低照度图像盲多模态质量评估

Miaohui Wang，会员，*IEEE*，Zhuowei Xu、Mai Xu，*IEEE*高级会员，以及 Weisi Lin，*IEEE*会士 *E,*

摘要——盲图像质量评价（BIQA）旨在自动、准确地预测视觉信号的客观评分，已广泛用于监测低照度应用中的产品和服务质量，涵盖智能手机摄影、视频监控、自动驾驶等领域。该领域近期的发展主要由单模态方案主导，但这与人类的主观评分模式并不一致，因为人类视觉感知会同时反映多种感官信息。本文提出一种独特的低照度图像盲多模态质量评估（BMQA）方法，将主观评估转化为客观评分。为探究多模态机制，我们首先构建了一个包含真实低光照失真的低照度图像多模态质量（MLIQ）数据库，其中包含图像—文本模态对。此外，我们专门设计了BMQA的关键模块，涵盖多模态质量表征、潜在特征对齐与融合，以及自监督和监督学习相结合的混合学习方法。大量实验表明，我们的BMQA在所提出的MLIQ基准数据库上达到了当前最先进的准确度。特别是，我们还构建了一个独立的单图像模态Dark-4K数据库，用于验证其在主流单模态应用中的适用性和泛化性能。Dark-4K上的定性和定量结果表明，只要提供一个用于生成文本描述的预训练模型，BMQA的性能就优于现有的BIQA方法。本文提出的框架和两个数据库，以及收集的BIQA方法和评估指标，均已在此公开。

索引词—低光照图　　像质量评估、视听质量数据库、质量语义描述、多模态学习。

arXiv:2303.10369v2 [cs.CV] 7 Oct 2023

1 引言

图TORAGE像的传输和处理不可避免 [22], ，尤其是在智能手机摄影、视频监控、自动驾驶 等领域。弱光环境下成像可能导致亮度不均、能见度差、色彩受损以及混合噪声增加，从而降低用户体验和产品价值 [47]。此外，低光照图像还会给主流视觉算法的性能带来各种挑战，包括目标检测 [8], 识别 [55],分类 [30],跟踪 [59], 评估 [49], 分割 [9],和增强 [67]。因此，开发可靠的低光照图像客观质量指标至关重要，这有助于满足各类工业产品和计算机视觉任务中的质量测量与检验需求。

低光照盲图像质量评价（BIQA）旨在自动、准确地估计客观评分，从而避免主观实验耗时、不稳定且无法自动化等问题。这对于工业产品的质量监控尤为重要 [47]。与此同时，人类视觉系统（HVS）是 BIQA 任务中视觉信号的最终接收者 [64], ，而人类视觉感知会同时通过多种感官信息体现。然而，现有的 BIQA 方法无论是手工设计的还是基于深度学习的，都很少考虑多模态信息 [34]，且仅适用于低照度图像。因此，如何利用多模态学习更准确地评估低照度图像的质量，是本研究最根本的动机。

图像质量的感知是一个复杂且主观的过程，涉及对视觉刺激的评估与解读。在对视觉信号进行质量评分时，人类能够同时感知多种感知信息 [1], [36]。经过后天训练，我们的大脑能够轻松建立不同模态数据之间的联系，并进一步形成对事物特征的全面表征 [39]。例如，当图像模态受到各种低照度噪声的影响时，其他辅助模态有望提供补充性的质量描述线索，例如对图像内容的文本描述或语义视觉理解 [45]。因此，多模态盲图像质量评估旨在创建一种模拟人类视觉系统的视觉指标，并学习能够表征人类视觉感知的更优质量描述符。

受上述讨论启发，我们提出了一种面向低光照图像的早期多模态盲图像质量评估范式。考虑到目前尚无配备多模态信息的低照度盲图像质量评估数据库，我们构建了首个多模态低照度图像质量（MLIQ）数据库。在图像模态中，低光照图像包含图像采集和处理步骤所产生的真实失真 [47]。在文本模态中，我们制定了质量感知原则，用于生成图像质量的语义描述；这些原则基于这样一个事实：相比给出量化值，人类更擅长描述质量认知 [58]。因此，基于文本的质量语义描述（QSD）可以提供补充信息

• *M. Wang* 和 *Z. Xu* 任职于中国深圳的深圳大学广东省智能信息处理重点实验室。电子邮箱：*wang.miaohui@gmail.com*

• *M. Xu* 任职于北京航空航天大学电子与信息工程学院，地址：中国北京。电子邮箱：*maixu@buaa.edu.cn*。

• *W. Lin* 任职于新加坡南洋理工大学计算机科学与工程学院，地址：新加坡南洋大道 *50* 号。电子邮箱：*wslin@ntu.edu.sg*。

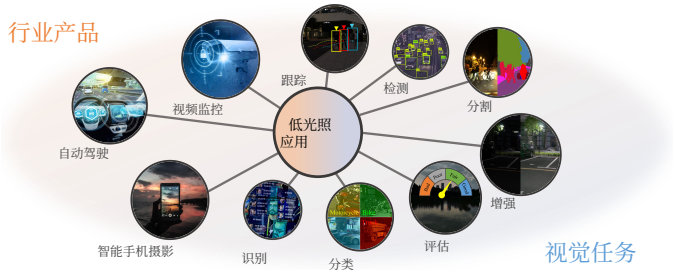


图1. **低光照图像的典型应用**。请放大电子版以查看更清晰的细节。

用于盲图像质量评估的建模。

此外，我们开发了一种独特的盲多模态质量评估（BMQA）方法，用于整合图像和文本特征。跨模态信息的整合有助于在拓展人类视觉感知广度的同时，保持客观视觉信号的表征深度，从而为学习图像质量描述符带来新的益处。数据模态的扩展有助于深度学习模型从不同角度丰富低层嵌入特征，进而提升预测性能的鲁棒性 [1]。大量实验验证了所提出的 BMQA 的有效性，展现了多模态学习在盲质量评估建模中的巨大潜力。

主要贡献有四点：

- 受人类视觉系统（HVS）的启发，我们提出将多模态学习应用于盲图像质量评估（BIQA）问题，通过整合视觉和文本特征来实现这一目标。据我们所知，这是首次尝试明确探索跨不同模态的低光照质量评估。
- 为验证多模态在 BIQA 任务中的可行性，我们首先构建了一个新的低光照图像 MLIQ 数据库，其中包含 3600 组图文数据对。此外，我们还对文本特征进行了统计分析，这有助于展示人类质量认知。
- 基于 MLIQ 数据库，我们进一步研究了多模态学习中的三个主要模块：图像-文本质量表征、潜在特征对齐和融合预测。为提高深度模型训练的效率，我们结合多模态自监督与监督，开发了一种有效的 BMQA 方法。
- 为展示 BMQA 的适用性，我们还建立了一个新的低光照图像质量数据库 Dark-4K，其中仅包含单一图像模态。Dark-4K 用于验证该方法在单模态评估场景下的适用性和泛化性能。实验表明，这种混合学习范式确保 BMQA 在 MLIQ 和 Dark-4K 数据库上均达到最先进性能。

2 相关工作

本节简要回顾近期从多个方面提出的盲图像质量评估方法，包括手工设计的BIQA方法（例如特定失真和通用方法）以及深度学习BIQA方法（例如基于监督学习和无监督学习的方法）。此外，我们还回顾基于多模态的质量评估在用户感知的体验质量（QoE）方面的探索。

2.1 单模态盲图像质量评估

2.1.1 手工设计的*BIQA*方法

手工设计的盲图像质量评估方法 [46]通常基于从专家经验和工程经验中提取特征。由于手工设计的盲图像质量评估方法对部署环境的要求较低（例如数据库规模、硬件平台、计算能力等），因此易于实现和部署。现有的手工设计盲图像质量评估方法可分为 1）特定失真型和 2）通用型。

特定失真型盲图像质量评估方法会同时考虑特定应用中的退化方式和失真类型来评估图像质量，如图 1 所示。例如，已有一些初步的BIQA模型专门针对低照度图像开发，包括亮度引导型 [52], 色彩丰富度启发型 [48],可见度诱导型 [47],以及基于比较学习的 [49]方法。

此外，还有其他有前景的特定失真型盲图像质量评估方法，包括屏幕内容 [46], 高动态范围 [18], 立体全景 [54],审美性 [10],光场 [42], 和水下图像 [68]。由于篇幅所限，本文不讨论其他类型的特定失真型盲图像质量评估方法。

通用BIQA。另一方面，通用方法利用常见的质量感知特征来量化图像失真。例如，基于自然场景统计（NSS）的BIQA方法基于这样的假设：高保真图像遵循某种先验统计特征 [63], ，而这些特征会因频域 [38] 和空间域 [35]中的质量退化而发生改变。人类视觉系统引导的BIQA基于这样的认识：视觉信号的最终接收者是人类，因此利用自由能原理 [62]和视觉敏感度 [28]等感知特征至关重要。

2.1.2 深度学习*BIQA*方法

深度学习BIQA方法 [49] 以端到端的方式直接从失真图像中学习质量特征。与手工设计的BIQA方法不同，这些方法会自动优化质量预测模型，并已展现出良好的性能。深度学习BIQA方法通常可分为：1）基于监督学习的方法；2）基于无监督学习的BIQA方法。

基于监督学习的方法。基于监督学习的盲图像质量评估方法 [61] 着重解决训练样本不足的问题。例如，基于样本的BIQA [2], [19], [20] 主要通过利用图像块级质量特征来预测图像级分数，从而扩大训练样本的规模。基于约束的BIQA [21], [65] 则在监督学习中同时优化多个损失函数，例如采用多个损失进行多尺度监督 [51], 、将新的归一化嵌入引入目标函数 [23], 、使用额外约束调整初始化参数 [69],，以及纳入从其他数据库中学习到的约束 [31], [64], [66]。

基于无监督学习。无监督学习型盲图像质量评估方法无需依赖真实的平均意见分标签即可提取潜在特征。例如，基于度量的盲图像质量评估方法 [32]

采用广泛使用的距离度量（例如，余弦相似度、瓦瑟斯坦距离，等等。）来提取潜在嵌入特征。基于域的盲图像质量评估方法 [29] 通常会设计域对齐约束，并根据在目标域中定义的误差度量，衡量源域中每个样本的质量差异。然而，基于域的盲图像质量评估方法通常需要严格的假设，当测试图像的失真类型未知时，模型要求往往难以满足。

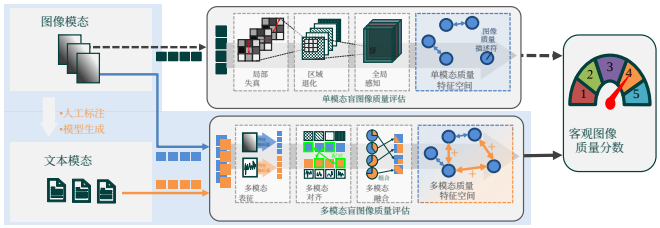


图 2. **单模态与多模态盲图像质量评估方法之间的关系**。人类通过语义描述而非量化值，更善于感知图像质量，这表明文本描述是盲图像质量评估建模中一种非常有用的模态。

2.2 多模态质量评估

多模态质量评估方法仍处于早期阶段，主要关注视频和音频信息 [34]。例如，可以表示并组合视音频特征，以预测体验质量分数。目前的探索主要考虑空间感知 [4], 、时间感知 [60], 和时空感知 [34]的特征组合。

此外，视频和音频失真不会相互影响（即，视频退化不会导致音频退化，反之亦然）。首先分别测量视音频的体验质量分数，然后进一步设计组合规则来预测联合质量。组合规则主要考虑加法 [16],、乘法[50], 和加权闵可夫斯基 [33]。由于忽略了跨媒体信息之间的交互，仅通过组合各自的质量分数很难准确估计联合体验质量。

2.3 动机

如 [34], [36], 所述，目前已有一些基于视音频对建立的多模态质量数据库。然而，包含图文主观标签的数据库却寥寥无几。现有视音频质量指标与我们的图文方法之间存在明显差异：

1）前者旨在预测视频及其对应音频的联合质量，而后者旨在预测图像的视觉质量。我们的 BMQA 主要研究辅助模态对图像质量预测的影响。2）现有的视频-音频质量评估方法很少考虑深度多模态学习，包括多模态特征提取、对齐、融合等。我们的 BMQA 研究各种跨媒体学习模块，涵盖从深度质量表征到质量融合预测。3）在许多质量评估应用中，音频模态通常不可用，而现有的视频-

音频质量评估方法变得不再适用。反之，文本模态可以利用现有语言模型轻松生成 [27], ，而在我们的 BMQA 中，缺少文本并不是问题。

值得注意的是，人类通过语义描述而非量化分数来评估图像质量的效果更好 [58], ，因此文本在盲图像质量评估的建模中可以成为一种非常有价值的模态。此外，本文使用的文本模态可以由许多流行的图像描述生成方法 [53] 以及近期的大语言模型（LLMs） [27]直接生成。

此外，我们还研究了额外模态信息（例如语义视觉理解 [44]）在存在时会如何影响盲图像质量评估（BIQA）的建模。借助我们的MLIQ数据库，我们可以预训练一个QSD诱导的图像描述生成模型，使我们的BMQA方案能够无缝集成到任何现有的单模态盲图像质量评估任务中。单模态盲图像质量评估方法与我们的BMQA之间的关系如图2所示。

最后，我们的BMQA提供了一个全新且前景广阔的研究视角。一方面，多模态数据的同质性表明，训练信息可以互为补充或共享，从而有助于学习描述性更强的质量特征。另一方面，多模态数据的异质性可以拓展训练信息的广度和深度，从而有望提升预测性能。为此，开发一种专用于低光照图像的多模态质量指标十分必要。

3 多模态低照度图像质量（MLIQ）数据库构建

在本节中，我们将介绍如何构建 MLIQ数据库。建立的数据库包含 RGB图像、主观质量分数以及基于QSD的文本，如图 3 所示。

3.1 多模态构建

3.1.1 基于*MOS*的图像模态

自然夜间图像数据库（NNID） [52] 是最新公开的无参考低照度图像数据库。该数据库包含 448 个不同的视觉场景，共计 2240 个样本，涵盖日常生活记录、智能交通、监控、城市灯光秀、航空摄影以及许多其他应用场景。NNID数据库的特征非常符合我们的实验需求。为便于获取用于 MLIQ 的低光照图像数据，我们将其作为基准图像来源之一。

此外，我们采用由两款新设备佳能 *EOS 6D* 和华为 *Mate 30 Pro*拍摄的另外1360张低光照图像，以扩大弱照明场景的覆盖范围。MLIQ共包含3600个低光照图像样本。这些低光照样本是在真实环境（例如室内和室外）中，使用五种不同的移动设备拍摄的。一台设备会以五种不同的设置拍摄同一视觉场景。这五种设置允许



图3. **我们的 MLIQ数据库中的部分图像-文本-MOS对示例**。我们提供了一些具有代表性的示例，用于分析和讨论文本属性，包括亮度、内容、颜色、模糊、噪声和饱和度。请在电子版中放大查看，以获得更清晰的细节。

在不同场景下有所不同。每个低光照样本的分辨率范围为 512×512 至 2048×2048。因此，MLIQ是规模最大的无参考低光照数据库，涵盖多种场景、大量样本、复杂噪声、多样化设备以及真实失真。

为了获得 MLIQ中每张低光照图像的平均意见分标签，我们进行了主观实验。遵循国际电信联盟-R BT.500-14 [3] 和 NNID [52]，，我们采用单刺激方式并构建图形用户界面来进行实验。使用创维 *28U1*，观看距离约为图像高度的三倍。我们邀请了26名参与者，其中女性16名，男性10名（年龄在18至35岁之间）。参与者需要根据11个离散质量分数对图像质量进行评分，分数范围为0至1，步长为0.1。评分最终对应五个质量等级，分别为 $[0, 0.1]$ 表示差， $[0.2, 0.3]$ 表示较差， $[0.4, 0.5]$ 表示一般，$[0.6, 0.7]$ 表示良好，以及 $[0.8, 1.0]$ 表示优秀。

图 4 展示了我们构建的 MLIQ 的统计分析结果。图 4 (a)–(c) 分别给出了拍摄设备、图像分辨率和内容应用场景的整体统计数据。图4 (d) 展示了 MOS 值的直方图分布。如图所示，MOS 值覆盖了视觉质量的整个量化范围，且各个等级的样本数量充足且分布较为均匀。这表明，我们的 MLIQ数据库涵盖了视觉质量的整个范围（从较差到优秀），并且能够很好地区分感知质量。此外，图4 (e) 给出了每幅图像评分的均值和标准差所计算出的95% 置信区间，作为一致性评估指标；置信区间主要分布在0.11 至0.17之间。这表明所有观察者对低光照图像的感知质量均达成了高度一致。因此，所提出的 MLIQ数据库被用作评估客观质量指标性能的真实值。

3.1.2 基于 *QSD* 的文本模态

表达人类视觉感知的一种简单而有效的方法，是请受试者描述并记录他们的语义视觉理解 [44]。文本描述可以提供额外的质量评估线索：一方面，对视觉理解的自由口头描述可能是最丰富的 QSD 来源，因为语言是人类交流中最灵活的方式。另一方面，语言描述有助于避免个人信息偏差，因为语言描述出现不一致的可能性远低于一致的可能性 [17]。

综合此前的一些研究，我们基于人类视觉系统的感知机制，设计了两条文本描述原则。这些原则能够展现从视网膜（即，直觉视觉感知）到初级视觉皮层（即，经验视觉感知）的前馈式视觉信息提取与聚合过程 [57]。这些原则用于指导标注人员生成语言描述。

直觉视觉感知。这一原则受到以往关于人类视觉系统的生理学和心理学实验启发，包括显著性检测和最小可觉差。它与早期视觉密切相关，关注视觉观察所产生的光学刺激与人类视觉系统经验之间的关系 [44]。直觉视觉主要关注一些基本的数据特征，包括整体亮度、颜色状态、纹理、显著物体，等。例如，图3 (c) 中的语言描述包含了一些质量属性，例如亮度信息“明亮”、颜色信息“黑色”和“蓝色”，以及观察到的物体信息“汽车”和“标志”。

经验视觉感知。这一原则受到视觉理论领域现代理论研究的启发。这些研究将经验主义视为构建人类感知的主导范式，例如格雷戈里的理论 [14]。这些研究表明，我们感知到的事物中，一部分来自眼前的物体，另一部分（甚至可能是更大的一部分）始终来自我们的大脑。经验主义原则与晚期视觉密切相关，重点探究观察者的观点如何参与质量认知。经验视觉主要涉及一些常见的现实世界知识以及经验性视觉理解，并着重关注低照度观察可能发生的现实场景 [13]。例如，受试者会使用

‘驾驶’而非‘乘坐’，用于描述‘船’，如图 3 (e) 所示。

然而，文字描述面临着若干挑战。本文将讨论其中三项与获取文本标签有关的挑战：

- **主观性**。由于个体偏好存在固有的主观性，对图像质量感知的语言描述会受到阻碍，因此难以建立一套标准化且普遍适用的图像质量描述词汇。

- **差异性**。语言描述存在显著差异。由于个人经历、文化背景、审美偏好和生理结构各不相同，人们对同一测试图像的感知和解读也各不相同。换言之，视觉吸引力因人而异。

- **表达力**。人类语言表达视觉属性和特质的能力有限。仅用文字描述，往往难以充分呈现图像内容的丰富性和复杂性。例如，要准确表达色调、纹理或光照条件的细微差异，可能颇具挑战。

为应对上述三项挑战，我们尝试开发一种切实可行的语言描述范式。基于质量的句子应包含一组感知属性，例如色彩准确度、噪声、清晰度和整体图像质量。在实验中，每位受试者需要为一张低光照图像提供有意义且简洁的语言描述。对每张图像内容进行听写时，需满足以下要求：

- 对于具有显著物体的图像（例如，图3 (e)中的船），尝试描述图像内容中的所有重要物体。对于具有显著场景的图像（例如，图3 (c)中的停车场景），尝试描述整体环境。对于没有显著内容的图像（例如，图3 (a)中包含许多小物体的建筑物），尝试描述吸引人的内容部分，包括物体和场景。

- 尝试利用相关的光照特征描述整体亮度，例如“明亮”、“亮”、“昏暗”或“暗”。

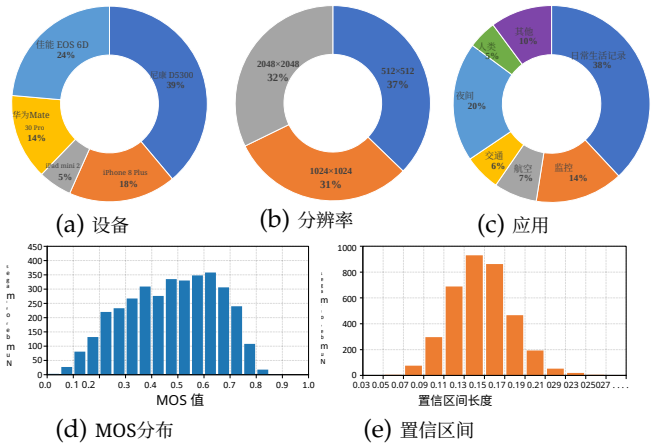
- 尝试描述每个物体的主要属性，例如颜色、亮度、纹理（例如，“木制”杆，见图3 (b)，以及“砖砌”人行道，见图3 (d)），等等。

- 尝试利用相关的色彩特征描述质量认知，例如“色彩丰富”、“鲜艳”、“模糊”、“有噪声”等。

在实验中，14名男性和10名女性受过训练的参参与者需要为每张低光照图像提供有意义且简洁的文字描述。最终，我们收集了3600条文本描述，文本标签的长度为6至45个单词。

3.2 数据库分析

在本节中，我们分析 MLIQ数据库中图像、文本和质量分数（即，MOS）之间的相关性。具体而言，图像模态代表一种视觉刺激，而相应的文本信息则代表对相关图像模态的主观质量认知和理解。基于统计分析



(a) 设备 (b) 分辨率 (c) 应用

(d) MOS分布 (e) 置信区间

图4. **对所提出的多模态MLIQ数据库进行统计分析**：(a) 拍摄设备，(b) 图像分辨率，(c) 应用场景，(d) MOS分布，以及 (e) 置信区间。

我们试图捕捉视觉信号与语言描述之间的内在联系。我们根据亮度、内容和颜色进行统计分析，然后讨论影响低照度图像质量感知的其他因素。

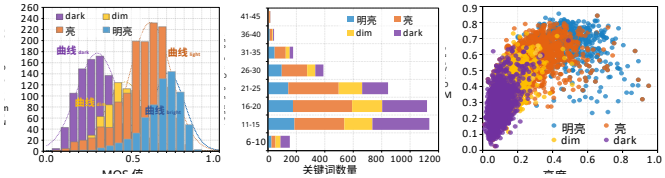
3.2.1 亮度

低光照图像主要存在亮度不足或不均匀的问题。低光照图像的质量水平敏感地取决于视觉亮度感知。文本信息包含描述亮度状态的质量线索（即关键词），能够有效提供补充信息。因此，探究亮度与质量之间的关系具有重要意义。

我们首先从亮度的角度探究基于QSD的文本如何表征图像质量。亮度质量特征包括“暗”“昏暗”“亮”和“明亮”，用于表征照明条件。我们计算了整个数据库中亮度特征的直方图，如图5(a)所示。可以看出，与“暗”“昏暗”“亮”和“明亮”对应的MOS值分别集中在0.2至0.4、0.4至0.5、0.5至0.6和0.6至0.8附近。直观而言，每种亮度特征的直方图都应遵循独立的高斯分布。因此，我们采用高斯函数拟合“暗”“昏暗”“亮”和“明亮”的直方图，其高斯中心分别为0.2797、0.4338、0.5744和0.6745。上述观察表明，亮度属性与图像质量之间存在很强的关联。

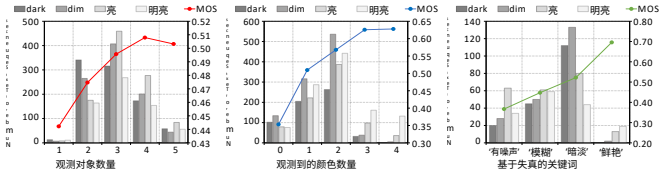
接下来，我们根据亮度分析图像、文本和质量分数之间的关系。图 5 (b). 如图所示，“昏暗”和“暗”表示低亮度，其文字描述长度往往较短。这与我们的经验一致：人们通常难以用较长的语言描述来描述非常暗的场景。

此外，我们计算每幅图像的平均亮度值作为客观亮度级别，并报告相应质量分数的散点图，如图5 (c)所示。我们进一步根据相应文本中包含的亮度关键词，为每个图中的点标注颜色。可以观察到，随着亮度等级的提高，



(a) 高斯拟合 (b) 关键词长度 (c) 主观质量

图5. **基于亮度的文本质量特征统计**：(a) 亮度关键词的直方图及高斯拟合，(b) 亮度关键词长度和数量的直方图，以及 (c) 亮度关键词与 MOS 值的散点分布图。



(a) 内容 (b) 颜色 (c) 关键词

图6. **其他文本质量特征的统计数据**：(a) 内容，(b) 颜色，以及(c)关键词。相关图像的数量以主纵坐标（左侧）表示，而平均意见分结果（用实线连接的点）则以次纵坐标（右侧）表示。

不同散点颜色显示，质量分数总体上也随之提高。这表明亮度特征能够有效表征视觉质量感知。

3.2.2 内容

由于曝光不足，低光照失真通常会导致视觉质量不完整或不清晰，进而带来令人不快的视觉体验。文本模态包含对观察到的物体的语言描述，可有效提供辅助信息，说明视觉注意力集中于哪些物体。因此，我们探究图像质量与内容之间的关系。

观察到的物体减少通常反映为文本模态中的物体描述减少，如图3(b) 所示。基于这一观察，我们统计观察到的物体数量，并报告相应的 MOS 值。图6 (a) 由堆叠柱状图和散点图组成，涵盖观察到的物体数量从 1 到 5 的情况。对于每种观察到的物体数量，堆叠柱状图报告各亮度等级下的图像数量，亮度等级分别为‘暗’、‘昏暗’、‘亮’和‘明亮’。

根据图像内容的统计数据，我们可以得出一些有趣的结论：1）图6(a)中的曲线表明，随着观测对象数量增加，质量分数往往会更高。这可能说明，视觉质量较好的图像通常包含更多可识别的观测对象。2）图 6 (a)中的堆叠直方图表明，亮度等级越低，观测对象数量越少。3）观测对象数量增加所带来的质量分数提升很小（从1到5仅增加0.0657 ），这说明难以通过观测对象数量敏锐地反映视觉体验。一种可能的原因是，低光照失真往往会造成细节丢失，而非显著物体丢失；同时，质量分数更多取决于显著内容本身。

3.2.3 颜色

低光照失真往往表现为色彩对比度低、饱和度低。观察到的颜色可能有效地 pro-
为视觉感知响应提供有用的文本信息。因此，我们研究图像质量与观察到的颜色之间的关系。

观察到的颜色受到损害，表现为颜色描述数量减少，如图所示。3(c)。受此启发，我们统计色彩特征的数量，并报告相应的 MOS 值。图6 (b)由堆叠柱状图和散点图组成，涵盖了颜色词数量从 0 到 4 的情况。对于每种数量的色彩特征，堆叠柱状图分别报告各亮度等级（包括‘暗’、‘昏暗’、‘亮’和‘明亮’）下的图像数量。

根据图像颜色的统计数据，我们还可以得出一些有趣的结论：1）图6(b)中的曲线表明，随着色彩特征数量增加，质量分数趋于升高。这可能表明，视觉感知体验较好的图像通常包含更多可辨识颜色，如图3 (c)所示。2）图6(b)中的堆叠直方图表明，视觉感知亮度越低，观察到的颜色数量往往越少。3）观察到的颜色数量增加所带来的质量分数提升幅度较大（从0增加到4时提升0.2747 ）。这可能表明，观察到的颜色数量能够灵敏地反映低光照失真的质量体验。

3.2.4 其他因素

低光照摄影还常常受到许多其他因素的影响，包括模糊、严重噪声和低饱和度 [7]。如图3(d)所示，若将相机设置为长曝光时间，低光照图像可能会因相机抖动而变得模糊。提高感光度会降低信噪比，同时增加曝光，如图3(e)所示。此外，曝光不足和曝光过度都会显著影响颜色饱和度，进而影响视觉体验。

考虑到文本标签可能包含一些直接描述这些退化特征的关键词，我们收集了基于失真的关键词，并报告了相应的 MOS 值。图6(c)展示了堆叠柱状图和散点图，涵盖了与失真相关的关键词“模糊”、“有噪声”、“暗淡”和“鲜艳”。

根据上述统计数据，我们可以得出一些有趣的结论：1）“模糊”、“有噪声”和“暗淡”等文本特征代表较差的视觉体验，而“鲜艳”代表良好的视觉体验。2）在“明亮”亮度等级下，“有噪声”所占比例较大，这表明低光照图像中的噪声更容易被察觉。3）不同亮度等级下“模糊”所占比例相近，这表明模糊与亮度关系不大。4）在“暗”和“昏暗”亮度等级下，“暗淡”所占比例较大；而在“亮”和“明亮”亮度等级下，“鲜艳”所占比例较大。这些观察结果与人眼偏好高饱和度颜色这一事实相符，如图 3 （f）所示。

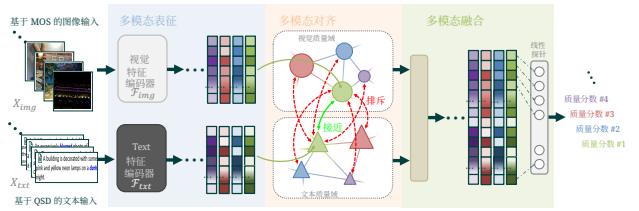


图 7. **所提出的 BMQA 整体学习框架**。该框架由三个关键模块组成，包括多模态特征表征、潜在特征对齐和融合预测。

4 所提出的盲多模态质量评估（BMQA）框架

基于所提出的MLIQ数据库，我们设计了一种独特的深度学习BMQA，如图 7所示。我们针对BIQA任务中的多模态学习主要挑战展开研究，包括特征表示、对齐和融合。最后，我们将介绍我们的学习机制。

4.1 多模态质量表征

本节所说的多模态质量表征，是指提取并整合能够利用补充性质量描述线索的有效特征。由于异构数据具有显著不同的特性，图像和文本空间之间存在较大差异 [39]。因此，我们为二者设计了两种不同的特征质量表征。

4.1.1 文本质量特征表征

在本小节中，我们从文本特征的角度探讨质量认知。文本质量表征 $\mathbf{F}_{txt}$可定义为

$$\mathbf{F}_{txt} = \mathcal{F}_{txt}\left(X_{txt};\, \boldsymbol{\theta}_{txt}\right), \tag{1}$$

其中， $\mathcal{F}_{txt}$ 表示文本特征提取器， X_{txt} 表示输入的质量描述句子， $\boldsymbol{\theta}_{txt}$ 表示相应的权重。 X_{txt}=$<x_1, \cdots, x_N>$ 表示长度为 N 的口语词语。

接下来，我们从 X_{txt} 中提取潜在文本特征。给定一个词语 $x_k \in < x_1, \cdots, x_N >$，我们采用 $\mathcal{F}_{text}$ 为每个词语提取语义质量特征。在 BMQA 中， $\mathcal{F}_{txt}$ 可表示如下:

$$\begin{aligned}\mathcal{F}_{txt}\left(x_k\right) =& \sigma\left[\mathcal{F}_{lp}^{1}\left(x_n, < x_1, \cdots, x_n, 0, \cdots >;\, \boldsymbol{\theta}_{lp}^{1}\right)\right] \\ & \oplus \mathcal{F}_{lp}^{2}\left(x_n, < x_1, \cdots, x_n, 0, \cdots >;\, \boldsymbol{\theta}_{lp}^{2}\right),\end{aligned} \tag{2}$$

其中， $\mathcal{F}_{lp}^{1}$ 和 $\mathcal{F}_{lp}^{2}$ 表示两个线性投影（例如，全连接或注意力投影）， $\boldsymbol{\theta}_{lp}^{1}$ 和 $\boldsymbol{\theta}_{lp}^{2}$ 表示相应的权重， σ 表示非线性激活函数（例如，修正线性单元或高斯误差线性单元）， $\oplus$ 表示融合操作（例如，加法或拼接）。目前，对于高效语义处理 [40] 仍未达成共识。主要原因是，不同方法构建了不同的词语关联（即，对 $\mathcal{F}_{lp}^{1}$、 $\mathcal{F}_{lp}^{2}$、 σ 和 $\oplus$ 采用不同的设置），并且它们在自然语言处理（NLP）方面各有不同优势。

最后，听写中所有词语之间的关联将共同构成$\mathbf{F}_{txt}$的特征表示。在 BMQA 中，它由所有词语特征的线性组合生成:

$$\mathbf{F}_{txt} = \mathcal{F}_{lc}\left[\mathcal{F}_{txt}\left(x_1\right), \cdots, \mathcal{F}_{txt}\left(x_N\right);\, \boldsymbol{\theta}_{lc}\right], \tag{3}$$

其中， $\mathcal{F}_{lc}$ 表示一个全连接层， $\boldsymbol{\theta}_{lc}$ 表示 $\mathcal{F}_{lc}$的权重。在实验中，我们选取 3 个具有代表性的网络作为 $\mathcal{F}_{txt}$的骨干网络，包括词袋（记作 BoW）、循环神经网络（记作RNN）和 Transformer（记作TransF）。

4.1.2 图像质量特征表征

近期研究已证明，深度神经网络（DNN）在许多视觉任务中具有可行性和优越性。受此启发，我们通过深度学习映射获得图像质量表征$\mathbf{F}_{img}$：

$$\mathbf{F}_{img} = \mathcal{F}_{img}\left(X_{img};\, \boldsymbol{\theta}_{img}\right), \tag{4}$$

其中， $\mathcal{F}_{img}$ 表示图像特征编码器， X_{img} 表示输入图像， $\boldsymbol{\theta}_{img}$ 表示相应的权重。

现有的深度学习图像质量表征 [12]主要依赖于空间域中特征的提取与整合。 $\mathcal{F}_{img}$ 通常设计为由多个DNN模块构成的层级堆叠。具体而言，令 $\mathcal{F}_{img}^{l}$ 表示 $\mathcal{F}_{img}$中第 l层的一个基本DNN模块。在BMQA中， $\mathcal{F}_{img}^{l}$ 通常由依次执行的线性计算和非线性激活函数组成，可表示为:

$$\mathcal{F}_{img}^{l}\left(X_{img}^{l}\right) = \sigma_l\left[\mathcal{F}_{lp}^{1}\left(X_{img}^{l};\, \boldsymbol{\theta}_{lp}^{1}\right)\right] \oplus \mathcal{F}_{lp}^{2}\left(X_{img}^{l};\, \boldsymbol{\theta}_{lp}^{2}\right), \tag{5}$$

其中， $\mathcal{F}_{lp}^{1}$ 和 $\mathcal{F}_{lp}^{2}$ 表示两个线性投影（例如，全连接、注意力投影或卷积）， $\boldsymbol{\theta}_{lp}^{1}$ 和 $\boldsymbol{\theta}_{lp}^{2}$表示相应的权重， σ 表示非线性激活函数（例如，修正线性单元或高斯误差线性单元）， $\oplus$ 表示融合操作（例如，加法或拼接）。

近期研究表明，深度模型采用不同的骨干网络会产生不同的学习能力。这说明网络架构可能会对**特征**$_{img}$的提取产生重要影响。因此，我们选择了 5 种具有代表性的网络作为 $\mathcal{F}_{img}$ 的骨干网络，包括 VGG、ResNet（记为 RN）、EfficientNet（记为EN）、Vision-in-Transformer（记为 ViT）和 ConvNeXT（记为CNXT）。

4.2 图像与文本质量特征对齐

在 BMQA 中，多模态质量对齐是指寻找两种模态之间对应的质量表征关系。图像和文本均针对同一图像生成，因此质量描述可以高度一致 [5]。例如，文本关键词可以直接指示视觉注意力所关注的图像区域 [25]，，从而提升图像特征编码器的学习性能。

现有方法通过设计不同模态之间的约束（称为跨模态约束 [1]）来构建多模态对齐。在构建训练目标时，需要定义一种特殊的度量指标来衡量两种模态之间的差异。如果两种模态来自同一传感器的不同视角，则度量指标定义为绝对值误差，例如平均绝对误差（MAE）和均方误差（MSE）。如果两种模态来自不同的传感器，则度量指标定义为相对值误差，例如余弦相似度 [37]。

图像和文本是异构模态。因此，我们采用余弦相似度来衡量相对差异，并设计注意力池化来实现多模态质量对齐。具体而言，我们通过学习注意力分布来提取共享的质量信息，并将$\widehat{}$对齐后的视觉特征$\mathbf{F}^{img}$表示为：

$$\begin{cases} \widehat{\mathbf{F}}_{img} = \mathcal{N}_{ln}\left(\mathcal{F}_{\text{fw}}\left(\ddot{\mathbf{F}}_{img};\ \boldsymbol{\theta}_{\text{fw}}\right) + \ddot{\mathbf{F}}_{img}\right); \\ \ddot{\mathbf{F}}_{img} = \mathcal{N}_{ln}\left(\mathcal{F}_{\text{ms}}\left(\mathbf{F}_{img};\ \boldsymbol{\theta}_{\text{k}}, \boldsymbol{\theta}_{\text{q}}, \boldsymbol{\theta}_{\text{v}}, \sigma\right) + \mathbf{F}_{img}\right) \end{cases}, \tag{6}$$

其中，$\mathcal{F}_{\text{fw}}$表示参数权重为$\boldsymbol{\theta}_{\text{fw}}$的前馈层，$\mathcal{N}_{ln}$表示层归一化。$\mathcal{F}_{ms}$表示一个32头自注意力池化模块。$\boldsymbol{\theta}_{\text{k}}$、$\boldsymbol{\theta}_{\text{q}}$和$\boldsymbol{\theta}_{\text{v}}$分别表示$\mathcal{F}_{ms}$中键、查询和值嵌入的参数权重，我们将投影维度设为2048。$\sigma(\cdot) = \texttt{Softmax}(\cdot/\sqrt{d})$表示用于softmax的缩放$\mathcal{F}_{\text{ms}}$。公式（6）表明，与共享质量信息相关的特征元素将获得比其他元素更高的注意力值。因此，对齐后的质量信息将是所有特征元素的加权组合。

$\widehat{}$对齐后的文本特征$\mathbf{F}_{txt}$可表示为：

$$\begin{cases} \widehat{\mathbf{F}}_{txt} = \mathcal{N}_{ln}\left(\mathcal{F}_{\text{fw}}\left(\ddot{\mathbf{F}}_{txt};\ \boldsymbol{\theta}'_{\text{fw}}\right) + \ddot{\mathbf{F}}_{txt}\right); \\ \ddot{\mathbf{F}}_{txt} = \mathcal{N}_{ln}\left(\mathcal{F}_{\text{ms}}\left(\mathbf{F}_{txt};\ \boldsymbol{\theta}'_{\text{k}}, \boldsymbol{\theta}'_{\text{q}}, \boldsymbol{\theta}'_{\text{v}}, \sigma\right) + \mathbf{F}_{txt}\right) \end{cases}, \tag{7}$$

其中$\boldsymbol{\theta}'_{\text{k}}$、$\boldsymbol{\theta}'_{\text{q}}$和$\boldsymbol{\theta}'_{\text{v}}$表示嵌入参数权重。$\boldsymbol{\theta}'_{\text{fw}}$表示文本特征对齐的前馈参数权重。

最后，通过余弦相似度衡量两个对齐后的1024-$\widehat{}$维宽$\widehat{\mathbf{F}}^{img}$与$\widehat{\mathbf{F}}_{txt}$之间的相对差异，可表示为

$$\mathcal{D}_{cos}\left(\widehat{\mathbf{F}}_{img}, \widehat{\mathbf{F}}_{txt}\right) = \frac{\widehat{\mathbf{F}}_{img} \odot \widehat{\mathbf{F}}_{txt}}{\left\|\widehat{\mathbf{F}}_{img}\right\| \left\|\widehat{\mathbf{F}}_{txt}\right\|}, \tag{8}$$

其中$\odot$表示内积运算，$\|\cdot\|$表示欧几里得距离。$\mathcal{D}_{cos}\left(\widehat{\mathbf{F}}^{img}、\widehat{\mathbf{F}}_{txt}\right)$将作为第 4.4.1 节中多模态自监督学习的度量指标。

4.3 图像与文本质量特征融合

在 BMQA 中，使用两种模态的主要优势在于，可以从不同视角描述图像质量 [11]。例如，质量分数可以直接通过主观测试获得，而文本信息能够反映这些分数。另一个典型情况是，高度相似场景的语言描述通常不可避免地相似，而图像中的空间细节将有助于进行更细粒度的评分决策。这两种情况表明，融合图像和文本质量表征有助于预测更准确的质量分数。因此，多模态质量融合 [1] 被用于整合两种模态，以在我们的 BMQA 中预测质量分数。

为尽可能保留质量信息，我们通过拼接操作整合两种异质模态特征。接下来，我们采用一个宽度为2048的线性探测器[37]，$\mathcal{F}_{fuse}$进行融合，并预测最终的质量分数s_{pred}，其公式如下：

$$s_{pred} = \mathcal{F}_{fuse}\left[\texttt{concat}\left(\widehat{\mathbf{F}}_{img}, \widehat{\mathbf{F}}_{txt}\right);\ \boldsymbol{\theta}_{fuse}\right], \tag{9}$$

其中 concat$(\cdot)$ 表示拼接操作，$\boldsymbol{\theta}_{fuse}$表示 $\mathcal{F}_{fuse}$的参数权重。

算法 1 针对图文情形提出的 BMQA 方案，BMQA^{image}-$_{text}$。

输入：图像样本 X_{img} 和质量描述文本 X_{txt}。输出：质量分数 s_{pred}。1：提取公式 (2)-(3) 中的文本特征表示 F_{txt}；2：提取公式 (4)-(5) 中的图像特征 F^{img}；$\frown$3：分别获取公式 (6) 中对齐后的图像特征 F^{img}，以及公式 (7) 中 F_{txt} 里的$\widehat{}$对齐后的文本特征；$\frown$ $\frown$4：融合 F^{img} 和 F_{txt}，并根据公式 (9) 预测客观质量分数s_{pred}。

4.4 深度多模态学习机制

BMQA 利用跨模态特征提供的补充性质量描述，学习预测图像质量分数。本节将介绍我们的多模态学习机制，包括多模态自监督和监督。

4.4.1 多模态质量自监督

现有方法表明，无需主观分数也能潜在地学习图像质量状态。换言之，可以从失真图像样本中捕捉并学习质量状态。多模态信息有望增强深度模型学习，因为两种模态提供的补充信息有助于质量表征学习。例如，与「明亮」图像相比，人们更喜欢「暗」图像。

来自同一实例的两种模态所共享的质量信息可提供补充信息。因此，我们学习一个嵌入空间，使来自同一实例的特征对彼此接近，而来自不同实例的特征对彼此远离。具体而言，$\hat{}$给定当前训练批次 B中的第 i 个 $\mathbf{F}^{img(i)}$，计算成对概率 $\mathcal{P}(\cdot)$，以找到匹配的第 j 个文本特征：

$$\mathcal{P}_{img}(i,j) = \frac{\exp\left(\mathcal{D}_{cos}\left(\widehat{\mathbf{F}}_{img(i)}, \widehat{\mathbf{F}}_{txt(j)}\right)/\tau\right)}{\sum_{k\in B}\exp\left(\mathcal{D}_{cos}\left(\widehat{\mathbf{F}}_{img(i)}, \widehat{\mathbf{F}}_{txt(k)}\right)/\tau\right)}, \tag{10}$$

其中，τ 表示温度参数，用于控制分布的集中程度。

同时，为第 j 个文本特征$\mathbf{F}^{txt(j)}$找到匹配的$\hat{}$第 i 个图像特征的成对概率为：

$$\mathcal{P}_{txt}(j,i) = \frac{\exp\left(\mathcal{D}_{cos}\left(\widehat{\mathbf{F}}_{img(i)}, \widehat{\mathbf{F}}_{txt(j)}\right)/\tau\right)}{\sum_{k\in B}\exp\left(\mathcal{D}_{cos}\left(\widehat{\mathbf{F}}_{img(k)}, \widehat{\mathbf{F}}_{txt(j)}\right)/\tau\right)}. \tag{11}$$

因此，多模态质量对齐需要满足：每一对$\widehat{}$ $\widehat{}$$\mathbf{F}^{img}$和$\widehat{\mathbf{F}}_{txt}$之间的成对概率达到最大值，即：$\mathcal{P}_{img}(i,i) = \max_{k\in B}\mathcal{P}_{img}(i,k)$和$\mathcal{P}_{txt}(j,j) = \max_{k\in B}\mathcal{P}_{txt}(j,k)$。整体学习目标是最大化联合概率，这等价于最小化负对数似然：

$$\min_{\substack{\boldsymbol{\theta}_{\text{img}},\ \boldsymbol{\theta}_{\text{fw}},\boldsymbol{\theta}_{\text{k}},\boldsymbol{\theta}_{\text{q}},\boldsymbol{\theta}_{\text{v}};\\ \boldsymbol{\theta}_{\text{txt}},\ \boldsymbol{\theta}'_{\text{fw}},\boldsymbol{\theta}'_{\text{k}},\boldsymbol{\theta}'_{\text{q}},\boldsymbol{\theta}'_{\text{v}}}} -\log\left[\sum_{i\in B}\mathcal{P}_{img}(i,i) + \sum_{j\in B}\mathcal{P}_{txt}(j,j)\right]. \tag{12}$$

4.4.2 多模态质量监督

为了获得更好的融合模型 $\mathcal{F}_{fuse}$，在公式（9）中，我们采用均方误差 $\|\cdot\|_2^2$ 来衡量预测分数 s_{pred} 与真实主观分数 s_{gt} 之间的差异。我们的目标是预测出尽可能接近主观平均意见分的分数，优化问题可表述为：

$$\min_{\boldsymbol{\theta}_{fuse}} \left\| \mathcal{F}_{fuse}\left[\texttt{concat}\left(\widehat{\mathbf{F}}_{img}, \widehat{\mathbf{F}}_{txt}\right); \boldsymbol{\theta}_{fuse}\right] - s_{gt} \right\|_2^2. \quad (13)$$

4.5 总体算法

给定一对图像数据 X_{img} 和文本数据 X_{txt}，我们根据式$\mathbf{F}_{img}$ （4）得到图像特征，并根据式（$\mathbf{F}_{txt}$ ）3得到文本特征。之后，BMQA 根据式（ ^ ^6）**和式**（img 7）**分别学习输出对齐特征**F_{txt} 和F。质量特征对齐的学习通过多模态自监督实现。最后，BMQA 根据式（9）学习融合对齐特征，其中质量特征融合的学习通过多模态监督实现。

我们的 BMQA 详细算法总结于**算法** 1。

5 验证与讨论

本节在两个最新的基准低照度图像数据库上开展了大量实验。具体而言，我们分别在图文数据库 MLIQ（即，BMQA$^{image\text{-}text}$）和仅含图像的数据库 Dark-4K（即，BMQA^{image_only}）上验证 BMQA 的有效性。此外，我们展示了与 25 种竞争方法的比较结果，其中包括 8 种手工设计的BIQA方法和 17 种深度学习BIQA方法。下文将详细介绍实验验证、分析和讨论。

5.1 实验方案

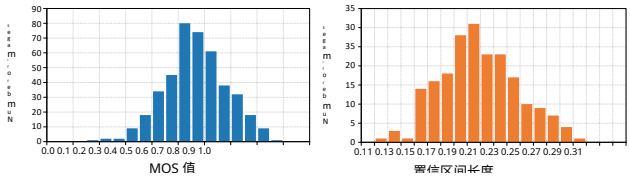
5.1.1 基准数据库

实验中使用了三个数据库：预训练数据库 MS-COCO、新的多模态图像质量数据库 MLIQ，以及仅包含单一图像模态的新型低光照数据库 Dark-4K。

MS-COCO。我们在 MS-COCO 图像描述数据库上对 BMQA 模型进行预训练 [43]，，该数据库包含82783 个带有图像描述标注的图像实例。每条图像描述至少包含 8 个词。MS-COCO 的划分严格遵循官方指定的设置。

MLIQ。我们使用提出的 MLIQ数据集训练 BMQA 模型，该数据集包含3600 对低光照图像及其对应的文本描述。详情见第3节。实验中，按照拍摄场景将 MLIQ 随机划分为**8:1:1**。具体而言，训练集包含 576 个场景和 2880 个样本，验证集包含72个场景和360个样本，测试集也包含72个场景和360个样本。

Dark-4K。为验证跨数据集性能，我们进一步建立了一个新的超高清（UHD）低光照数据库，用于跨数据集验证，命名为 Dark-4K。



(a) MOS分布 (b) 置信区间

图8. **所提出的单模态 Dark-4K 数据库的统计分析**：(a) MOS值直方图；(b) 置信区间分布。

Dark-4K 的原始图像采集自 [6]。Dark-4K 由424 张原始低光照图像组成，这些图像由两款消费级电子产品拍摄：*Sony* α*7S-II* 和*Fujifilm X-T2*。这两款相机采用了不同的成像硬件：*Sony*配备全画幅拜耳传感器，*Fujifilm* 则配备 APS-C X-Trans传感器。Dark-4K 支持对由不同滤色阵列生成的低光照图像进行质量评估。

Dark-4K 的主观实验采用与第3.1.1节所述相同的设置。标注 MOS 结果的直方图以及主观评分的95% 置信区间如图 8 (a) 和图8(b) 所示。

5.1.2 评价度量指标

在视觉质量评估中，皮尔逊线性相关系数（PLCC）、斯皮尔曼秩相关系数（SRCC）和均方根预测误差（RMSE）是三种常用的评估指标 [64]。PLCC用于衡量客观预测与主观分数之间的线性关系，SRCC反映预测结果的单调性，而RMSE衡量预测的准确性。对于效果良好的方法，PLCC和SRCC接近1，而RMSE接近0。

5.1.3 训练细节

所有实验均在一台计算服务器上进行，服务器配备*Intel(R) Xeon(R)Gold6226R* CPU@*2.90GHz*、38GB RAM，以及*NVIDIA A100-PCIE*GPU@*40GB*×6。

接下来，我们将介绍各个训练阶段的设置，包括：1）预训练、2）自监督训练，以及3）监督训练。请注意，我们使用*Adam* 作为*Python*工具箱*PyTorch*中的优化求解器。

自监督预训练（阶段$_{\mathrm{PT}}$）。我们首先在 MS-COCO数据库上对图像特征提取器 $\mathcal{F}_{img}$ 和文本特征提取器 $\mathcal{F}_{txt}$ 进行预训练。我们将768个样本作为一个批次（即， 128×6），并对模型进行 50 个训练轮次的预训练。初始学习率设为 $1e$-3，并按余弦 调度进行衰减。

自监督训练（阶段$_{\mathrm{SS}}$）。我们使用所提出的 MLIQ数据库中的低光照图像与相应基于QSD的文本配对来训练 BMQA 模型。此时，我们将批量大小减小至 256，并将训练轮次总数设为 20，初始学习率固定为 $4e$-5。实验中我们发现，训练不足（例如，损失大于 1.0）或训练过度（例如，损失小于 0.6）的模型都无法取得更好的性能。根据经验，我们会随机选取损失约为 0.8 的最终模型。

监督训练（阶段$_{\mathrm{ST}}$）。对于 VGG、ResNet 和 Effi- cientNet，我们将初始学习率设为 $8e$-5，并降低

表1**MLIQ数据库上 15 $\mathcal{F}_{img}$ 与 $\mathcal{F}_{txt}$ 组合的总体性能比较.**

$\mathcal{F}_{img}$	$\mathcal{F}_{txt}$	皮尔逊线性相关系数 ↑	SRCC ↑	均方根误差 ↓
VGG-19	BoW	0.8538	0.8532	0.0885
	RNN	0.8670	0.8636	0.0869
	TransF	0.8731	0.8685	0.0847
RN-50	BoW	0.8743	0.8750	0.0845
	RNN	0.8859	0.8851	0.0824
	TransF	0.8989	0.8877	0.0830
EN-B4	BoW	0.8760	0.8719	0.0834
	RNN	0.8898	0.8886	0.0810
	TransF	0.8987	0.8922	0.0825
ViT-B32	BoW	0.8802	0.8774	0.0827
	RNN	0.8879	0.8871	0.0822
	TransF	0.9089	0.9040	0.0816
CNXT-B	BoW	0.8822	0.8844	0.0832
	RNN	0.8970	0.8984	0.0817
	TransF	0.9121	0.9065	0.0802

学习率，在第0.95 倍，并在第*150*和第*250*个训练轮次时进行调整。对于 ViT 和 ConvNeXT，我们将初始学习率设得更低，为 $6.4e$-5 ，并同样在第*150*和第*250*个训练轮次时将学习率降低0.95 倍。最后，我们将批量大小设为 16，并训练质量分数预测器 300 个训练轮次。

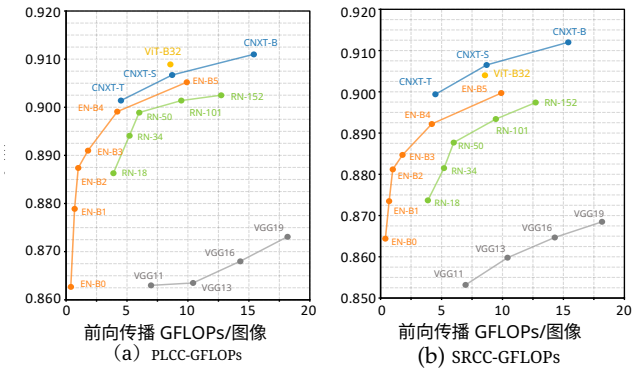
5.2 特征表示验证

为验证两种异构模态特征提取器的有效性，我们在若干具有代表性的 $\mathcal{F}_{txt}$ 和 $\mathcal{F}_{img}$ 模型上进行了实验。具体而言，我们分别采用 VGG-19、ResNet-50（简称RN-50）、EfficientNet（简称EN-B4）、Vision-in-Transformer（简称ViT-B32）和 ConvNeXT（简称CNXT-B）作为图像特征编码器。此外，我们分别采用词袋（简称BoW）、循环神经网络（简称RNN）和 Transformer（简称TransF）作为文本特征编码器。

鉴于 BMQA 的预测准确率主要与特征表示有关，我们首先探讨不同网络骨干对 $\mathcal{F}_{img}$的影响，包括 VGG19、RN-50、EN-B4、ViT-B32 和 CNXT-B。同时，我们采用 BoW、RNN 和 TransF 作为 $\mathcal{F}_{txt}$。表1 给出了 15 $\mathcal{F}_{txt}$与 $\mathcal{F}_{img}$ 组合的整体比较结果。如图所示，全部 15 种 BMQA 变体均取得了令人满意的性能，这验证了我们多模态范式出色的鲁棒性。此外，采用更强大的特征编码器时，相关组合的预测准确率更高。

此外，考虑到图像是 BIQA 任务中的主要模态，我们进一步探究了不同网络变体 $\mathcal{F}_{img}$的影响。图9给出了五种 $\mathcal{F}_{img}$ 模型的 PLCC 和 SRCC 曲线。如图所示，我们可以得出一些有趣的结论：1）同一网络架构的不同变体对最终结果的影响很小

预测结果。例如，将 ResNet 网络从最重的 RN-152 替换为最轻的 RN-18，PLCC 仅从 0.8863 降至 0.9025 ，SRCC 则从 0.8737 降至 0.8974 ；2）轻量级模型也取得了具有竞争力的性能。例如，EN-B0 仅需 0.39G 浮点运算次数，便可达到 0.8627 的 PLCC 分数和 0.8532 的 SRCC 分数。这表明，我们的 BMQA 框架能够使轻量级模型保持优异性能。



（a） PLCC-GFLOPs (b) SRCC-GFLOPs

图9.**不同图像特征提取器的性能比较** $\mathcal{F}_{img}$：(a) 不同网络变体的 PLCC-GFLOPs 曲线；(b) 不同网络变体的 SRCC-GFLOPs 曲线。

5.3 消融研究

如上所述，BMQA 基于两种模态，通过三个训练阶段构建而成，包括预训练、多模态自监督和多模态监督。鉴于此，我们开展了消融实验，以验证两种不同模态在不同学习阶段的有效性。详细结果见表 2。

多模态自监督。 为展示两种模态的作用，我们首先验证了方法 (a) 和方法 (b) 中单一模态的性能，其中表2中的"-"表示未使用相关模态或训练。换言之，BMQA 分别采用 RN-50 作为图像编码器 $\mathcal{F}_{img}$ 或采用 TransF 作为文本编码器 $\mathcal{F}_{txt}$。如结果所示，方法 (a) 优于方法 (b)，这证实了图像数据是视觉质量评估任务中的主要模态。

为验证阶段ss的效果，我们还设计了方法 (c)，在方法 (a) 的基础上进行自监督学习，以实现 TransF ($\mathcal{F}_{txt}$) 与 RN-50 ($\mathcal{F}_{img}$) 之间的特征对齐，如第 4.4.1节所述。如结果所示，方法 (c) 的效果优于方法 (a)，这表明引入多模态自监督有助于深度模型训练。方法 (d) 与方法 (b) 之间也观察到了相同的现象。

多模态监督。为验证阶段st的效果，我们还采用方法 (e) 进行了实验，该方法使用 RN-50 作为 $\mathcal{F}_{img}$，TransF 作为 $\mathcal{F}_{txt}$，以端到端方式进行多模态质量监督。可以看出，方法 (e) 的结果优于方法 (c) 和方法 (d)，这表明引入多模态监督学习能够提升 BMQA 的性能。

预训练策略。 为了展示 $\mathcal{F}_{txt}$ 和 $\mathcal{F}_{img}$之间的特征对齐性能，我们进一步在 MS-COCO数据库上对方法 (e)进行预训练。具体而言，我们采用了三种不同的预训练策略：1）方法 (f)验证预训练对分类任务的效果（记为 CL），其中 RN-50经过预训练以区分 $\mathcal{F}_{img}$ 的类别，而 TransF经过预训练以区分 $\mathcal{F}_{txt}$的类别。2）方法 (g)验证预训练对复原任务的效果（记为 RE），其中 RN-50经过预训练以复原 $\mathcal{F}_{img}$ 中缺失的图像内容，而 TransF经过预训练以补全 $\mathcal{F}_{txt}$中缺失的词语信息。3）方法 (h)验证预训练对 $\mathcal{F}_{img}$ 与

表2**各训练阶段贡献的消融研究**：自监督预训练阶段PT、自监督训练阶段SS和监督训练阶段ST。

方法	阶段PT	阶段SS		阶段ST		皮尔逊线性相关系数↑	SRCC↑	均方根误差↓
		$\mathcal{F}_{img}$	$\mathcal{F}_{txt}$	$\mathcal{F}_{img}$	$\mathcal{F}_{txt}$			
(a)	-	-	-	RN-50	-	0.7964	0.7934	0.1082
(b)	-	-	-	-	TransF	0.7504	0.7466	0.1130
(c)	-	RN-50	TransF	RN-50	-	0.8567	0.8556	0.0902
(d)	-	RN-50	TransF	-	TransF	0.8209	0.8183	0.0932
(e)	-	RN-50	TransF	RN-50	TransF	0.8614	0.8610	0.0896
(f)	CL	RN-50	TransF	RN-50	TransF	0.8707	0.8709	0.0868
(g)	RE	RN-50	TransF	RN-50	TransF	0.8753	0.8721	0.0871
(h)	FM	RN-50	TransF	RN-50	TransF	0.8989	0.8877	0.0830

$\mathcal{F}_{txt}$之间特征对齐的效果，其中 RN-50和 TransF经过预训练，以对齐图像和文本特征。实验表明，方法 (h)的结果优于方法 (f)和 (g)，这说明在预训练中引入多模态质量对齐能够提升盲图像质量评估（BIQA）的性能。

5.4 总体性能比较

为展示整体质量预测性能，我们将 BMQA 与 25 种具有代表性的 BIQA 方法在 MLIQ数据库上进行比较。为保证公平，所有比较方法均采用第 5.1 节所述的相同实验设置。根据质量特征提取的类型，这些盲图像质量评估方法分为两类：1）手工设计方法；2）深度学习方法。

手工设计的BIQA方法。 考虑到 MLIQ 中的低照度图像具有混合多重失真特征，基于手工设计特征的对比方法由以下四类组成：

- **GP**：通用型盲图像质量评估方法，包括*Mittal2012TIP*[35], *Liu2014SPIC* [26], 和*Zhang2015TIP* [63]。
- **MD**：针对多重失真的失真特定型盲图像质量评估方法，包括*Li2016SPL* [24]。
- **CD**：针对对比度失真的失真特定型盲图像质量评估方法，包括*Gu2017TCB* [15]。
- **LL**：针对低光照失真的失真特定型盲图像质量评估方法，包括*Xiang2020TMM* [52], *Wang2021ICME* [48],和*Wang2022TII* [47]。

需要指出的是，*Xiang2020TMM* [52]的实现尚未公开，因此我们提供了自行重新实现的结果。为确保公平比较，表 3 中的其他所有结果均通过运行相应作者发布的实现获得。

深度学习BIQA方法。考虑到 MLIQ 中的低照度图像具有一些特定失真，基于深度学习特征的对比方法可分为四类：

- **AS**：基于分配的监督型BIQA，包括 *Kang2014CVPR* [19]， *Bosse2018TIP* [2]， 和*Ke2021ICCV* [20]。
- **GS**：基于生成的监督型BIQA，包括*Ying2020CVPR* [61]。
- **MT**：基于多任务的监督型BIQA，包括*Kim2019TNNLS* [21]， *Yan2019TMM* [56],*Zhang2020TCSVT* [65]。
- **多目标**：基于多目标的有监督 BIQA，包括 *Wu2020TIP* [51], *Li2020ACMMM*[23],*Su2020CVPR* [41], *Zhu2020CVPR* [69]。
- **MC**：采用跨数据集学习的约束的基于多目标的有监督 BIQA，包括*Ma2021ACMMM*[31]， *Zhang2021TIP* [66]， 和*Zhang2022TPAMI* [64]。
- **MU**：基于度量的无监督 BIQA，包括 *Wang2022ACMMM* [49] 和 *Madhusudana2022TIP*[32]。
- **DU**：基于领域的无监督 BIQA，包括*Liu2019TPAMI* [29]。

性能比较。表3汇总了25 种代表性 BIQA 方法的整体结果。此外，我们还提供了三种 BMQA 变体的结果，包括 BMQA^{image}-$text$ $_{\text{RN-50+TransF}}$、BMQA^{image}-$text$ $_{\text{ViT-B32+TransF}}$ 和 BMQA^{image}-$text$ $_{\text{CNXT-B+TransF}}$，它们分别以 TransF 作为 $\mathcal{F}_{txt}$，并以 RN-50、ViT-B32 和 CNXT-B 作为 $\mathcal{F}_{img}$。

从表3可以看出，三种 BMQA 的平均预测准确率显著优于其他25 种具有竞争力的 BIQA 方法。表现最佳的 BMQA^{image}-$text$ $_{\text{CNXT+TransF}}$ 取得了0.9121的 PLCC 分数、0.9065的 SRCC 分数，以及0.0802的 RMSE 分数，在真实失真的 BIQA任务中表现十分出色。此外，实验结果还表明，我们的 BMQA 在设备-I、设备-V以及整个数据库上的性能没有显著差异，这说明不同的拍摄设备在多模态学习中只会产生轻微的副作用。

5.5 进一步讨论

本节主要讨论我们提出的 BMQA^{image_only} 方法在当前主流单图像模态场景中的适用性和泛化能力。

5.5.1 动机

从第5.2 节到5.4节，我们已从特征表示、消融研究和整体性能等方面验证了 BMQA^{image}-$text$ 的有效性。然而，所提出的 BMQA 面临着缺少文本描述这一挑战。换言之，问题在于，当缺少基于 QSD 的文本模态时，如何使用我们的 BMQA。为应对这一挑战，我们进一步提出了一种可行方案，并在另一个独立的低光照数据库 Dark-4K 上进行了验证。该数据库仅包含图像样本及其平均意见分标签。

5.5.2 额外模态生成

考虑到在盲图像质量评估任务中，图像质量主要可通过相关语言信息来表征，我们认为 X_{txt} 也可以由近期的图像描述模型'show and tell'（SAT）替代 [53]。因此，当文本模态不可用时，BMQA 会采用图像描述模型生成X_{txt}的特征表示。

据我们所知，现有的图像描述模型中，很少有专门针对 BIQA任务进行训练的，因此生成的字幕与视觉质量体验的相关性较低，可能不符合第 3.1.2 节所述的文本生成原则。因此，我们基于 QSD 训练了一个专用的字幕生成模型 $\mathcal{F}_{cap}$，基于 MLIQ 数据库（例如，按 8:1:1 划分）。

表3**所提出的BMQA方法与**25 **MLIQ数据库上最先进的盲图像质量评估方法**的性能比较。针对不同设备，手工设计和深度学习盲图像质量评估方法中的最佳结果以**粗体**显示，而我们提出的BMQA方法的最佳结果以**下划线**标示。

方法类型	设备-I（尼康 *D5300*）			设备II（*iPhone 8 Plus*）			设备III（*iPad mini2*）			设备IV（佳能*EOS*）			设备-V（华为*Mate*）			整个数据库		
手工设计的BIQA方法 类型 皮尔逊系																		
Mittal2012TIP [35] 通用型	0.7833	0.7657	0.1073	0.7865	0.7695	0.0906	0.8005	0.7706	0.0802	0.7273	0.7136	0.1247	0.7709	0.7500	0.1100	0.7797	0.7716	0.1095
Liu2014SPIC [26] 通用型	0.7848	0.7666	0.1072	0.7894	0.7710	0.0904	0.7777	0.7593	0.0904	0.7301	0.7178	0.1265	0.7678	0.7454	0.1109	0.7779	0.7689	0.1101
Zhang2015TIP [63] 通用型	0.6537	0.6704	0.1303	0.6888	0.6747	0.1069	0.7883	0.8328	0.0837	0.5836	0.5832	0.1496	0.6736	0.6560	0.1286	0.6186	0.6346	0.1379
Li2016SPL [24] 多重失真型	0.8267	0.8242	0.0961	0.7626	0.7335	0.0936	0.7935	0.7651	0.0821	0.8442	0.8338	0.0986	0.8767	0.8676	0.0824	0.8317	0.8273	0.0972
Gu2017TCB [15] 对比度失真型	0.8486	0.8446	0.0911	0.7957	0.7918	0.0893	0.7845	0.7878	0.0844	0.7721	0.7764	0.1171	0.7726	0.7658	0.1105	0.7726	0.7658	0.1105
Xiang2020TMM [52] 低光照失真型	0.8265	0.8221	0.1003	**0.8642**	**0.8530**	0.0885	0.8105	0.8004	0.0871	0.8323	0.8146	0.0730	**0.8789**	**0.8642**	**0.0881**	0.8063	0.7950	0.1011
Wang2021ICME [48] 低光照失真型	0.8619	0.8539	0.0868	0.8304	0.8179	0.0798	0.8002	0.7759	0.0805	0.8584	0.8538	0.0942	0.8386	0.8234	0.0954	0.8365	0.8350	0.0969
Wang2022TII [47] 低光照失真型	**0.8647**	**0.8571**	**0.0845**	0.8401	0.8373	**0.0815**	**0.8336**	**0.8459**	**0.0713**	**0.9085**	**0.8892**	**0.0789**	0.8577	0.8488	0.0936	**0.8578**	**0.8488**	**0.0904**
深度学习BIQA方法 类型 皮尔逊线性																		
Kang2014CVPR [19] 分配型监督	0.8103	0.8197	0.1580	0.7954	0.7987	0.1537	0.6822	0.6992	0.1792	0.8301	0.8629	0.1681	0.8783	0.8874	0.1575	0.8095	0.8167	0.1604
Bosse2018TIP [2] 分配型监督	0.9077	0.9154	0.0918	0.7395	0.7198	0.0979	0.6824	0.6305	0.1281	0.8695	0.8596	0.1198	0.8516	0.8615	0.1164	0.8182	0.8198	0.1070
Ke2021ICCV [20] 分配型监督	0.8614	0.8464	0.0884	0.8343	0.8254	**0.0822**	0.8297	0.8208	0.0738	0.8956	0.8835	0.0800	0.8362	0.8312	**0.0944**	0.8504	0.8487	0.0923
Ying2020CVPR [61] 生成型监督	0.8458	0.8427	0.0926	**0.8561**	**0.8468**	0.0893	0.8117	0.7953	0.0854	0.8628	0.8138	**0.0700**	0.8842	0.8616	0.0846	0.8450	0.8455	0.0921
Kim2019TNNLS [21] 多任务	0.9096	0.9166	0.0863	0.7424	0.7213	0.0963	0.6828	0.6305	0.1333	0.8713	0.8609	0.1213	0.8515	0.8633	0.1180	0.8217	0.8234	0.1063
Yan2019TMM [56] 多任务	0.8244	0.8276	0.1631	0.8030	0.8105	0.1575	0.7156	0.6992	0.1755	0.8690	0.8775	0.1700	0.8398	0.8535	0.1624	0.8256	0.8370	0.1639
Zhang2020TCSVT [65] 多任务	0.8720	0.8845	0.1075	0.7118	0.7141	0.1035	0.6681	0.5929	0.1250	0.8423	0.8340	0.1163	0.7631	0.7558	0.1213	0.8241	0.8261	0.1035
Wu2020TIP [51] 多目标	0.8713	0.8786	0.1249	0.7132	0.7188	0.1163	**0.8516**	**0.8277**	0.0721	0.9265	**0.9356**	0.0976	0.8537	0.8498	0.0989	0.8299	0.8371	0.1099
Li2020ACMMM [23] 多目标	0.8561	0.8514	0.0877	0.7981	0.7894	0.0882	0.8074	0.8275	**0.0706**	0.8866	0.8740	0.0869	0.8174	0.8028	0.0995	0.8236	0.8223	0.0988
Su2020CVPR [41] 多目标	0.8645	0.8581	0.0843	0.7933	0.7898	0.0791	0.8171	0.8226	0.0664	**0.9303**	0.9250	0.0677	0.8401	0.8427	0.0893	0.8520	0.8510	0.0882
Zhu2020CVPR [69] 多目标	0.9248	0.9311	0.0721	0.7709	0.7626	0.0922	0.6819	0.6283	0.1211	0.8890	0.8814	0.1096	0.8513	0.8603	0.1031	0.8519	0.8547	0.0948
Ma2021ACMMM [31] MC	0.8594	0.8623	0.0871	0.7971	0.7927	0.0877	0.8049	0.8017	0.0771	0.8631	0.8592	0.0903	0.8092	0.7997	0.0987	0.8271	0.8275	0.0982
Zhang2021TIP [66] MC	0.8848	0.8838	0.0880	0.8487	0.8466	0.0832	0.8479	0.8051	0.0708	0.8476	0.8482	0.1052	0.8147	0.8135	0.1008	0.8485	0.8605	0.0929
Zhang2022TPAMI [64] MC	**0.9277**	**0.9320**	**0.0713**	0.7808	0.7712	0.0911	0.6811	0.6403	0.1117	0.8938	0.8870	0.1037	**0.8908**	**0.8977**	0.0975	**0.8596**	**0.8615**	**0.0910**
Liu2019PTPAMI [29] DU	0.8507	0.8451	0.1105	0.8185	0.8100	0.0979	0.8096	0.8269	0.0644	0.8300	0.8121	0.1095	0.7593	0.7346	0.1206	0.8103	0.8140	0.1081
Wang2022ACMMM [49] MU	0.8657	0.8608	0.0870	0.8210	0.8206	0.0854	0.8265	0.8122	0.0747	0.8764	0.8665	0.0862	0.8311	0.8215	0.0994	0.8383	0.8373	0.0973
Madhusudana2022TIP [32] MU	0.9179	0.9254	0.0784	0.7708	0.7656	0.0935	0.7152	0.6569	0.1099	0.8908	0.8842	0.1042	0.8474	0.8579	0.1017	0.8495	0.8531	0.0943
BMQA$^{image\text{-}text}$ MU	0.9348	0.9187	0.0695	0.9064	0.8813	0.0686	0.9075	0.8712	0.0638	0.9364	0.9140	0.0726	0.9011	0.8860	0.0816	0.8989	0.8877	0.0830
BMQA$^{image\text{-}text}_{\text{RN50+TransF}}$ MU	0.9320	0.9264	0.0715	**0.9131**	**0.9042**	0.0664	**0.9108**	**0.9067**	**0.0598**	**0.9488**	0.9282	0.0737	**0.9041**	**0.8913**	0.0828	0.9089	0.9040	0.0816
BMQA$^{image\text{-}text}_{\text{CNXT-B+TransF}}$ MU	**0.9404**	**0.9310**	**0.0692**	0.9128	0.9022	**0.0653**	0.8920	0.8744	0.0607	0.9464	**0.9358**	**0.0717**	0.8959	0.8850	**0.0808**	**0.9121**	**0.9065**	**0.0802**



图 10. **在我们的 Dark-4K 数据库上生成文本模态的示例**。每个示例均包含一个原始图像样本（左侧）、一个质量认知注意力图（右侧）、在MS-COCO数据库上训练的预训练 SAT 字幕模型（$\mathbf{SAT}^{\mathbf{MS\text{-}COCO}}$）生成的预测字幕结果，以及在 MLIQ数据库上训练的$\mathbf{SAT}^{\mathbf{MLIQ}}$ 生成的基于低光照QSD的字幕结果。

具体而言，每个图像样本都会输入 $\mathcal{F}_{cap}$，并将相应的文本描述用作训练标签。学习目标是最大化预测字幕与人工标注字幕的联合成对概率，或者等价地，最小化训练集上的对数似然：

$$\min \mathbb{E}\left\{-\log \sum_{1<k<N} \mathcal{P}\left[\mathcal{F}_{cap}\left(X_{img};\, \boldsymbol{\theta}_{cap}\right), x_k\right]\right\}, \quad (14)$$

其中，$\boldsymbol{\theta}_{cap}$ 表示 $\mathcal{F}_{cap}$的模型权重。

因此，给定图像 X_{img}，文本模态由以下方式生成

$$< x'_1, \cdots, x'_N >= \mathcal{F}_{cap}\left(X_{img};\, \boldsymbol{\theta}_{cap}\right), \quad (15)$$

其中，$< x'_1, \cdots, x'_N >$表示预测的基于 QSD 的文本模态。

为了验证 $\mathcal{F}_{cap}$的可行性，我们直接采用 SAT 模型作为 $\mathcal{F}_{cap}$的骨干网络。在实验中，我们采用由 [53](记为 $\text{SAT}_{MS-COCO}$)，并在 MLIQ数据库上针对 QSD 微调一个专用的图像描述模型，训练 100 个训练轮次（记为SAT^{MLIQ}）。最终，测试集上的图像描述匹配性能 TOP-5 准确率达到 89。74%，BLEU-4 分数达到29。36%。

图 10 展示了 SAT 基于 QSD 的一些描述结果^{MLIQ} 和 $\text{SAT}_{MS-COCO}$ 在 Dark-4K 数据库上的结果。如图所示，SAT^{MLIQ} 包含更多与视觉体验相关的质量词语，例如亮度状态、显著物体和颜色状态。为了进一步探究这些关键词与原始图像之间的关系，我们提供了与这些关键词对应的注意力图。如图所示，感兴趣区域（ROI）与质量关键词一致。因此，我们认为，经过良好训练的基于 QSD 的描述模型能够胜任生成 MLIQ 中提供的文本模态。

5.5.3 图像单模态情形下的跨数据库验证

为验证 $\mathcal{F}_{cap}$ 在式（15）中的作用，我们在 Dark-4K 数据库上进行了跨数据库验证。为确保公平比较，所有盲图像质量评估方法（包括我们的 BMQA 模型）均未经过微调或重新训练。换言之，所有训练好的模型均与第 5.4 节中的完全相同。值得注意的是，由于 $\mathcal{F}_{cap}$ 在 Dark-4K 数据库中不存在，因此使用一个专门训练的模型 $\mathcal{F}_{cap}$ 替代$\mathcal{F}_{cap}$，用于**算法** 1中。除此之外，其他设置与第5.1.3节相同。

表4 给出了所有17 种深度学习盲图像质量评估方法在皮尔逊线性相关系数、SRCC 和均方根误差方面的总体比较结果。如图所示，我们的 BMQA 方法在跨数据集验证中达到了最先进的性能。最佳 BMQA$^{image\text{-}only}_{\text{CNXT-B+TransF}}$ 获得了皮尔逊线性相关系数分数0.9156，SRCC

表 4**所提出的 BMQA 与**17 **Dark-4K 数据库上的深度学习BIQA方法的性能比较**。深度学习BIQA方法的最佳结果以**粗体**突出显示，我们的 BMQA 的最佳结果以**下划线**突出显示。

方法类型		整个数据库		
深度学习BIQA方法	Type	皮尔逊线性相关系数	SRCC ↑	均方根误差 ↓
Kang2014CVPR [19]	AS	0.5434	0.5605	0.1307
Bosse2018TIP [2]	AS	0.6552	0.6442	0.1028
Ke2021ICCV [20]	AS	**0.8630**	**0.8382**	0.1195
Ying2020CVPR [61]	GS	0.6453	0.5802	0.0631
Kim2019TNNLS [21]	MT	0.5968	0.6030	0.1205
Yan2019TMM [56]	MT	0.5588	0.5610	0.1322
Zhang2020TCSVT [65]	MT	0.5187	0.5581	0.1346
Wu2020TIP [51]	MO	0.6216	0.5877	0.1248
Li2020ACMMM [23]	MO	0.6138	0.6358	0.1107
Su2020CVPR [41]	MO	0.7914	0.7829	0.0755
Zhu2020CVPR [69]	MO	0.7791	0.7616	0.0931
Ma2021ACMMM [31]	MC	0.7713	0.7419	0.0810
Zhang2021TIP [66]	MC	0.8079	0.8131	0.0697
Zhang2022TPAMI [64]	MC	0.8538	0.8358	**0.0593**
Liu2019TPAMI [29]	DU	0.6319	0.6352	0.1002
Wang2022ACMMM [49]	MU	0.8526	0.8357	0.0687
Madhusudana2022TIP [32]	MU	0.8293	0.8240	0.0669
BMQA$^{image\text{-}only}_{\text{RN+TransF}}$	MU	0.9026	0.8874	0.0624
BMQA$^{image\text{-}only}_{\text{ViT-B32+TransF}}$	MU	0.9117	0.8989	**0.0593**
BMQA$^{image\text{-}only}_{\text{CNXT-B+TransF}}$	MU	**0.9156**	**0.9085**	0.0605

分数为 0.9085，RMSE 分数为0.0605。实验结果验证了我们 BMQA 框架的适用性和泛化性能。

6 结论与未来工作

本文从主观和客观两个角度，对低光照图像的盲多模态质量评估（BMQA）进行了新的研究。具体而言，我们建立了首个低光照图像多模态质量评价数据库，并为多模态基准的构建设计了两项质量感知原则。此外，我们通过探索质量特征表示、对齐与融合以及多模态学习，研究了 BMQA 框架。实验结果验证了，与两种不同的低光照基准数据库上的最先进方法相比，我们的 BMQA 方法具有有效性。我们相信，这项工作能够为图像和视频质量评价带来新的研究视角。

未来，BMQA 可从多个方面开展进一步研究和验证。首先，我们展示了多模态数据如何提升低照度应用场景下盲图像质量评估（BIQA）的性能，而非多模态数据如何共同决定视觉质量体验。后者更具挑战性，因此需要部署更多传感器并融合多传感器数据。

其次，我们设计了两项质量语义描述（QSD）原则，并且在多模态学习中仅考虑了一些基于QSD的文本特征。然而，图像质量感知可能非常复杂，因此还需要进一步研究更多可解释属性例如清晰度、构图和审美性。

第三，我们采用多模态自监督和监督机制来训练基线模型。目前的设计针对低光照图像简单且高效，但对于 BMQA 框架而言，图像与文本模态之间的交互仍有待进一步探索，这将有助于提取更高效的特征表示。

最后但同样重要的是，我们构建了一个多模态低照度图像数据库，并提出了一种有效的无参考质量指标，因为这一方向更具挑战性，也更加迫切。然而，多模态驱动的质量评估的可行性仍需在更多基准数据库上验证，包括面向特定失真和通用应用的数据库。此外，还需研究并验证其在全参考任务中的可行性，因为这种质量评估范式仍应用于许多现实场景。

参考文献

[1] Tadas Baltruˇsaitis、Chaitanya Ahuja 和 Louis-Philippe Morency。多模态机器学习：综述与分类。*IEEE*模式分析与机器智能汇刊，41(2)：423–443，2018年。[2] Sebastian Bosse、Dominique Maniry、Klaus-Robert M¨uller、Thomas Wiegand 和 Wojciech Samek。用于无参考和全参考图像质量评估的深度神经网络。*IEEE*图像处理汇刊，27(1)：206–219，2018年。[3] I BT。电视图像质量主观评估方法，文件《ITU-R BT.500-14建议书》（2019年10月）。国际电信联盟，日内瓦，瑞士，2020年。[4] Yuqin Cao、Xiongkuo Min、Wei Sun 和 Guangtao Zhai。用于全参考和无参考音视频质量评估的深度神经网络。载于*IEEE*国际图像处理会议（*ICIP*），第1429–1433页，2021年。[5] Bingzhi Chen、Qi Cao、Mixiao Hou、Zheng Zhang、Guangming Lu 和 David Zhang。具有时间和语义一致性的多模态情绪识别。*IEEE*音频、语音与语言处理汇刊，29(1)：3592–3603，2022年。[6] Chen Chen、Qifeng Chen、Jia Xu 和 Vladlen Koltun。学习在暗处看见。载于*IEEE*计算机视觉与模式识别会议（*CVPR*），第3291–3300页，2018年。[7] Liang Chen、Jiawei Zhang、Jinshan Pan、Songnan Lin、Faming Fang 和 Jimmy S Ren。为夜间模糊图像学习非盲去模糊网络。载于*IEEE*计算机视觉与模式识别会议（*CVPR*），第10542–10550页，2021年。[8] Ziteng Cui、Guo-Jun Qi、Lin Gu、Shaodi You、Zenghui Zhang 和 Tatsuya Harada。用于暗光目标检测的正交切线正则化多任务AET。载于*IEEE* 国际计算机视觉会议（*ICCV*），第2553–2562页，2021年。[9] Xueqing Deng、Peng Wang、Xiaochen Lian 和 Shawn Newsam。NightLab：一种用于夜间分割的双层架构与难度检测。载于*IEEE*计算机视觉与模式识别会议（*CVPR*），第16938–16948页，2022年。[10] Yubin Deng、Chen Change Loy 和 Xiaoou Tang。图像美学评价：一项实验性综述。*IEEE*信号处理杂志，34(4)：80–106，2017年。[11] Sidney K D’mello 和 Jacqueline Kory。多模态情感检测系统的综述与元分析。*ACM* 计算调查，47(3)：1–36，2015年。[12] Yuming Fang、Liping Huang、Jiebin Yan、Xuelin Liu 和 Yang Liu。全向图像的感知质量评估。载于*AAAI*人工智能会议（*AAAI*），第580–588页，2022年。[13] Ian E Gordon。视觉感知理论。心理学出版社，2004年。[14] Richard Langton Gregory。作为假设的知觉。伦敦皇家学会哲学汇刊*B*辑：生物科学，290(1038)：181–197，1980年。[15] Ke Gu、Jun Zhou、Jun-Fei Qiao、Guangtao Zhai、Weisi Lin 和 Alan Conrad Bovik。屏幕内容图像的无参考质量评估。*IEEE*图像处理汇刊，26(8)：4005–4018，2017年。[16] David S Hands。基础多媒体质量模型。*IEEE*多媒体汇刊，6(6)：806–816，2004年。[17] Alan Hanjalic 和 Li-Qun Xu。情感视频内容表示与建模。*IEEE*多媒体汇刊，7 154，2005 1 143( ): – 。[18] Mingxing Jiang、Liquan Shen、Min Hu、Ping An 和 Fuji Ren。用于电子显示器的色调映射高动态范围图像及多曝光融合图像的盲质量评估器。*IEEE*消费电子汇刊，67(4)：350–362，2021年。[19] Le Kang、Peng Ye、Yi Li 和 David Doermann。用于无参考图像质量评估的卷积神经网络。载于*IEEE*计算机视觉与模式识别会议（*CVPR*），第1733–1740页，2014年。

[20] Junjie Ke、Qifei Wang、Yilin Wang、Peyman Milanfar 和 Feng Yang。多尺度图像质量Transformer。载于 IEEE 国际计算机视觉会议（ICCV），第 5148–5157 页，2021年。[21] Jongyoo Kim、Anh-Duc Nguyen 和 Sanghoon Lee。基于深度卷积神经网络的盲图像质量预测器。IEEE 神经网络与学习系统汇刊，30(1)：11–24，2019年。[22] Chongyi Li、Chunle Guo、Ling-Hao Han、Jun Jiang、Ming-Ming Cheng、Jinwei Gu 和 Chen Change Loy。使用深度学习进行低照度图像与视频增强：综述。IEEE 模式分析与机器智能汇刊，44 9416，2022 12 9396( )：–。[23] Dingquan Li、Tingting Jiang 和 Ming Jiang。用于图像质量评估的范数中的范数损失：收敛更快、性能更优。载于 ACM 多媒体国际会议（MM），第 789–797 页，2020年。[24] Qiaohong Li、Weisi Lin 和 Yuming Fang。梯度域中多重失真图像的无参考质量评估。IEEE 信号处理快报，23(4)：541–545，2016年。[25] Chunxiao Liu、Zhendong Mao、Tianzhu Zhang、Anan Liu、Bin Wang 和 Yongdong Zhang。聚焦注意力：一种用于多模态学习的焦点注意力机制。IEEE 多媒体汇刊，24(1)：103–115，2022年。[26] Lixiong Liu、Bao Liu、Hua Huang 和 Alan Conrad Bovik。基于空间熵和频谱熵的无参考图像质量评价。Elsevier《信号处理：图像通信》，29(8)：856–863，2014年。[27] Pengfei Liu、Weizhe Yuan、Jinlan Fu、Zhengbao Jiang、Hiroaki Hayashi 和 Graham Neubig。预训练、提示与预测：自然语言处理中提示方法的系统综述。ACM 计算调查，55(9)：1–35，2023年。[28] Tsung-Jung Liu 和 Kuan-Hsien Liu。采用宽感知域评分器集成方法的无参考图像质量评价。IEEE 图像处理汇刊，27(3)：1138–1151，2018年。[29] Xialei Liu、Joost Van De Weijer 和 Andrew D Bagdanov。通过自监督排序学习利用卷积神经网络中的无标签数据。IEEE 模式分析与机器智能汇刊，41(8)：1862–1878，2019年。[30] Yuen Peng Loh 和 Chee Seng Chan。借助纯暗光数据集了解低光照图像。Elsevier《计算机视觉与图像理解》，178：30–42，2019年。[31] Rui Ma、Hanxiao Luo、Qingbo Wu、King Ngi Ngan、Hongliang Li、Fanman Meng 和 Linfeng Xu。记忆与重用：通过相关性感知增量学习进行跨任务盲图像质量评估。载于 ACM 多媒体国际会议（MM），第 5248–5256 页，2021年。[32] Pavan C Madhusudana、Neil Birkbeck、Yilin Wang、Balu Adsumilli 和 Alan C Bovik。使用对比学习进行图像质量评估。IEEE 图像处理汇刊，31(1)：4149–4161，2022年。[33] Helard A Becerra Martinez 和 Mylene CQ Farias。结合音频与视频指标评估视听质量。Springer《多媒体工具与应用》，77(18)：23993–24012，2018年。[34] Xiongkuo Min、Guangtao Zhai、Jiantao Zhou、Mylene CQ Farias 和 Alan Conrad Bovik。视听信号主观与客观质量评估研究。IEEE 图像处理汇刊，29(5588)：6054–6068，2020年。[35] Anish Mittal、Anush Krishna Moorthy 和 Alan Conrad Bovik。空间域中的无参考图像质量评价。IEEE 图像处理汇刊，21(12)：4695–4708，2012年。[36] Margaret H Pinson、Lucjan Janowski、Romuald P´epion、Quan Huynh-Thu、Christian Schmidmer、Phillip Corriveau、Audrey Younkin、Patrick Le Callet、Marcus Barkowsky 和 William In-gram。受试者与环境对视听主观测试的影响：一项国际研究。IEEE 信号处理选题期刊，6(6)：640–651，2012年。[37] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。通过自然语言监督学习可迁移的视觉模型。载于 PMLR 机器学习国际会议（ICML），第 8748–8763 页，2021 年。[38] Michele A Saad 、 Alan C Bovik 和 Christophe Charrier。盲图像质量评价：一种基于 DCT 域自然场景统计的方法。IEEE 图像处理汇刊，21(8)：3339–3352，2012年。[39] Guoli Song、Shuhui Wang、Qingming Huang 和 Qi Tian。基于高斯过程潜变量模型的协调式多模态学习。IEEE 模式分析与机器智能汇刊，43(3)：858–872，2019年。[40] Nisan Stiennon、Long Ouyang、Jeffrey Wu、Daniel Ziegler、Ryan Lowe、Chelsea Voss、Alec Radford、Dario Amodei 和 Paul F Christiano。利用人类反馈学习摘要生成。神经信息处理系统（NeurIPS）进展，第 3008–3021 页，2020年。[41] Shaolin Su、Qingsen Yan、Yu Zhu、Cheng Zhang、Xin Ge、Jinqiu Sun 和 Yanning Zhang。借助自适应超网络，在自然场景中盲评图像质量。载于 IEEE计算机视觉与模式识别会议（CVPR），第 3667–3676 页，2020年。[42] Yu Tian、Huanqiang Zeng、Junhui Hou、Jing Chen、Jianqing Zhu 和 Kai-Kuang Ma。一种基于对称性和深度特征的光场图像质量评估模型。IEEE视频技术电路与系统汇刊，31(5):2046–2050，2021年。[43] Oriol Vinyals、Alexander Toshev、Samy Bengio 和 Dumitru Erhan。Show and tell：2015年MS COCO图像描述挑战赛的经验总结。IEEE模式分析与机器智能汇刊，39(4):652–663，2016年。[44] Nicholas Wade 和 Mike Swanston。视觉感知：入门。心理学出版社，2013年。[45] Guotao Wang、Chenglizhao Chen、Deng-Ping Fan、Aimin Hao 和 Hong Qin。从语义类别到注视点：一种新颖的弱监督视听显著性检测方法。载于 IEEE计算机视觉与模式识别会议（CVPR），第 15119–15128 页，2021年。[46] Miaohui Wang、Yijing Huang、Jiaxin Lin、Wuyuan Xie、Guanghui Yue、Shiqi Wang 和 Leida Li。通过前景感知和背景抑制测量屏幕图像质量。IEEE仪器与测量汇刊，70 11，2021 1: – 。[47] Miaohui Wang、Yijing Huang、Jian Xiong 和 Wuyuan Xie。自然场景中的低光照图像：一种由可见性感知引导的全新盲质量指标。IEEE工业信息学汇刊，1(1):1–11，2022年。[48] Miaohui Wang、Yijing Huang 和 Jialin Zhang。通过弱光照分析盲评夜间图像质量。载于 IEEE多媒体与博览会国际会议（ICME），第 1–6 页，2021年。[49] Miaohui Wang、Zhuowei Xu、Yuanhao Gong 和 Wuyuan Xie。S-CCR：用于自然场景低光照图像质量推断的超完备比较表征。载于 ACM 多媒体国际会议（MM），第 5219–5227 页，2022年。[50] Stefan Winkler 和 Christof Faller。低码率多媒体内容的视听感知质量。IEEE多媒体汇刊，8(5):973–980，2006年。[51] Jinjian Wu、Jupo Ma、Fuhu Liang、Weisheng Dong、Guangming Shi 和 Weisi Lin。利用级联深度神经网络进行端到端盲图像质量预测。IEEE图像处理汇刊，29(1):7414–7426，2020年。[52] Tao Xiang、Ying Yang 和 Shangwei Guo。盲夜间图像质量评估：主观方法与客观方法。IEEE多媒体汇刊，22(5):1259–1272，2019年。[53] Kelvin Xu、Jimmy Ba、Ryan Kiros、Kyunghyun Cho、Aaron Courville、Ruslan Salakhudinov、Rich Zemel 和 Yoshua Bengio。Show, attend and tell：利用视觉注意力生成神经图像描述。载于 PMLR机器学习国际会议（ICML），第 2048–2057 页，2015年。[54] Mai Xu、Lai Jiang、Chen Li、Zulin Wang 和 Xiaoming Tao。基于视口的CNN：一种用于评估 360度视频质量的多任务方法。IEEE模式分析与机器智能汇刊，44(4):2198–2215，2022年。[55] Yuecong Xu、Haozhi Cao、Kezhi Mao、Zhenghua Chen、Lihua Xie 和 Jianfei Yang。为动作识别领域自适应对齐相关性信息。IEEE 神经网络与学习系统汇刊，1(1):1–12，2022年。[56] Bo Yan、Bahetiyaer Bare 和 Weimin Tan。自然度感知深度无参考图像质量评估。IEEE多媒体汇刊，21(10):2603–2615，2019年。[57] Qi Yang、Zhan Ma、Yiling Xu、Zhu Li 和 Jun Sun。通过图相似度推断点云质量。IEEE模式分析与机器智能汇刊，44(6):3015–3029，2022年。[58] Wen Yang、Jinjian Wu、Shiwei Tian、Leida Li、Weisheng Dong 和 Guangming Shi。结合层次语义退化的细粒度图像质量描述。IEEE图像处理汇刊，31(1):3578–3590，2022年。[59] Junjie Ye、Changhong Fu、Guangze Zheng、Ziang Cao 和 Bowen Li。DarkLighter：照亮黑暗，实现无人机跟踪。载于IEEE智能机器人与系统国际会议（IROS），第 3079–3085 页，2021年。[60] Zhenqiang Ying、Deepti Ghadiyaram 和 Alan Bovik。临场感视频质量评估。载于 Springer欧洲会议

计算机视觉（欧洲计算机视觉国际会议），第 327–347 页，2022年。[61] Zhenqiang Ying、Haoran Niu、Praful Gupta、Dhruv Mahajan、Deepti Ghadiyaram 和 Alan C. Bovik。从图像块到整幅图像（PaQ-2-PiQ）：描绘图像质量的感知空间。载于*IEEE*计算机视觉与模式识别会议（*CVPR*），第 3575–3585 页，2020年。[62] Guangtao Zhai、Xiaolin Wu、Xiaokang Yang、Weisi Lin 和 Wenjun Zhang。基于自由能原理的心理视觉质量指标。*IEEE*图像处理汇刊，21(1):41–52，2012年。[63] Lin Zhang、Lei Zhang 和 Alan C Bovik。特征增强的完全盲图像质量评价器。*IEEE*图像处理汇刊，24(8):2579–2591，2015年。[64] Weixia Zhang、Dingquan Li、Chao Ma、Guangtao Zhai、Xiaokang Yang 和 Kede Ma。盲图像质量评价的持续学习。*IEEE*模式分析与机器智能汇刊，1(1):1–15，2022年。[65] Weixia Zhang、Kede Ma、Jia Yan、Dexiang Deng 和 Zhou Wang。使用深度双线性卷积进行盲图像质量评价。神经网络。*IEEE*视频技术电路与系统汇刊，30(1):36–47，2020年。[66] Weixia Zhang、Kede Ma、Guangtao Zhai 和 Xiaokang Yang。《实验室与自然场景中的不确定性感知盲图像质量评价》。*IEEE*图像处理汇刊，30(1):3474–3486，2021年。[67] Yonghua Zhang、Xiaojie Guo、Jiayi Ma、Wei Liu 和 Jiawan Zhang。《超越低光照图像增亮》。*Springer*国际计算机视觉期刊，129(4):1013–1037，2021年。[68] Yannan Zheng、Weiling Chen、Rongfu Lin、Tiesong Zhao 和 Patrick Le Callet。《UIF：一种用于水下图像增强的客观质量评价方法》。*IEEE*图像处理汇刊，1:5456–5468，2022年。[69] Hancheng Zhu、Leida Li、Jinjian Wu、Weisheng Dong 和 Guangming Shi。《MetaIQA：用于无参考图像质量评价的深度元学习》。载于*IEEE*计算机视觉与模式识别会议（*CVPR*），第14143–14152页，2020年。