. Acta Optica Sinica, 2018, 38(4): 0411003. (in Chinese)

[31] 赵圆圆, 施圣贤. 融合多尺度特征的光场图像超分辨率方法[J]. 光电工程, 2020, 47(12): 200007.
ZHAO Y Y, SHI S X. Light-field image super-resolution based on multi-scale feature fusion[J]. Opto-Electronic Engineering, 2020, 47(12): 200007. (in Chinese)

[32] KARAMAN M, O´DONNELL M. Subaperture processing for ultrasonic imaging[J]. IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, 1998, 45 (1): 126-135.

[33] CALLOWAY T M, DONOHOE G W. Subaperture autofocus for synthetic aperture radar[J]. IEEE Transactions on Aerospace and Electronic Systems, 1994, 30(2): 617-621.

[34] XIANG J, YU M, JIANG G, et al. Blind light field image quality assessment with tensor color domain and 3D shearlet transform[J]. Signal Processing, 2023, 211: 109083.

[35] CUI Y, JIANG G, YU M, et al. Stitched wide field of view light field image quality assessment: Benchmark database and objective metric[J]. IEEE Transactions on Multimedia, 2024, 26: 5092-5107.

[36] PAN Z, YU M, JIANG G, et al. Combining tensor slice and singular value for blind light field image quality assessment[J]. IEEE Journal of Selected Topics in Signal Processing, 2021, 15(3): 672-687.

[37] DAFFERTSHOFER A, LAMOTH C J C, MEIJER O G, et al. PCA in studying coordination and variability: a tutorial[J]. Clinical Biomechanics, 2004, 19(4): 415-428.

[38] HORE A, ZIOU D. Image quality metrics: PSNR vs. SSIM[C]//2010 20th International Conference on Pattern Recognition. Piscataway: IEEE, 2010: 2366-2369.

[39] SARA U, AKTER M, UDDIN M S. Image quality assessment through FSIM, SSIM, MSE and PSNR—A comparative study[J]. Journal of Computer and Communications, 2019, 7(3): 8-18.

[40] SETIADI D R I M. PSNR vs SSIM: Imperceptibility quality assessment for image steganography[J]. Multime-

dia Tools and Applications, 2021, 80(6): 8423-8444.

[41] PAUDYAL P, BATTISTI F, CARLI M. Reduced reference quality assessment of light field images[J]. IEEE Transactions on Broadcasting, 2019, 65(1): 152-165.

[42] 黄虹, 张建秋. 一个图像质量盲评估的统计测度[J]. 电子学报, 2014, 42(7): 1419-1423.
HUANG H, ZHANG J Q. A statistical measure for blind image quality assessment[J]. Acta Electronica Sinica, 2014, 42(7): 1419-1423. (in Chinese)

[43] 王长淼, 李晖, 张水平, 等. 基于深度学习的光场显微像差校正[J]. 光学学报, 2024, 44(14): 90-99.
WANG C M, LI H, ZHANG S P, et al. Light field microscopic aberration correction based on deep learning[J]. Acta Optica Sinica, 2024, 44(14): 90-99. (in Chinese)

[44] 梁丹, 张海苗, 邱钧. 基于自监督学习的光场空间域超分辨成像[J]. 激光与光电子学进展, 2024, 61(4): 172-184.
LIANG D, ZHANG H M, QIU J. Self-supervised learning for spatial-domain light-field super-resolution imaging [J]. Laser & Optoelectronics Progress, 2024, 61(4): 172-184. (in Chinese)

[45] SHAO S, XING L, XU R, et al. MDFM: Multi-decision fusing model for few-shot learning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32 (8): 5151-5162.

[46] SANDIĆ-STANKOVIĆ D, KUKOLJ D, LE CALLET P. Multi-scale synthesized view assessment based on morphological pyramids[J]. Journal of Electrical Engineering, 2016, 67(1): 3-11.

[47] 叶佳, 张建秋, 胡波. 客观评估彩色图像质量的超复数奇异值分解法[J]. 电子学报, 2007, 35(1): 28-33.
YE J, ZHANG J Q, HU B. Hyper complex singular value decomposition approach to objectively assessing color image quality[J]. Acta Electronica Sinica, 2007, 35(1): 28-33. (in Chinese)

[48] SHI L, ZHAO S, CHEN Z. BELIF: Blind quality evaluator of light field image with tensor structure variation index[C]//2019 IEEE International Conference on Image Processing (ICIP). Piscataway: IEEE, 2019: 3781-3785.

[49] ZHANG Z, TIAN S, ZOU W, et al. Deeblif: Deep blind light field image quality assessment by extracting angular and spatial information[C]//2022 IEEE International Conference on Image Processing. Piscataway: IEEE, 2022: 2266-2270.

[50] XIANG J, YU M, CHEN H, et al. VBLFI: Visualization-based blind light field image quality assessment[C]//2020 IEEE International Conference on Multimedia and Expo. Piscataway: IEEE, 2020: 1-6.

[51] ZHANG W, ZHAI G, WEI Y, et al. 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. Piscataway: IEEE, 2023: 14071-14081.

[52] DING L, GOSHTASBY A. On the Canny edge detector [J]. Pattern Recognition, 2001, 34(3): 721-725.

[53] 李季瑀, 付章杰, 王帆. Canny-Gauss通用域图像隐写算法[J]. 计算机学报, 2024, 47(1): 213-230.
LI J Y, FU Z J, WANG F. Canny-Gauss universal domain image steganography algorithm[J]. Chinese Journal of Computers, 2024, 47(1): 213-230. (in Chinese)

[54] PEREZ-LOMBARD L, ORTIZ J, MAESTRE I R. The map of energy flow in HVAC systems[J]. Applied Energy, 2011, 88(12): 5020-5031.

[55] GAO W, ZHANG X, YANG L, et al. An improved Sobel edge detection[C]//2010 3rd International Conference on Computer Science and Information Technology. Piscataway: IEEE, 2010, 5: 67-71.

[56] WU J, CUI Z, SHENG V S, et al. A comparative study of SIFT and its variants[J]. Measurement Science Review, 2013, 13(3): 122-131.

[57] ATHAR S, WANG Z. Degraded reference image quality assessment[J]. IEEE Transactions on Image Processing, 2023, 32: 822-837.

[58] TIAN Y, ZENG H, XING L, et al. A multi-order derivative feature-based quality assessment model for light field image[J]. Journal of Visual Communication and Image Representation, 2018, 57: 212-217.

[59] MA J, ZHANG X, JIN C, et al. Light field image quality assessment using natural scene statistics and texture degradation[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(3): 1696-1711.

[60] SHI L, ZHOU W, CHEN Z, et al. No-reference light field image quality assessment based on spatial-angular measurement[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 30(11): 4114-4128.

[61] DENDI S V R, CHANNAPPAYYA S S. No-reference video quality assessment using natural spatiotemporal scene statistics[J]. IEEE Transactions on Image Processing, 2020, 29: 5612-5624.

[62] TU Z, WANG Y, BIRKBECK N, et al. UGC-VQA: Benchmarking blind video quality assessment for user generated content[J]. IEEE Transactions on Image Pro-

cessing, 2021, 30: 4449-4464.
[63] ZHANG L, ZHANG L, BOVIK A C. A feature-enriched completely blind image quality evaluator[J]. IEEE Transactions on Image Processing, 2015, 24(8): 2579-2591.
[64] ZHOU W, SHI L, CHEN Z, et al. Tensor oriented no-reference light field image quality assessment[J]. IEEE Transactions on Image Processing, 2020, 29: 4070-4084
[65] 刘玉轩, 张力, 艾海滨, 等. 光场相机三维重建研究进展与展望[J]. 电子学报, 2022, 50(7): 1774-1792.
LIU Y X, ZHANG L, AI H B, et al. Progress and prospect of 3D reconstruction based on light field cameras[J]. Acta Electronica Sinica, 2022, 50(07): 1774-1792. (in Chinese)
[66] 周广福, 文成林, 高敬礼. 基于小波变换与稀疏傅里叶变换相结合的光场重构方法[J]. 电子学报, 2017, 45(4): 782-790.
ZHOU G f, WEN C L, GAO J L. Light field reconstruction based on wavelet transform and sparse Fourier Transform[J]. Acta Electronica Sinica, 2017, 45(4): 782-790. (in Chinese)

作者简介

王汉灵　男, 1999 年 4 月出生于福建省福州市. 现为中国地震局工程力学研究所博士研究生.
E-mail: 2209187058@qq.com

柯　逍　男, 1983 年 10 月出生于福建省福州市. 现为福州大学计算机与大数据学院教授、博士生导师. 主要研究方向为计算机视觉与机器学习.
E-mail: kex@fzu.edu.cn

江澳鑫　男, 1999 年出生于福建省福州市. 现为福州大学计算机技术硕士研究生. 主要研究方向为计算机视觉.
E-mail: jax991220@163.com

郭文忠　男, 1979 年生出生于福建. 现为福州大学副校长, 福州大学计算机科学与技术学科带头人、博士生导师.