arXiv:2602.07403v1 [eess.IV] 7 Feb 2026

监控人脸图像质量评估：多维数据集与轻量级模型

Yanwei Jiang、Wei Sun、Yingjie Zhou、Xiangyang Zhu、Yuqin Cao、Jun Jia、
Yunhao Li、Sijing Wu、Dandan Zhu、Xingkuo Min、Guangtao Zhai，*IEEE*会士

摘要——监控人脸图像通常在非受控条件下采集，由于低分辨率、运动模糊、遮挡和光照不佳等因素，会导致严重的质量下降。尽管近年来应用于监控摄像机的人脸复原技术能够显著提升视觉质量，但它们往往会损害保真度（即身份保持特征），这与监控图像的首要目标——可靠的身份验证——直接冲突。现有的人脸图像质量评价（FIQA）主要侧重于视觉质量或面向识别的评估，因而未能同时兼顾对监控应用至关重要的视觉质量与保真度。为弥合这一差距，我们提出了首项关于监控人脸图像质量评估（SFIQA）的综合性研究，聚焦监控场景所固有的独特挑战。具体而言，我们首先构建了*SFIQA-Bench*，这是一个面向监控人脸图像的多维质量评估基准，包含由三种广泛部署的监控摄像机在真实场景中采集的5,004幅监控人脸图像。我们开展主观实验，收集六个维度的质量评分，包括噪声、锐度、色彩丰富度、对比度、保真度和整体质量，涵盖SFIQA的关键方面。此外，我们提出*SFIQA-Assessor*，一种轻量级多任务人脸图像质量评估模型，通过跨视图特征交互联合利用互补的人脸视图，并采用可学习任务令牌引导多个质量维度的统一回归。在所提出数据集上的实验结果表明，与最先进的通用图像质量评估（IQA）和面部图像质量评估方法相比，我们的方法取得了最佳性能，验证了其在实际监控应用中的有效性。

索引词—**监控人脸图像，图像质量评估，人脸保真度，深度神经网络**

I. 引言

在数字时代，公共监控系统已迅速迅速普及，已成为现代智能基础设施的基石。监控人脸图像（SFIs）作为这些系统的核心视觉输入，在公共安全、身份验证和访问控制场景中发挥着不可或缺的作用 [1]。然而，与在受控环境中拍摄的人像图像不同，监控人脸图像通常是在高度不受控的条件下获取的，容易受到低分辨率、运动模糊、遮挡、非理想拍摄角度以及光照不足或不均匀等因素的影响，导致图像质量严重下降。这些质量退化给人眼检查和自动人脸识别系统处理监控人脸图像带来了重大挑战。

Yanwei Jiang、Yingjie Zhou、Yuqin Cao、Jun Jia、Yunhao Li、Sijing Wu、Xiongkuo Min 和 Guangtao Zhai 任职于上海交通大学图像通信与网络工程研究所，地址：中国上海 200240。（电子邮箱：{jiang-yan-wei、zyj2000、caoyuqin、jia-jun0302、lyhsjtu、wusijing、minxiongkuo、zhaiguangtao}@sjtu.edu.cn）。Wei Sun 任职于华东师范大学通信与电子工程学院，地址：中国上海 200241。（电子邮箱：wsun@cee.ecnu.edu.cn）。Dandan Zhu 任职于华东师范大学计算机科学与技术学院，地址：中国上海 200333。（电子邮箱：ddzhu@mail.ecnu.edu.cn）。Xiangyang Zhu 任职于上海人工智能实验室，地址：中国上海 200232。（电子邮箱：zhuxiangyang@pjlab.org.cn）。通信作者：Wei Sun（wsun@cee.ecnu.edu.cn）。本研究得到国家自然科学基金委员会（62301316、62225112）资助。版权所有 © 2026 IEEE。允许将本材料用于个人用途。但如需将本材料用于其他用途，必须发送电子邮件至 pubs-permissions@ieee.org，向 IEEE 获取许可。



图 1。（a）高质量参考人脸图像。（b）通过 JPEG压缩和模糊生成的对应退化版本。（c）和（d）由两种具有代表性的人脸复原方法生成的复原图像 [2],[3]。虽然复原提升了感知质量，但也可能引入幻觉纹理或伪结构，从而损害身份保真度。

鉴于这些挑战，近期人脸复原领域的进展 [2]–[5]已应用于监控影像，旨在提升视觉质量。虽然这些技术在一定程度上能够改善退化图像的感知效果，但也可能引入伪结构或幻觉细节，从而损害人脸保真度（见图 1），直接削弱监控成像的首要目标：可靠的身份验证。这一问题凸显了迫切需求：开发一种可靠且全面的质量评估模型，能够有效评估监控人脸图像的感知质量和人脸保真度，确保现实场景中与人脸相关的应用值得信赖。

现有的人脸图像质量评估（FIQA）研究主要关注两个方面。第一方面是感知人脸图像质量评估，例如 TransFQA [6],和 DSL-FIQA [7], ，旨在根据

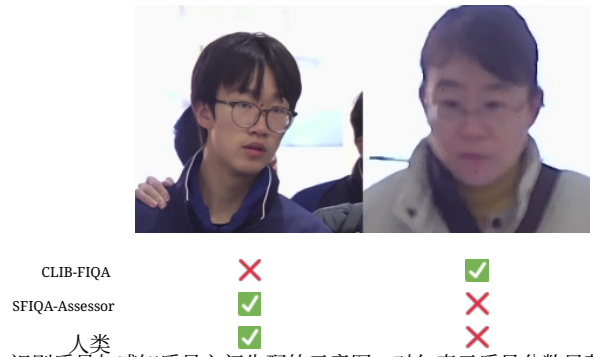

图2. 识别质量与感知质量之间失配的示意图。对勾表示质量分数显著高于叉号。CLIB-FIQA分数反映识别质量，而SFIQA-Bench分数代表感知质量。与CLIB-FIQA相比，所提出的SFIQA-Assessor生成的分数与人类判断更加一致。

人类的主观评分评估人脸图像的感知视觉质量。这些研究中的人脸图像主要来源于互联网或社交媒体平台，并由人类观察者标注质量分数。然而，这些数据集缺少在现实监控场景中采集的人脸图像，也未考虑人脸增强技术对人脸保真度的影响，因此限制了它们在监控环境中的适用性。

第二类是面向识别的FIQA，其代表方法包括FaceQNet [8]，、SER-FIQ [9]，和CLIB-FIQA [10]，，旨在根据人脸图像在人脸识别系统中的实用性评估其识别质量。这些方法通常使用人脸识别数据集中的人脸图像作为评估对象。在这种设置下，质量标签通常源自人脸识别模型，而非人类判断。例如，测试图像与其对应注册图像之间的匹配分数通常被用作质量的伪真值指标。然而，由于现代人脸识别系统对各种失真具有内在的鲁棒性，识别质量与感知质量之间往往存在失配。也就是说，识别质量高的图像不一定具有较高的感知质量，如图2所示。此外，现有面向识别的FIQA研究也很少考虑人脸修复技术对人脸保真度的影响。这些局限使基于识别的质量指标不太适用于监控应用。

为弥补这些不足，我们首次对监控人脸图像质量评估开展了全面研究，系统探讨了在现实监控场景中评估感知质量和人脸保真度所面临的挑战。为实现这一目标，我们首先构建了**SFIQA-Bench**，这是一个针对监控人脸图像的多维质量评估基准。不同于以往主要从互联网收集人脸图像的图像质量评估研究（其成像技术可能已经过时，且拍摄条件与监控环境中的条件存在显著差异），我们的数据集由三种广泛部署的监控摄像机拍摄的真实场景监控图像组成。这些摄像机配备了集成式人脸增强模块，可在拍摄环境较差时自动启动。我们总共收集了 5,004 张监控人脸图像，涵盖三个具有代表性的场景：室内人员监控、室外行人监控，以及面向车内人员的智能交通系统（ITS）监控，涵盖环境光照变化、拍摄角度等多种环境条件。这确保了该基准具有广泛的泛化能力和很强的实用价值。

由于监控人脸图像通常是在低照度、极端俯仰角、运动等具有挑战性的条件下拍摄的，因此往往呈现出多种复杂失真。为系统地研究其感知质量，我们从四个维度评估监控人脸图像：**噪声**、**锐度**、**色彩丰富度**和**对比度**，从而识别视觉退化的具体成因，并支持可解释的感知质量评估。为应对人脸增强技术引入的非自然伪影，我们增加了**保真度**维度，用于评估图像是否包含不真实、不自然或幻觉式人脸内容。我们还纳入**整体质量** 维度，以分析各因素的相对重要性及其与整体质量的相关性。为获得可靠的标注，我们组织了一项涉及100名参与者的大规模主观实验。在受控观看条件下，每张图像均由至少25名受试者从全部六个维度进行评分，从而确保标注质量高且一致。

此外，我们提出了**SFIQA-Assessor**，这是一种基于深度神经网络（DNN）的轻量级模型，专为监控人脸图像的实时多维质量评估而设计。模型以三种互补的人脸视图作为输入：包含人脸和背景的原始图像、聚焦于人脸区域的裁剪后的人脸图像，以及突出关键面部特征的眼嘴区域。为整合这些视图，我们提出了**面部感知质量特征编码器**，从三个视图中提取多尺度特征，并通过轻量级跨视图注意力模块将其融合，从而高效捕捉视图间依赖关系并形成统一质量表征。随后，我们设计了**任务感知质量解码器**，其中结合了任务自注意力模块和交叉注意力模块。首先，通过任务自注意力细化一组可学习任务令牌，以建模任务间关系；随后将其用作交叉注意力中的查询，从统一表征中提取任务特定特征。最后，使用独立的**回归头** 将各任务特定特征映射为相应的质量分数。实验结果表明，SFIQA-Assessor在所有质量维度上均持续优于最先进方法，同时保持极低的计算复杂度，验证了其在真实监控应用中的实用价值。

我们的贡献总结如下：

- 我们构建了 SFIQA-Bench，这是首个专为监控人脸图像设计的多维质量评估基准。该基准包含 5,004 张图像，这些图像采集自三种具有代表性的现实监控场景，并使用三种广泛部署的相机型号，确保了广泛的泛化能力和实用价值。
- 我们在 SFIQA-Bench 上开展了全面的主观实验，收集了六个质量维度的人工评分：噪声、锐度、色彩丰富度、对比度、保真度和整体质量，从而提供可解释且

细粒度标签，用于评估监控人脸图像的感知质量和人脸保真度。

- 我们提出了 SFIQA-Assessor，这是一种轻量级的人脸图像质量评估模型，可从多个人脸视图中提取并融合多尺度特征，以评估六个质量维度，在实现实时高效的同时达到最先进的性能。

II. 相关工作

A. 图像质量评估

多年来，图像质量评估（IQA）[1] 已得到广泛研究，并提出了大量数据集 [11]–[21]和模型 [21]–[48] 。早期的 IQA 数据集主要关注 JPEG 压缩、高斯噪声、模糊等合成失真。LIVE [11],、CSIQ [12], 和 TID2013 [15] 等常用数据集通常包含数量有限的高质量参考图像，每张图像都经过几种常见合成失真的处理，因此失真图像总数从数百张到数千张不等。KADID-10k [18] 数据集大幅扩展了合成 IQA 基准的规模，其中包含 81 张原始图像和 25 种失真，失真样本超过 10,000 张。近年来，IQA 研究的重点已转向真实场景图像，这些图像由相机直接拍摄或从互联网收集，并呈现由真实世界采集和传输条件造成的真实失真。具有代表性的真实场景 IQA 数据集包括 BID [13], 、CLIVE [16], 、KonIQ-10k [19], 、SPAQ [20],和 FLIVE [17],, 其数据集规模从数百张到一万多张图像不等。表 I 提供了这些数据集的详细信息。

图像质量评估模型大致可分为两大类：手工特征方法和基于深度神经网络的方法。手工特征方法通常依赖自然场景统计（NSS） [22]–[24], [49] ，或纹理 [50]–[52], 、边缘 [53]–[55], 、自由能线索 [56]等其他人工设计的特征。DIIVINE [24], 、BRISQUE [23],、NIQE [22], 等经典手工特征方法，主要针对合成失真图像设计。然而，由于建模能力有限，这些方法在真实场景中的泛化能力往往较差。相比之下，基于深度神经网络的方法利用深度神经网络强大的表征学习能力提取质量感知特征，在合成失真和真实失真方面均取得了更好的表现。早期的基于深度神经网络的模型 [25], [26],[57] 通常采用浅层卷积神经网络（CNN），通过处理图像块来学习局部质量表示。近期的研究 [27]–[33], [58], [59] 则引入多尺度特征提取器 [27], [28], [58], 、注意力机制 [29], [31],[59], 、CLIP变体提供的语义先验 [30], [33], 以及大型多模态模型（LMMs） [32] 等先进架构，以提升鲁棒性和准确性。为进一步提高泛化能力，一些研究探索了多数据集训练 [33], [35] ，或采用无监督学习策略 [34], ，从而能够利用大规模未标注数据，并提升模型在不同条件下的适应性。

B. 人脸图像质量评价

现有的人脸图像质量评估研究主要分为两类：感知人脸图像质量评估和面向识别的FIQA。

感知人脸图像质量评估关注人脸图像的感知视觉质量，与传统图像质量评估研究更为接近。Su 等人 [60] 提出了GFIQA-20K，这是一个人脸图像质量评估数据集，包含来自互联网的20,000张标注有主观质量评分的人脸图像，并提出了一种生成先验引导的图像质量评估模型，用于人脸质量预测。Jo 等人 [61]提出了IFQA，这是一种新颖的可解释的人脸质量评估方法，采用生成对抗框架来模拟人脸复原，并提供与人类感知一致的像素级质量评分。IFQA重点关注眼睛、鼻子和嘴巴等关键人脸区域，在多种架构和复杂场景下均优于传统的通用及人脸专用质量指标。Chahine等人 [62]提出了一个肖像图像质量评估数据集，包含在50个不同场景中采集的5,116张图像，并附有专家对细节、曝光和整体质量的成对标注。他们进一步提出了SEM-HyperIQA，一种适配场景特定质量尺度的语义感知模型。Sun 等人 [63] 提出了一种用于肖像图像质量评估的双分支网络，利用独立的Swin Transformer主干网络分别提取全图与人脸区域特征，并通过学习排序策略进行优化。Liu等人 [6] 构建了一个同时包含真实场景图像与合成失真的人脸图像的人脸图像质量评估数据集，并提出了Trans-FQA，一种双分支Transformer 模型，能够捕获全局、人脸及部件级上下文以及失真感知线索，以提高预测准确率。Chen 等人 [7]提出了CGFIQA-40K，这是一个均衡且多样的人脸图像质量评估数据集，包含40,000张图像，旨在缓解肤色与性别偏差。他们还提出了DSL-FIQA，该方法采用双集合退化表征学习策略，同时利用真实失真与合成失真以增强泛化能力，并引入面部标志点引导的注意力，以聚焦感知重要区域。

面向识别的人脸图像质量评估旨在通过预测与识别性能相关的质量分数，估计人脸图像在识别任务中的效用。FaceQnet [8],[64] 将该任务建模为有监督回归问题，并使用预训练人脸识别模型的相似度分数作为伪标签。为避免依赖真实质量标注，SER-FIQ [9] 提出了一种无监督方法，在识别网络中使用随机失活进行多次随机前向传播，并将嵌入一致性作为质量代理指标。SDD-FIQA [65] 采用自监督策略，计算类内与类间相似度分布之间的 Wasserstein 距离，以此反映图像区分不同身份的能力。CLIB-FIQA [10] 进一步利用视觉语言模型（例如 CLIP [66]），根据模糊和遮挡等可解释的视觉退化因素生成伪质量标签，并采用置信度校准机制来减轻噪声监督的影响。

[1]本文聚焦于无参考图像质量评估（亦称盲图像质量评估）方法，因为人脸图像质量评估任务通常属于这一类别。

表 I 优秀IQA/FIQA数据集汇总

数据集	Task	图像类型	来源	评价维度	# 图像	分辨率	标签	失真
LIVE [11]	IQA	通用	专业相机	整体质量	779	$\leq 768 \times 512$	DMOS	合成
CSIQ [12]	IQA	通用	专业相机	整体质量	866	512×512	MOS	合成
TID2013 [15]	IQA	通用	专业相机	整体质量	3,000	512×384	MOS	合成
KADID-10k [18]	IQA	通用	专业相机	整体质量	10,125	512×384	DMOS	合成
BID [13]	IQA	通用	专业相机	整体质量	6,585	2272×1704	MOS	真实
CLIVE [16]	IQA	通用	智能手机	整体质量	1,162	1024×768	MOS	真实
KonIQ-10k [19]	IQA	通用	互联网	整体质量	10,073	1024×768	MOS	真实
SPAQ [20]	IQA	通用	智能手机	亮度、色彩丰富度、对比度 噪声程度、锐度、整体质量	11,125	$\leq 768 \times 512$	MOS	真实
FLIVE [17]	IQA	通用	互联网	整体质量	40,000	640×640	MOS	真实
FERET [67]	FIQA	Face	专业相机	识别质量	14,126	256×384	ID	真实
LFW [68]	FIQA	Face	互联网	识别质量	13,233	250×250	ID	真实
CASIA-WebFace [69]	FIQA	Face	互联网	识别质量	500,000	-	ID	真实
VGGFace2 [70]	FIQA	Face	互联网	识别质量	331万	-	ID	真实
GFIQA-20K [60]	FIQA	Face	互联网	整体质量	20,000	512×512	MOS	真实
PIQ23 [62]	FIQA	肖像	智能手机	曝光、细节、整体质量	5,116	1900×2400	MOS	真实
CGFIQA-40K [7]	FIQA	Face	互联网	整体质量	39,312	512×512	MOS	真实
TransFQA 数据集 [6]	FIQA	Face	互联网	整体质量	42,125	$\geq 512 \times 512$	MOS	混合
SFIQA-Bench	FIQA	Face	监控摄像头	噪声、锐度、色彩丰富度 对比度、保真度、整体质量	5,004	$\leq 768 \times 768$	MOS	真实

III. SFIQA-BENCH

本节将详细介绍我们的监控人脸图像质量评估基准SFIQA-Bench，包括图像采集流程、主观质量评估和基准分析。

A. 数据采集

我们收集真实场景监控图像，用于构建SFIQA-Bench，以捕捉由多样化环境和摄像设备所产生的真实失真。数据涵盖三种典型场景：在受控照明（开/关）和简化背景下的室内人员监控；背景复杂的室外行人监控；以及针对车内人员的智能交通系统（ITS）监控。后两种场景均包含白天和夜间条件，以反映真实场景中的照明变化。我们使用了三种不同的设备：Huawei M2241-10-QLI-E2-B 双镜头摄像机、HikVision DS-2CD7C888MWD 双镜头摄像机，以及 HikVision DS-2CD7V887MWD 四镜头监控摄像机。每台设备均以其内置的智能模式运行，在人员经过时自动采集人脸图像，并仅保存人脸区域。此外，当检测到环境条件较差时，摄像机便会启动人脸自动增强模式，以提升视觉质量。图3展示了在不同监控场景下采集的监控人脸图像示例。摄像机安装高度为 2.5 至 3.5 米，俯仰角介于 15° 至 30° 之间。在室内和室外场景中，摄像机分别安装在大学校园内的 6 个办公地点和 12 个室外位置。ITS 场景的监控图像则与交警部门合作采集。图像采集过程持续了两个月，共获得 11,164张原始人脸图像。随后，我们进行了人工审核和数据清洗。我们剔除了偏航角过大的图像（大于 60°），以及人脸小于 96×96像素的图像。对于存在明显意外遮挡的图像，例如人脸被手、他人的身体部位或



图3. SFIQA-Bench中真实监控人脸图像的代表性示例。这些图像涵盖六种典型场景：室内、室外，以及白天和夜间采集的车内监控图像。

树叶等静态物体也被剔除。此外，还排除了描绘同一身份的过度相似图像。对于口罩，我们选择不剔除戴口罩人脸，因为口罩已成为日常生活中的常见事物。

经过这一清理过程，最终数据集包含 5,004 张监控人脸图像。图4和图5展示了若干关键分布，包括场景类型（室内、室外和 ITS 车载场景）、光照条件（弱光和正常光照）以及图像分辨率直方图。这些分布表明，我们的 SFIQA-Bench 涵盖了多样的监控人脸图像，符合真实采集条件。

B. 主观实验

收集监控人脸图像后，我们进行了严格的人工主观评估，以获取每张图像的质量分数。

1）评价维度：我们从多个维度评估监控人脸图像的质量，因为单一的总体评分不足以捕捉

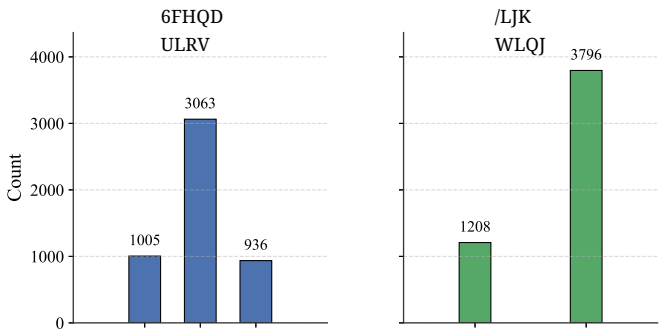

图 4. SFIQA-Bench 中部分属性的分布。

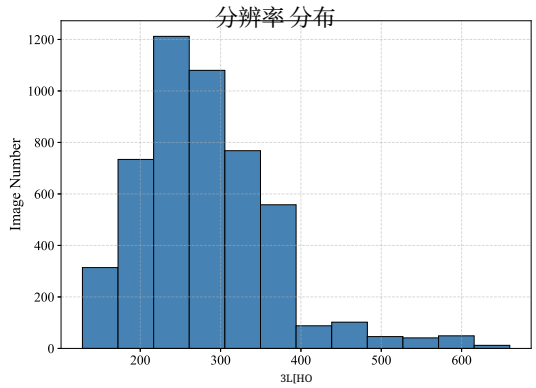

图 5. SFIQA-Bench 中图像分辨率的直方图。所有图像均为正方形，因此仅显示分辨率的一边。

此类图像中多样且复杂的失真。例如，监控场景中的光照条件变化很大，从完全黑暗（0 勒克斯）到强烈阳光（超过 10,000 勒克斯）都有。这些变化会导致图像对比度出现显著差异。在低光照环境中，通常会使用红外光来呈现面部纹理。然而，将红外图像与 RGB 图像结合可能会造成色偏，并产生错位或鬼影状纹理。主体移动时也常会出现运动模糊。此外，内置的人脸增强模块可能会生成不自然或虚构的面部特征，从而降低人脸图像的保真度。

为全面评估常见失真，我们定义了六个质量维度：**噪声**、**锐度**、**色彩度**、**对比度**、**保真度**和**整体质量**。前四个维度反映感知质量属性，而保真度则衡量人脸内容的真实性与逼真度。整体质量代表综合评估，涵盖上述所有质量方面。各维度的定义详述如下：

- **噪声** 评估视觉伪影的程度，尤其是在低照度条件下，过多的噪声可能会遮挡重要的面部细节。
- **锐度** 评估面部结构的清晰度与明确程度，这些特征可能会因运动模糊、失焦或低分辨率而受损。
- **色彩丰富度** 评估色彩呈现的准确性，这可能会受到红外干扰、光照不佳或交通信号灯等彩色光源的影响或霓虹灯的影响，导致出现不自然的色调或色偏。
- **对比度** 评估亮部与暗部之间的动态范围。光照变化常会造成过曝或欠曝区域，从而降低整体对比度。
- **保真度**通过识别增强算法引入的伪结构或幻觉细节，评估人脸内容的真实性；这些伪结构或细节可能会损害图像的逼真度和可信度。
- **总体质量** 通过综合上述所有维度，在平衡失真严重程度与人脸内容保真度的基础上，评估整体感知质量。

为更好地理解这些质量维度，我们在图 6 中展示了若干具有代表性的监控人脸图像，每幅图像在相应维度上均呈现出高质量或低质量。

2）主观实验方法：我们采用五点评分的绝对类别评分（ACR）量表对质量分数进行标注，遵循 [19], [71] 中的标准。五个类别——差、较差、一般、良好和优秀——分别对应 1 至 5 分。SFIQA-Bench 的主观实验分为两个阶段：**试点研究**和**正式研究**。在试点研究中，参与者需要完成培训环节。他们首先观看预先录制的教学视频，视频讲解监控人脸图像中的常见失真，以及评估各种质量维度的标准。随后，他们对 20 个经过专家标注的训练样本进行评分。具体而言，专家团队由一名教授、一名专门从事图像质量评估的博士后研究员、两名具有图像质量评估/人脸图像质量评估研究经验的博士生，以及两名来自领先监控设备制造商的研发工程师组成。团队所有成员均在图像质量评估/人脸图像质量评估领域发表过同行评审论文，或参与过与监控成像系统和质量评估相关的产业或学术项目。专家分数通过协作式专家标注流程获得。参与者每次查看一幅图像，在仔细观察后提供 6 个分数，然后再进行下一幅图像的评分。如果参与者在任一质量维度上与专家分数的斯皮尔曼等级相关系数（SRCC）低于 0.9，则必须重复培训，直至达到该阈值。

在正式评估阶段，我们将所有测试图像分为25个场次，每个场次包含200幅测试图像（其中一个场次包含204幅）。为确保可靠性，每个场次加入了5幅“标准图像”（预先标注为最佳质量（评分5分）或最差质量（评分1分）），以及5幅重复图像，用于检查评分一致性。总计每位参与者在每个场次评定208幅图像。如果参与者对标准图像的评分与参考分数相差超过1分，或对重复图像的评分差异超过1分，则将这些评分视为异常值。若异常评分的比例超过10%，则剔除该参与者在该场次中的数据。

我们招募了100名参与者进行主观实验，每场实验由25名参与者评分，共获得750,600条主观评分。根据标准图像

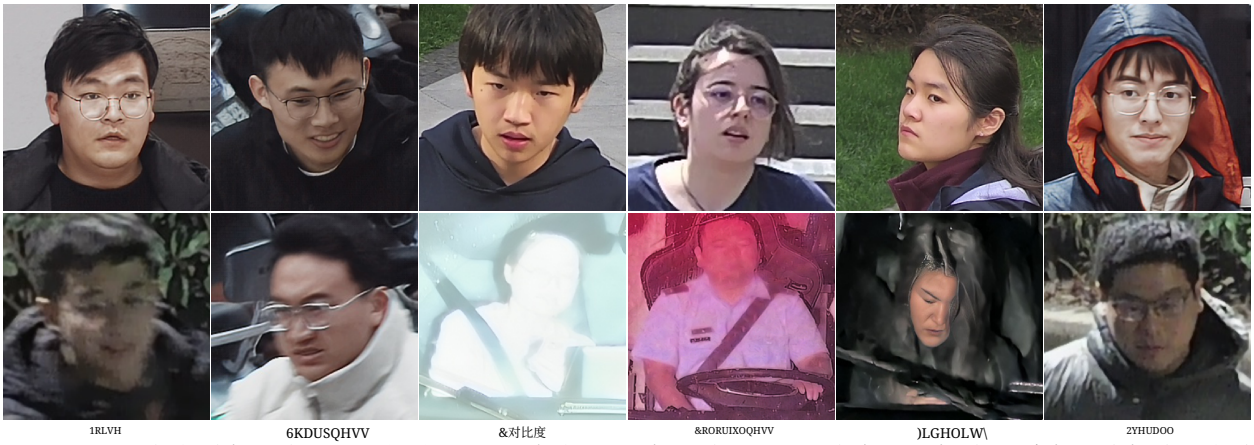

图6. 展示 SFIQA-Bench 各质量维度的代表性监控人脸图像。顶行展示高质量示例，底行则分别展示噪声、锐度、对比度、色彩丰富度、保真度和整体质量方面的低质量对应示例。

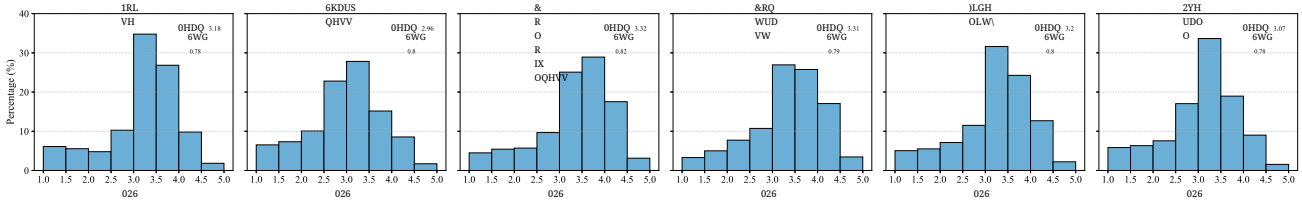

图7. .....的平均意见分分布 SFIQA-Bench 的六个质量维度：噪声、锐度、色彩丰富度、对比度、保真度和整体质量 ty.

以及重复图像检查，我们因评分不一致而剔除了1名参与者和3场实验的数据。随后，我们按照 ITU-T BT.500-13 建议对主观评分进行处理 [72]，，其中包括一种基于评分标准差的异常值检测策略，以及计算平均意见分（MOS）的标准流程。具体而言，我们将与均值偏差显著的个别评分识别为异常值并予以剔除；异常值评分超过5%的参与者则不纳入后续分析。在我们的实验中，共剔除了894条评分（占总评分的0.125%），没有参与者被排除。ITU标准规定的原始MOS计算方法包括将个体评分归一化为Z分数，再线性缩放至 [0, 100] 范围，并对有效参与者的评分取平均。在本研究中，我们省略了Z分数缩放，以使最终分数更好地对应五个预定义的质量类别。按照上述流程，我们获得了SFIQA-Bench中每张监控人脸图像各个维度的真实质量分数。

C. 平均意见分分析

我们从三个方面分析 SFIQA-Bench 中监控人脸图像的平均意见分：

1) ***MOS***分布： 图 7 展示了六个质量维度的 MOS 分布。在所有维度中，MOS 值大致呈正态分布或轻微左偏，大多数评分集中在 2.5 到 4.0 之间。平均 MOS 从 2.96（锐度）到 3.32（色彩丰富度）不等，表明大多数监控人脸图像的质量处于中等水平，且不同维度之间存在明显差异。

值得注意的是，锐度的平均得分最低，这表明面部纹理和结构的损失仍是监控人脸图像中的主要失真。这主要是由行人运动、车内监控和低分辨率等因素造成的。相比之下，色彩丰富度和对比度的 MOS 值相对较高，表明现代监控摄像机在这些方面往往表现更好。噪声和保真度呈现相似的分布，这可能是因为噪声常见于低照度条件下，而这又会触发人脸增强模块。这些增强处理，尤其是基于深度学习的处理，可能会引入伪结构，从而损害面部保真度。

有趣的是，整体质量的分布比各个单独维度的分布更接近高斯分布，但仍略微呈左偏。这可以解释为，整体质量综合了全部五个单独维度，从而产生自然的平均效应，平滑了偏度。总体而言，结果凸显了基准中存在的多种质量问题，并强调了在现实监控场景中进行多维质量评估的必要性。

*2)*不同维度之间的相关性： 图 8 展示了六个质量维度之间的皮尔逊相关系数（PLCC）。整体质量与全部五个单独维度均呈现强相关，其中与保真度的相关性最高，其次是锐度和噪声。这表明，面部结构的真实性、图像清晰度和噪声水平是影响整体质量主观判断的最重要因素。

在五个感知维度中，保真度与噪声和锐度都呈现出很强的相关性。保真度与噪声之间的关系已在

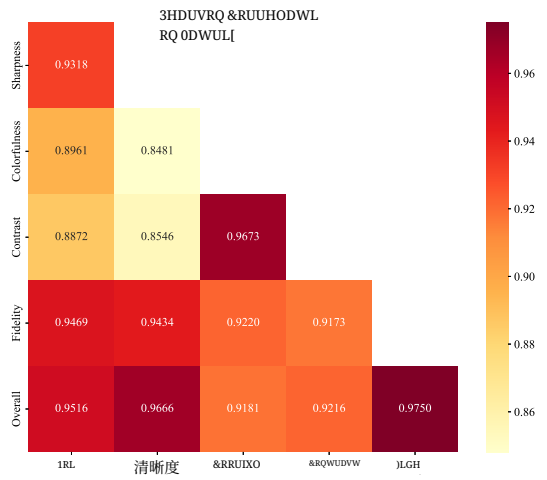

图8. 不同维度的PLCC值矩阵。红色表示PLCC值高，黄色表示低。

MOS分布分析中讨论过。至于保真度与锐度之间的联系，这可能是因为面部结构由细粒度纹理组成，而这些纹理本身会影响感知清晰度。对比度和色彩丰富度也呈现出很强的相关性，这表明这些属性往往会同步变化，可能是因为它们都受环境照明强度和色彩渲染的影响。相比之下，色彩丰富度和对比度与噪声和锐度的相关性较弱，这或许可以用它们性质不同来解释——颜色和对比度是更具整体性的感知属性，而噪声和锐度则与局部结构失真有关。

最后，我们采用最小二乘法 [73]进行多元线性回归分析，以量化各质量维度对整体质量的贡献。拟合模型表示如下：

$$\begin{aligned}\text{Overall} = {} & 0.0765 \cdot \text{Noise} + 0.4134 \cdot \text{Sharpness} \\ & + 0.0554 \cdot \text{Colorfulness} + 0.1463 \cdot \text{Contrast} \\ & + 0.3076 \cdot \text{Fidelity} \end{aligned} \tag{1}$$

该模型的决定系数较高（R^2 =0.9778），表明五个感知维度解释了总体质量评分方差的97.78%。回归结果凸显了锐度和保真度对整体质量的主导贡献，其次是对比度、噪声和色彩丰富度。这些发现总体上与PLCC分析结果一致。

3) 不同场景下的***MOS***分布： 为分析环境和拍摄条件如何影响感知质量，我们在图9中展示了不同场景下MOS的概率密度分布，包括白天与夜间、行人与车辆，以及室内与室外条件。

在白天与夜间的对比中，白天拍摄的图像在所有质量维度上的平均意见分通常都较高。具体而言，白天图像在锐度、保真度、对比度和整体质量方面的得分明显偏高，这体现了充足环境光照带来的益处。相比之下，夜间图像的噪声增加，且锐度降低，这可能是由于照度不足以及相机增益调整所致，进而损害了面部结构表征的质量。

就车辆与行人而言，在车载监控场景（如驾驶员监测）中拍摄的图像，质量往往低于行人场景中拍摄的图像。概率密度分布图显示，在噪声、锐度、保真度和整体质量方面，车辆图像的得分分布均偏向较低分数。这是因为车内环境通常存在运动模糊、车窗反射、光照不均匀和倾斜拍摄角度等挑战性因素。相比之下，行人图像通常在更开放、稳定的条件下拍摄，因此质量分数相对较高。

就室内与室外而言，室内场景的质量分数通常高于室外场景，尤其是在保真度、锐度和对比度方面。室内光照更加稳定且受控，因此图像曝光更一致，失真也更少。相比之下，室外场景更容易受到光照强度变化、逆光和环境干扰的影响，导致多个维度的分布范围更广，平均意见分也更低。

IV. SFIQA-Assessor

本节介绍 SFIQA-Assessor，这是一种面向监控人脸图像的多维客观质量评估模型。考虑到 SFIQA 模型通常部署在资源受限的客户端设备上，例如监控摄像机，因此我们对模型进行了精心优化，使其在保持较低计算复杂度的同时实现高准确率，从而非常适合实际部署。

A. 数据预处理

先前研究表明，在人脸图像的感知质量评估中，人脸区域比非人脸区域更为重要 [6]。在各种面部组成部分中，眼睛与嘴巴对感知到的人脸质量影响最大 [6], [74]。为充分利用这一关键信息，我们执行预处理流程，以提取紧凑人脸区域和眼嘴区域。具体而言，对于由监控摄像机自动采集的原始监控人脸图像（SFI） $\boldsymbol{x}_o$，我们首先使用 Ultra-Light-Fast-Generic-Face-Detector-1MB 工具 [75] 检测紧致裁剪的人脸边界框，并获取紧凑人脸区域 $\boldsymbol{x}_f$。接下来，我们应用人脸解析算法 [76]对 $\boldsymbol{x}_f$ 进行分割，提取面部组成部分，仅保留眼睛与嘴巴区域，从而得到$\boldsymbol{x}_{e,m}$。最后，三种人脸视图——原始宽松裁剪图像 $\boldsymbol{x}_o$、紧凑人脸区域 $\boldsymbol{x}_f$以及眼嘴区域$\boldsymbol{x}_{e,m}$——均输入我们的模型，使其能够同时利用背景上下文信息和感知上最显著的面部特征。这三个输入的示例见图 11。

B. 面部感知质量特征编码器

获取三个人脸视图后，我们设计了一种具备人脸感知能力的质量特征编码器，重点提取人脸区域中与质量相关的特征，从而实现有效的多维质量评价。

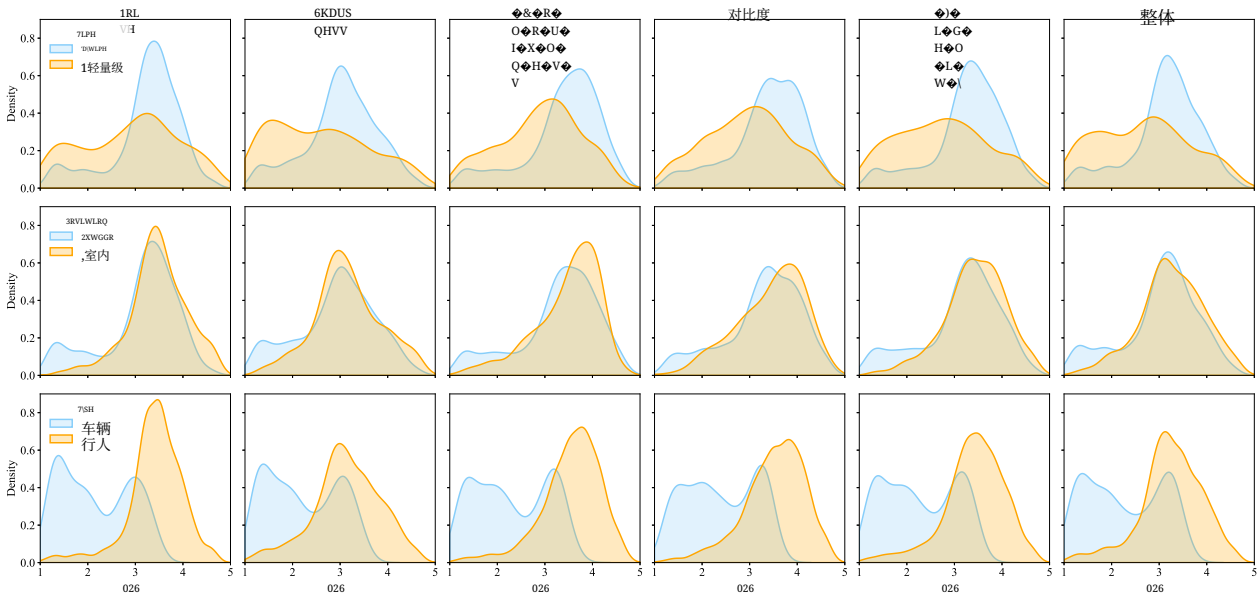

图 9. SFIQA-Bench 中不同监控场景下平均意见分的概率密度分布，包括白天与夜间、室内与室外，以及车辆与行人场景。

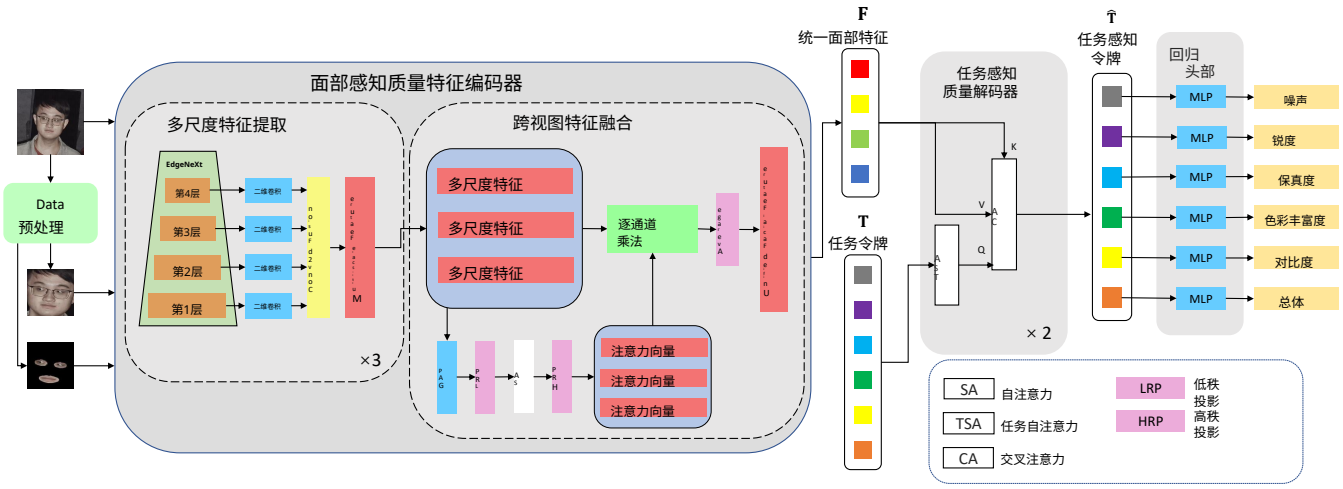

图 10. 我们提出的框架概览。该方法包含预处理模块、面部感知质量特征编码器、任务感知解码器和预测头。输入图像经过预处理，以提取紧凑的人脸区域和眼嘴区域。面部感知质量特征编码器提取多尺度、跨视图的面部质量特征 **F**。任务令牌 **T** 通过任务特定注意力机制和交叉注意力机制，在解码器中与统一面部特征 **F** 融合。随后，所得任务感知令牌被送入回归头，以预测六项质量分数：噪声、锐度、保真度、色彩丰富度、对比度和总体质量。



图 11. 预处理后的图像：原始宽松裁剪图像、紧凑人脸区域、眼嘴区域。

1) 多尺度特征提取：我们采用轻量级 CNN 模型 EdgeNeXt [77], 作为默认骨干网络，以提取提取丰富的视觉特征图（其他主干网络将在第 V-G 节讨论）。既有研究 [78] 表明，CNN 的浅层主要响应低级视觉特征，而深层则捕获高级语义信息。在我们的多维质量评估任务中，对比度、噪声、色彩丰富度和锐度等维度与低级特征密切相关 [23],，而保真度则更多依赖高级语义表征 [79]。因此，我们采用分层多尺度特征提取策略，聚合 EdgeNeXt 所有中间层的特征图，以有效支持多维质量预测。

具体而言，对于三个 $\boldsymbol{x}_o$、 $\boldsymbol{x}_f$和$\boldsymbol{x}_{e,\ m}$人脸视图，我们将每个视图输入骨干模型 $\mathcal{B}$ ，以获得多尺度

特征表示。令 N_s 表示 $\mathcal{B}$中的阶段数。尺度为 i的特征图表示如下：

$$\mathbf{F}_j^{(i)} = \mathcal{B}^{(i)}(\boldsymbol{x}_j), \tag{2}$$

其中$\mathbf{F}_j^{(i)}$表示视图 j中骨干网络第 i阶段的特征图，$i \in \{1, 2, \ldots, N_s\}$表示骨干网络的阶段索引，而 $j \in \{o, f, (e,\ m)\}$对应三个人脸视图。

对于从各个视图中提取的多尺度特征，我们首先使用多层感知机（MLP），将不同尺度的特征图投影到统一的通道维度：

$$\hat{\mathbf{F}}_j^{(i)} = \mathrm{MLP}_{\mathrm{D_i}\to\mathrm{D_o}}(\mathbf{F}_j^{(i)}), \tag{3}$$

其中，D_i 和 D_o 分别表示输入和输出通道维度。随后，沿通道维度拼接所有尺度的投影特征图，并使用 1×1 CNN进行融合：

$$\mathbf{F}_j = \mathrm{Conv}^{1\times 1}_{(\mathrm{N_s}*\mathrm{D_o}\to\mathrm{D_o})}(\mathrm{Concat}(\hat{\mathbf{F}}_j^{(1)}, \hat{\mathbf{F}}_j^{(2)}, ..., \hat{\mathbf{F}}_j^{(N_s)})), \tag{4}$$

其中$\mathbf{F}j$ 表示视图 j的最终融合多尺度特征表示。

2）跨视图特征融合：随后，我们融合从所有视图中提取的多尺度特征，以获得统一的面部感知质量特征表示。为实现有效的跨视图交互，我们采用基于注意力的融合策略，整合不同视图的特征。然而，标准注意力机制的计算复杂度为 $O(n^2)$，呈二次增长，这与我们的轻量化设计原则相悖。为此，我们采用**低秩通道注意力**方法。具体而言，首先通过全局平均池化沿空间维度将每个特征图压缩为全局语义向量：

$$f_j = \mathrm{GAP}(\mathbf{F}_j) = \frac{1}{H\,W}\sum_{h=1}^{H}\sum_{w=1}^{W}\mathbf{F}_j[:, h, w] \ \in \mathbb{R}^{D_o}, \tag{5}$$

其中，H 和 W 分别表示特征图的高度和宽度。随后，使用可学习的线性变换将该向量投影到较低维的嵌入空间中。此过程称为低秩投影（LRP）：

$$\widetilde{f}_j = \mathbf{W}_d f_i \in \mathbb{R}^{D_l}. \tag{6}$$

其中，$\mathbf{W}_d \in \mathbb{R}^{D_o\times D_l}$ 是可学习的投影矩阵，D_l是每个投影特征向量的潜在维度。来自所有视图的$\widetilde{}$投影向量 $\{f_i\}$沿序列（视图）维度进行拼接，以形成注意力模块的输入嵌入：

$$\widetilde{\mathbf{F}} = [\widetilde{f}_o; \widetilde{f}_f; \widetilde{f}_{e,m}] \in \mathbb{R}^{3\times D_l}. \tag{7}$$

接下来，我们引入一个自注意力模块，以建模从三个不同人脸视图中获得的全局向量 $\mathbf{F}$ 之间的相互依赖关系。通过对这一序列应用注意力机制，网络能够学习捕捉不同视图间的互补信息。自注意力机制定义如下：

$$\mathbf{G} = \mathrm{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_h}}\right)V,$$
$$\mathbf{Q} = \mathbf{K} = \mathbf{V} = \widetilde{\mathbf{F}}, \quad \mathbf{G} \in \mathbb{R}^{3\times D_l}, \tag{8}$$

其中，$\mathbf{G}$表示经注意力融合的特征序列。为实现跨视图特征融合，必须将$\mathbf{G}$进行上投影，使其与原始通道数相匹配。该上投影过程可表示如下，称为高秩投影（HRP）：

$$\mathbf{U} = \mathbf{W}_u\mathbf{G} \in \mathbb{R}^{3\times D_o}, \tag{9}$$

其中，$\mathbf{W}_u \in \mathbb{R}^{D_l\times D_o}$ 是可学习的投影矩阵。

最后，执行跨视图融合操作，以获得统一的人脸特征表示$\mathbf{F}$，其计算方式如下：

$$\mathbf{F} = \frac{1}{3}\sum_{j=1}^{3}\mathbf{F}_j \odot u_j \in \mathbb{R}^{D_o\times H\times W}, \tag{10}$$

其中uj 表示来自$\mathbf{U}$的第 j个注意力向量，$\odot$表示在空间维度上广播的逐通道乘法。

C. 任务感知质量解码器

由于我们的模型处理的是一个多任务学习问题，涉及多个维度上质量分数的回归，我们设计了任务感知质量解码器，以将统一面部特征表示$\mathbf{F}$ 解耦为与每个质量维度相对应的任务特定特征向量。具体而言，我们引入一组可学习任务令牌$\mathbf{T}$，它们在交叉注意力模块中作为查询，与共享面部特征表示$\mathbf{F}$进行交互。如III-C节所述，与不同质量维度相关的任务并非完全独立，而是呈现出固有的相关性。为了捕捉这些关系，任务令牌$\mathbf{T}$首先通过任务自注意力模块进行处理，从而实现跨任务的信息交换。更新后的任务令牌$\mathbf{T}'$ 计算如下：

$$\mathbf{T}' = \mathrm{Attn}(\mathbf{Q} = \mathbf{T}, \mathbf{K} = \mathbf{T}, \mathbf{V} = \mathbf{T}). \tag{11}$$

更新后的令牌$\mathbf{T}'$ 随后在交叉注意力模块中用作查询，以关注面部特征表示$\mathbf{F}$。交叉注意力模块的输出记为任务-ˆ感知令牌$\mathbf{T}$，计算如下：

$$\hat{\mathbf{T}} = \mathrm{Attn}(\mathbf{Q} = \mathbf{T}', \mathbf{K} = \mathbf{F}, \mathbf{V} = \mathbf{F}). \tag{12}$$

为提升性能，解码器（包含任务自注意力模块和交叉注意力模块）会应用两次。第一次解码时，任务令牌与共享面部特征交互，以获取粗粒度任务特定线索。第二次解码进一步细化这些令牌：重新建模任务间关系，并再次关注共享特征，从而为每个质量维度生成更稳定且更具判别性的表示。经过这一两阶段解码过程后，ˆ最终的任务感知令牌 $\mathbf{T}$ 随之得到。这些令牌不仅保留了通过任务自注意力机制学习到的任务不变表示，还通过交叉注意力交互捕获了任务特定面部特征。

表 II 不同图像质量评估方法在所提出的 SFIQA-BENCH 数据集上的比较。粗体表示第 1 名，下划线表示第 2 名。黄色高亮的方法代表通用图像质量评估方法，灰色高亮的方法代表面向识别的人脸图像质量评价方法，蓝色高亮的方法代表感知型人脸图像质量评价方法。

方法	噪声		锐度		色彩丰富度		对比度		保真度		总体		参数 (M)	GMACs	延迟 (ms)
	斯皮尔曼等级相关系数	PLCC	斯皮尔曼等级相关系数	PLCC	斯皮尔曼等级相关系数	PLCC	斯皮尔曼等级相关系数	PLCC	斯皮尔曼等级相关系数	PLCC	斯皮尔曼等级相关系数	PLCC			
BRISQUE [23]	0.7553												-	-	-
NIQE [22]	0.0242												-	-	-
HyperIQA [80]	0.7765												27.38	107.83	26.41
TReS [81]	0.8482												152.45	500.04	158.67
MUSIQ [27]	0.8828												78.59	115.37	72.09
MANIQA [31]	0.8058												135.75	2580.2	402.03
StairIQA [28]	0.9224												30.49	10.43	31.02
CLIPIQA [30]	0.6591												152.45	35.25	20.82
LIQE [33]	0.9162												151.28	1506.88	296.10
TOPIQ [29]	0.8874												45.2	20.88	22.74
SER-FIQ [9]	0.5058												43.72	6.35	29.15
SDD-FIQA [65]	0.5872												30.78	6.31	24.87
IFQA [61]	0.8626												105.17	26.58	28.37
DN-PIQA [63]	0.8424												173.69	89.32	80.28
DSL-FIQA [7]	**0.9311**												252.69	2427.2	518.90
SFIQA-Assessor-XXS													**5.28**	**1.39**	**8.26**
SFIQA-Assessor-XS	0.9178												8.26	2.05	9.34
SFIQA-Assessor-S	0.9254												17.76	3.75	10.59

D. 回归头

我们采用一个包含 128、128 和 1 个神经元的三层多层感知机作为回归头，将每个任务特定令牌映射到相应的质量分数维度：

$$\hat{q}_k = \mathrm{MLP}(\hat{t}_k), \quad k \in [1, 2, .., 6], \tag{13}$$

其中$\hat{} t_k$ 表示$\hat{}\mathbf{T}$中的第 k个任务令牌，而$\hat{} q_k$ 是输入人脸图像在第 k个维度上的输出质量分数。

对于损失函数，我们假设所有任务具有同等重要性。因此，我们计算每个质量维度的预测分数与真值标签之间的均方误差（MSE），并取其均值作为最终损失：

$$\mathcal{L} = \frac{1}{6}\sum_{k=1}^{6}(q_k - \hat{q}_k)^2, \tag{14}$$

其中 $\mathcal{L}$ 表示一张人脸图像的总损失， q_k表示第 k个质量维度的平均意见分。

V. 实验

A. 验证数据集

我们首先在 SFIQA-Bench 上评估 SFIQA-Assessor。为进一步评估其泛化能力，我们在两个广泛使用的感知人脸图像质量评估数据集上开展实验：GFIQA-20K [60]和 CGFIQA-40K [7]。这些数据集的详细介绍见表 I。对于 SFIQA-Bench，我们进行五折交叉验证，并按照 7:1:2 的比例将每一折划分为训练集、验证集和测试集。对于 GFIQA-20K，我们遵循 DSL-FIQA [7],采用的数据划分协议，按照 7:1:2 的比例划分训练集、验证集和测试集。对于 CGFIQA-40K，我们采用官方提供的训练集、验证集和测试集划分方式。

表 III GFIQA-20K 和 CGFIQA-40K 上的性能表。

方法	GFIQA-20K		CGFIQA-40K	
	SRCC	PLCC	SRCC	PLCC
HyperIQA [80]	0.8923	0.8846	0.9766	0.9629
TReS [81]	0.7967	0.7670	0.8505	0.8452
MUSIQ [27]	0.9243	0.9203	0.9694	0.9701
MANIQA [31]	0.9639	0.9589	0.9521	0.9561
CLIPIQA [30]	0.8484	0.8519	0.8062	0.8109
StairIQA [28]	0.9624	0.9665	0.9831	0.9843
LIQE [33]	0.9622	0.9405	0.9839	0.9567
TOPIQ [29]	0.9641	0.9581	0.9844	0.9806
SER-FIQ [9]	0.4873	0.5239	0.6037	0.6272
SDD-FIQA [65]	0.5633	0.5938	0.7030	0.7257
IFQA [61]	0.9514	0.9533	0.9779	0.9758
DN-PIQA [63]	0.9430	0.9379	0.9714	0.9724
DSL-FIQA [7]	**0.9696**	**0.9699**	**0.9861**	**0.9865**
SFIQA-Assessor-S	0.9654	0.9666	0.9842	0.9846

B. 对比方法

我们将所提出的 SFIQA-Assessor 与两类图像质量评估方法进行比较：通用图像质量评估方法和面向人脸的图像质量评估方法。通用方法包括 BRISQUE [23],NIQE [22],StairIQA [28], Hyper-IQA [80], MANIQA [31],MUSIQ [27], CLIP-IQA [30],TOPIQ [29],LIQE [33], 和 TReS [81]。面向人脸的方法包括 SER-FIQ [9], SDD-FIQA [65], IFQA [61],DN-PIQA [63], 和 DSL-FIQA [7]。其中，SER-FIQ 和 SDD-FIQA 是面向识别的FIQA方法，而 IFQA、DN-PIQA 和 DSL-FIQA 是感知型人脸图像质量评价方法。为确保公平性，除 NIQE 和 SER-FIQ [2] 外，所有方法均在三个测试数据集上重新训练。

C. 训练设置

SFIQA-Assessor 基于 PyTorch 实现。我们采用 EdgeNeXt [77]作为轻量级骨干网络，并探索了三个变体：EdgeNeXt-S、EdgeNeXt-XS、

[2]NIQE 和 SER-FIQ 是无需主观评价（无监督）的评估方法，不需要训练。

以及 EdgeNeXt-XXS，分别记为 SFIQA-Assessor-S、SFIQA-Assessor-XS 和 SFIQA-Assessor-XXS。除非另有说明，默认配置为 SFIQA-Assessor-S。EdgeNeXt 骨干网络由 $N_s = 4$ 个阶段组成，N_s 遵循所选骨干网络的阶段配置。输出特征和潜在特征的通道维度（D_o 和 D_l）分别设为 128 和 64。对于自注意力模块，嵌入维度为 64，包含 4 个注意力头。任务特定自注意力模块和交叉注意力模块的嵌入维度为 128，包含 8 个注意力头。输入的人脸图像将调整为224×224。使用在 ImageNet [82] 上预训练的权重进行初始化，然后在训练过程中针对目标数据集对整个网络进行端到端微调。模型使用 Adam优化器进行训练，初始学习率为 5×10^{-5}，批量大小为 4，并在单块 NVIDIA GeForce RTX 3090 GPU 上运行。

D. 评估标准

我们采用斯皮尔曼秩相关系数（SRCC）和皮尔逊线性相关系数（PLCC）报告所有方法的性能。我们使用三项指标评估计算复杂度：延迟、参数数量（参数）和千兆乘加运算次数（GMACs）。为避免冷启动效应，我们在模型充分预热后测量延迟。具体而言，我们测量模型对 100 张测试图像的推理时间并计算平均值。该测量涵盖算法所需的所有预处理步骤。

E. 性能比较

表 II 汇总了 SFIQA-Assessor 与各对比方法在 SFIQA-Bench 上的结果。从表中可以得出几项重要结论。首先，BRISQUE 和 NIQE 等传统手工设计的图像质量评估方法在 SFIQA-Bench 上表现不佳，这表明基于自然场景统计的特征无法有效量化人脸图像的感知质量。此外，与感知型人脸图像质量评价方法相比，HyperIQA、TReS 和 MANIQA 等通用的基于深度神经网络的图像质量评价方法表现也较差。这主要是因为这些模型并未针对面部特征进行专门优化，因而难以捕捉人脸图像中细微的失真。

其次，面向识别的人脸图像质量评价方法的表现不如感知人脸图像质量评估方法。这凸显了面向识别的质量（侧重于人脸验证或识别置信度）与感知质量（与人类视觉感知相一致）之间的差距。

在所有评估方法中，感知人脸图像质量评估方法 DSL-FIQA 在噪声、锐度、对比度和整体质量方面表现最佳。然而，其实际应用价值受到极高计算成本的限制——参数超过 2 亿，且需要 2,000 GFLOPs——因此不太适合实际部署。相比之下，ourSFIQA-Assessor-Sachieves 的总体表现位居第二，较 DSL-FIQA 的斯皮尔曼等级相关系数仅下降 0.52%。值得注意的是，它实现这一表现所需的参数更少，乘加运算次数减少超过 $600\times$，延迟降低约 $40\times$，凸显了其在效果与效率之间的出色平衡。

在比较 SFIQA-Assessor 的不同变体时，性能下降幅度仍然很小。具体而言，相较于 SFIQA-Assessor-S，SFIQA-Assessor-XS 和 SFIQA-Assessor-XXS 在总体质量维度上的降幅分别仅为 0.37% 和 0.74%。然而，这些降幅换来了显著的效率提升。参数数量分别减少了 53.5% 和 70.3%，GMACs 分别减少了 45.3% 和 62.9%，延迟分别降低了 11.8% 和 22%，从而证实了所提出架构的可扩展性和部署友好性。

表 III 报告了在 GFIQA-20K 和 CGFIQA-40K 上的性能。从结果中可以看出，我们得出的结论与表 II 中的结论相似，进一步证实了所提出方法的有效性和效率。此外，所有方法在 GFIQA-20K 和 CGFIQA-40K 上的性能都显著高于在 SFIQA-Bench 上的性能，这表明 SFIQA-Bench 包含更复杂、更具挑战性的失真。这凸显了该基准作为更困难、更全面的基准，对推动人脸图像质量评估研究所作出的贡献。

F. 跨数据集评估

我们在跨数据集设置下评估两种表现最佳的人脸图像质量评估方法 DSL-FIQA 和我们提出的 SFIQA-Assessor，以考察模型在不同数据集上训练时的泛化能力。实验结果汇总于表 IV，其中在 SFIQA-Bench 上使用总体质量维度进行评估。从表中可以看出，当模型在 GFIQA-20K 或 CGFIQA-40K 上训练并在另一个数据集上评估时，性能始终较高，所有斯皮尔曼等级相关系数（SRCC）均超过 0.90。这表明这两个数据集的图像分布非常相似。相比之下，当模型在 GFIQA-20K 或 CGFIQA-40K 上训练，并在 SFIQA-Bench 上测试时，性能显著下降，这进一步表明 SFIQA-Bench 包含更复杂的失真和更多样化的内容，凸显了我们所提出的基准在评估监控人脸图像质量方面的重要性。有趣的是，在 SFIQA-Bench 上训练并在 GFIQA-20K 或 CGFIQA-40K 上测试时，模型仍能取得中等到较强的性能，这表明在更具挑战性的 SFIQA-Bench 上训练，能够使模型合理地泛化到现有的人脸图像质量评估数据集。最后，值得注意的是，在 SFIQA-Bench 上测试时，我们的 SFIQA-Assessor 大幅优于 DSL-FIQA，验证了我们的模型设计在评估监控人脸图像质量方面的有效性。

G. 消融研究

在本节中，我们进行消融研究，以评估模型各个组成部分的有效性。分析内容包括对比不同的骨干网络和各种特征融合策略，以及评估其他单个模块的贡献。

表 IV 代表性面部图像质量评估方法的跨数据集性能比较。第一行表示训练数据集，第二行表示对应的测试数据集。每组训练–测试数据集均报告斯皮尔曼等级相关系数（SRCC）和 PLCC。对于 SFIQA-Bench，报告的是总体质量维度的结果。

方法	SFIQA-Bench				GFIQA-20K				CGFIQA-40K			
	GFIQA-20K		CGFIQA-40K		SFIQA-Bench		CGFIQA-40K		SFIQA-Bench		GFIQA-20K	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
DSL-FIQA [7]	0.7573	0.7788	0.8338	0.8263	0.5511	0.5601	0.9710	0.9615	0.5626	0.5802	0.9491	0.9392
SFIQA-Assessor-S	0.6666	0.7092	0.7499	0.7709	0.5890	0.6277	0.9767	0.9750	0.6029	0.6228	0.9350	0.9376

表 V 不同骨干网络的性能比较。斯皮尔曼等级相关系数和 PLCC 为六个质量维度上的平均值。

主干网络	SRCC	PLCC	参数量（百万）	GMACs
EdgeNeXt-S [77]	**0.9317**	**0.9510**	17.76	3.75
MobileNetV4-Conv-M [83]	0.9041	0.9306	23.42	5.67
EfficientFormerV2-S1 [84]	0.9150	0.9378	19.05	**2.75**
RepViT-M1.0 [85]	0.9085	0.9352	**16.10**	3.36

表 VI
所提模型中各个模块性能的消融研究。“原始视图”“人脸视图”和“眼嘴视图”分别指原始宽松裁剪图像、紧凑人脸图像和眼睛和嘴部图像。“CFF”表示跨视图特征融合，“TSA”表示任务自注意力。斯皮尔曼等级相关系数和PLCC为六个质量维度上的平均值。

输入视图			融合与解码器		指标			
原始 View	Face View	眼睛-嘴巴 View	CFF	TSA	SRCC	PLCC	参数量（百万）	GMACs
✓	✗	✗	✓	✓	0.9186	0.9374	**6.91**	**1.32**
✓	✓	✗	✓	✓	0.9279	0.9482	12.34	2.54
✓	✗	✓	✓	✓	0.9219	0.9437	12.34	2.54
✓	✓	✓	✓	✗	0.9093	0.9267	17.01	3.74
✓	✓	✓	✗	✓	0.9265	0.9487	16.64	3.15
✓	✓	✓	✓	✓	**0.9317**	**0.9510**	17.76	3.75

1）主干网络：我们评估了多种广泛使用的高效主干网络，包括 MobileNetV4-Conv-M [83],EfficientFormerV2-S1 [83]，和 RepViT-M1.0 [85]，它们的参数量与 EdgeNeXt-S（本文默认使用的主干网络）相近或略大。性能对比见表 V。结果表明，在所有受评估的轻量级主干网络中，EdgeNeXt-S 的平均斯皮尔曼等级相关系数最高，同时计算成本较低。这凸显了 EdgeNeXt 的混合CNN-Transformer架构，以及其采用自适应核尺寸并以线性注意力复杂度高效提取面部质量特征的优势。

2）输入视图：我们在表 VI 中评估了不同输入视图的有效性。紧凑人脸视图和眼部与嘴部视图均有助于提升模型性能。移除其中任一视图都会导致性能下降，这证实了它们的有效性。值得注意的是，舍弃紧凑人脸视图造成的性能下降大于移除眼部与嘴部视图造成的下降，这表明紧凑人脸视图更为关键。尽管如此，眼部与嘴部视图明确提供了细粒度局部特征，与其他视图结合时，可进一步增强模型评估感知质量的能力。

3）融合与解码器：在 SFIQA-Assessor 中，我们引入跨视图特征融合（CFF）模块，以整合来自不同视图的面部特征，并在解码器中引入任务特定的自注意力（TSA）模块，以学习任务感知表征。我们在表 VI 中评估了这两个模块的有效性。结果表明，移除任一模块都会导致性能明显下降，凸显了它们在所提出框架中的重要性。CFF 模块有助于将多视图特征整合为统一表征，而 TSA 模块使模型能够自适应地关注任务相关质量信息。两个模块协同使用时性能最佳，证实了它们在提升特征表征和质量预测方面的有效性。

VI. 结论

本文对监控人脸图像质量评价进行了全面研究，针对现实监控场景带来的独特人脸图像质量评价挑战展开探讨。为此，我们推出了 SFIQA-Bench，这是一个大规模基准，包含多维主观标注，涵盖六个关键维度：噪声、锐度、色彩丰富度、对比度、保真度和整体质量。凭借丰富多样的监控图像内容和全面的评估维度，SFIQA-Bench 可作为在现实监控条件下评估人脸图像质量评价模型的可靠基础。此外，我们提出了 SFIQA-Assessor，这是一种轻量且有效的多任务质量评估模型，利用跨视图特征融合和任务引导注意力，联合预测多个质量维度。大量实验表明，我们的方法优于现有的通用图像质量评估和人脸图像质量评价方法。在研究中，我们观察到，锐度维度往往代表一组混合退化，而非单一、孤立的清晰度相关失真。例如，运动模糊、失焦和低分辨率都可能归入这一维度。未来，我们计划进一步拆解并深入研究这一维度。我们希望这些贡献能够为推进现实监控图像质量分析提供有价值的见解和工具。

参考文献

[1] H. Choung、P. David 和 T.-W. Ling，《监控场景中对人工智能驱动的面部识别技术的接受度：信任、安全与隐私感知的作用》，《社会中的技术》，第79卷，102721，2024年，第。[2] T. Yang、P. Ren、X. Xie 和 L. Zhang，《用于真实环境盲人脸修复的嵌入生成对抗网络先验的网络》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2021年，第672–681页。[3] Z. Yue 和 C. C. Loy，《DifFace：通过扩散误差收缩实现盲人脸修复》，载于*IEEE*《模式分析与机器智能汇刊》，2024年。

[4] S. Zhou、K. Chan、C. Li 和 C. C. Loy，《通过码本查找变换器实现鲁棒盲人脸修复》，载于《神经信息处理系统进展》，第35卷，第30 599–30 611页，2022年。[5] X. Wang、Y. Li、H. Zhang 和 Y. Shan，《利用生成式人脸先验实现真实环境中的盲人脸修复》，载于IEEE/CVF计算机视觉与模式识别会议论文集，2021年，第9168–9178页。[6] T. Liu、S. Li、M. Xu、L. Yang 和 X. Wang，《人脸图像质量评估：大规模数据库与变换器方法》，载于IEEE《模式分析与机器智能汇刊》，第46卷，第5期，第3981–4000页，2024年。[7] W.-T. Chen、G. Krishnan、Q. Gao、S.-Y. Kuo、S. Ma 和 J. Wang，《DSL-FIQA：通过双集合退化学习和关键点引导变换器评估人脸图像质量》，载于IEEE/CVF计算机视觉与模式识别会议论文集，2024年，第2931–2941页。[8] J. Hernandez-Ortega、J. Galbally、J. Fierrez、R. Haraksim 和 L. Beslay，《FaceQnet：基于深度学习的人脸识别质量评估》，载于国际生物特征识别会议，2019年，第1–8页。[9] P. Terhorst、J. N. Kolf、N. Damer、F. Kirchbuchner 和 A. Kuijper，《Ser-FIQ：基于随机嵌入鲁棒性的无监督人脸图像质量估计》，载于IEEE/CVF计算机视觉与模式识别会议论文集，2020年，第5651–5660页。[10] F.-Z. Ou、C. Li、S. Wang 和 S. Kwong，《CLIB-FIQA：通过置信度校准评估人脸图像质量》，载于IEEE/CVF计算机视觉与模式识别会议论文集，2024年，第1694–1704页。[11] H. R. Sheikh、M. F. Sabir 和 A. C. Bovik，《近期全参考图像质量评估算法的统计评估》，载于IEEE《图像处理汇刊》，第15卷，第11期，第3440–3451页，2006年。[12] E. C. Larson 和 D. M. Chandler，《最明显失真：全参考图像质量评估与策略的作用》，载于《电子成像期刊》，第19卷，第1期，第011 006–011 006页，2010年。[13] A. Ciancio、E. A. Da Silva、A. Said、R. Samadani、P. Obrador 等，《基于多特征分类器的数字图像无参考模糊评估》，载于IEEE《图像处理汇刊》，第20卷，第1期，第64–75页，2010年。[14] T. Virtanen、M. Nuutinen、M. Vaahteranoksa、P. Oittinen 和 J. H¨akkinen，《Cid2013：用于评估无参考图像质量评估算法的数据库》，载于IEEE《图像处理汇刊》，第24卷，第1期，第390–402页，2014年。[15] N. Ponomarenko、L. Jin、O. Ieremeiev、V. Lukin、K. Egiazarian、J. Astola、B. Vozel、K. Chehdi、M. Carli、F. Battisti 等，《图像数据库TID2013：特点、结果与展望》，载于《信号处理：图像通信》，第30卷，第57–77页，2015年。[16] D. Ghadiyaram 和 A. C. Bovik，《大规模在线众包主观与客观图像质量研究》，载于IEEE《图像处理汇刊》，第25卷，第1期，第372–387页，2015年。[17] Z. Ying、H. Niu、P. Gupta、D. Mahajan、D. Ghadiyaram 和 A. Bovik，《从图像块到图像（PAQ-2-PIQ）：映射图像质量的感知空间》，载于IEEE/CVF计算机视觉与模式识别会议论文集，2020年，第3575–3585页。[18] H. Lin、V. Hosu 和 D. Saupe，《KADID-10k：大规模人工失真IQA数据库》，载于国际多媒体体验质量会议，2019年，第1–3页。[19] V. Hosu、H. Lin、T. Sziranyi 和 D. Saupe，《KonIQ-10k：用于盲图像质量评估深度学习的生态有效数据库》，载于IEEE《图像处理汇刊》，第29卷，第4041–4056页，2020年。[20] Y. Fang、H. Zhu、Y. Zeng、K. Ma 和 Z. Wang，《智能手机摄影的感知质量评估》，载于IEEE/CVF计算机视觉与模式识别会议论文集，2020年，第3677–3686页。[21] K. Zhang、D. Zhu、X. Min、Z. Gao 和 G. Zhai，《质量与美学的协同评估：方法与综合基准数据集》，载于IEEE《电路与系统视频技术汇刊》，第34卷，第4期，第2536–2549页，2023年。[22] A. Mittal、R. Soundararajan 和 A. C. Bovik，《构建“完全盲”的图像质量分析器》，载于IEEE《信号处理快报》，第20卷，第3期，第209–212页，2012年。[23] A. Mittal、A. K. Moorthy 和 A. C. Bovik，《空间域中的无参考图像质量评估》，载于IEEE《图像处理汇刊》，第21卷，第12期，第4695–4708页，2012年。[24] A. K. Moorthy 和 A. C. Bovik，《盲图像质量评估：从自然场景统计到感知质量》，载于IEEE《图像处理汇刊》，第20卷，第12期，第3350–3364页，2011年。[25] L. Kang、P. Ye、Y. Li 和 D. Doermann，《用于无参考图像质量评估的卷积神经网络》，载于IEEE/CVF计算机视觉与模式识别会议，2014年，第1733–1740页。[26] S. Bosse、D. Maniry、K.-R. M¨uller、T. Wiegand 和 W. 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》，第27卷，第1期，第206–219页，2017年。[27] J. Ke、Q. Wang、Y. Wang、P. Milanfar 和 F. Yang，“Musiq：多尺度图像质量变换器”，载于 IEEE/CVF国际计算机视觉会议论文集，2021年，第5148–5157页。[28] W. Sun、X. Min、D. Tu、S. Ma 和 G. Zhai，“通过分层特征融合和迭代混合数据库训练评估真实场景图像的盲质量”，IEEE信号处理精选主题期刊，2023年。[29] C. Chen、J. Mo、J. Hou、H. Wu、L. Liao、W. Sun、Q. Yan 和 W. Lin，“TOPIQ：一种从语义到失真的自顶向下图像质量评估方法”，IEEE《图像处理汇刊》，第33卷，第2404–2418页，2024年。[30] J. Wang、K. C. Chan 和 C. C. Loy，“探索 CLIP 以评估图像的视觉观感”，载于 AAAI人工智能会议论文集，第37卷，第2期，2023年，第2555–2563页。[31] S. Yang、T. Wu、S. Shi、S. Lao、Y. Gong、M. Cao、J. Wang 和 Y. Yang，“MANIQA：用于无参考图像质量评估的多维注意力网络”，载于 IEEE/CVF计算机视觉与模式识别会议论文集，2022年，第1191–1200页。[32] H. Wu、Z. Zhang、W. Zhang、C. Chen、L. Liao、C. Li、Y. Gao、A. Wang、E. Zhang、W. Sun 等，“Q-align：通过离散文本定义等级训练大型多模态模型进行视觉评分”，载于国际机器学习会议，2024年，第54 015–54 029页。[33] W. Zhang、G. Zhai、Y. Wei、X. Yang 和 K. Ma，“通过视觉—语言对应关系进行盲图像质量评估：多任务学习视角”，载于 IEEE/CVF计算机视觉与模式识别会议论文集，2023年，第14 071–14 081页。[34] P. C. Madhusudana、N. Birkbeck、Y. Wang、B. Adsumilli 和 A. C. Bovik，“使用对比学习进行图像质量评估”，IEEE《图像处理汇刊》，第31卷，第4149–4161页，2022年。[35] Y. Gao、X. Min、Y. Zhu、X.-P. Zhang 和 G. Zhai，“盲图像质量评估：一种用于主观评分分布预测的模糊神经网络”，IEEE《视频技术电路与系统汇刊》，第34卷，第3期，第1641–1655页，2023年。[36] T. Zhou、S. Tan、B. Zhao 和 G. Yue，“结合知识引导注意力的多任务深度神经网络，用于盲图像质量评估”，IEEE《视频技术电路与系统汇刊》，2024年。[37] W. Zhou、Q. Yang、W. Chen、Q. Jiang、G. Zhai 和 W. Lin，“结构引导重采样的稠密三维点云盲质量评估”，ACM多媒体计算、通信与应用汇刊，第20卷，第8期，第1–21页，2024年。[38] S. Sun、T. Yu、J. Xu、W. Zhou 和 Z. Chen，“Graphiqa：学习用于盲图像质量评估的失真图表示”，IEEE《多媒体汇刊》，第25卷，第2912–2925页，2022年。[39] Y. Zhou、Z. Zhang、J. Jia、Y. Jiang、X. Liu、X. Min 和 G. Zhai，“谁是更好的模仿者：动画人物的主观与客观质量评估”，IEEE《视频技术电路与系统汇刊》，2025年。[40] J. Liu、W. Zhou、X. Li、J. Xu 和 Z. Chen，“Liqa：终身盲图像质量评估”，IEEE《多媒体汇刊》，2022年。[41] H. Wang、J. Liu、H. Tan、J. Lou、X. Liu、W. Zhou 和 H. Liu，“基于自适应图注意力的盲图像质量评估”，IEEE《视频技术电路与系统汇刊》，2024年。[42] W. Zhou、J. Xu、Q. Jiang 和 Z. Chen，“通过分析多频信息和局部—全局自然度对360度图像进行无参考质量评估”，IEEE《视频技术电路与系统汇刊》，第32卷，第4期，第1778–1791页，2021年。[43] J. Xu、W. Zhou 和 Z. Chen，“基于视口导向图卷积网络的盲全向图像质量评估”，IEEE《视频技术电路与系统汇刊》，第31卷，第5期，第1724–1737页，2020年。[44] W. Zhou、L. Shi、Z. Chen 和 J. Zhang，“面向张量的无参考光场图像质量评估”，IEEE《图像处理汇刊》，第29卷，第4070–4084页，2020年。[45] W. Zhou、Z. Chen 和 W. Li，“用于无参考立体图像质量评估的双流交互网络”，IEEE《图像处理汇刊》，第28卷，第8期，第3946–3958页，2019年。[46] Z. Chen、W. Zhou 和 W. Li，“立体视频盲质量评估：从深度感知到整体体验”，IEEE《图像处理汇刊》，第27卷，第2期，第721–734页，2017年。

[47] Y. Liu、S. Li、H. Duan、Y. Zhou、D. Fan 和 G. Zhai，“结合特征交互的多任务引导盲全向图像质量评估”，IEEE《视频技术电路与系统汇刊》，2025年。[48] J. Cao、S. Zhang、Y. Liu、F. Gao、K. Gu、G. Zhai、J. Dong 和 S. Kwong，“用于增强图像盲质量评估的多尺度局部与全局特征融合”，IEEE《视频技术电路与系统汇刊》，2025年。[49] C. Tang、X. Yang 和 G. Zhai，“通过统计分析和噪声注入估计自然图像中的噪声”，IEEE《视频技术电路与系统汇刊》，第25卷，第8期，第1283–1294页，2014年。[50] H. Liu、N. Klomp 和 I. Heynderickx，“用于评估图像感知振铃伪影的无参考度量”，IEEE《视频技术电路与系统汇刊》，第20卷，第4期，第529–539页，2009年。[51] N. D. Narvekar 和 L. J. Karam，“基于模糊检测累积概率的无参考图像模糊度量”，IEEE《图像处理汇刊》，第20卷，第9期，第2678–2683页，2011年。[52] B. Sebastian V、A. Unnikrishnan 和 K. Balakrishnan，“灰度共生矩阵：推广及若干新特征”，arXiv预印本，arXiv:1205.4831，2012年。[53] R. Ferzli 和 L. J. Karam，“基于恰可察觉模糊（jnb）概念的无参考客观图像清晰度度量”，IEEE《图像处理汇刊》，第18卷，第4期，第717–728页，2009年。[54] K. Bahrami 和 A. C. Kot，“一种基于最大局部变化的快速无参考图像锐度评估方法”，IEEE《信号处理快报》，第21卷，第6期，第751–755页，2014年。[55] C. T. Vu 和 D. M. Chandler，“S3：频谱与空间清晰度度量”，载于国际多媒体进展会议，2009年，第37–43页。[56] G. Zhai、X. Wu、X. Yang、W. Lin 和 W. Zhang，“基于自由能原理的心理视觉质量度量”，IEEE《图像处理汇刊》，第21卷，第1期，第41–52页，2011年。[57] W. Zhang、K. Ma、J. Yan、D. Deng 和 Z. Wang，“使用深度双线性卷积神经网络进行盲图像质量评估”，IEEE《视频技术电路与系统汇刊》，第30卷，第1期，第36–47页，2020年。[58] Y. Gao、X. Min、Y. Cao、X. Liu 和 G. Zhai，“无参考图像质量评估：根据图像质量分数分布获取平均意见分”，IEEE《视频技术电路与系统汇刊》，2024年。[59] X. Min、Y. Gao、Y. Cao、G. Zhai、W. Zhang、H. Sun 和 C. W. Chen，“探索真实场景图像质量评估中的丰富主观质量信息”，IEEE《视频技术电路与系统汇刊》，2025年。[60] S. Su、H. Lin、V. Hosu、O. Wiedemann、J. Sun、Y. Zhu、H. Liu、Y. Zhang 和 D. Saupe，“深入开展人脸图像质量评估：一种新型数据库和模型”，IEEE《多媒体汇刊》，第26卷，第2671–2685页，2023年。[61] B. Jo、D. Cho、I. K. Park 和 S. Hong，“IFQA：可解释人脸质量评估”，载于 IEEE/CVF冬季计算机视觉应用会议论文集，2023年，第3444–3453页。[62] N. Chahine、S. Calarasanu、D. Garcia-Civiero、T. Cayla、S. Ferradans 和 J. Ponce，“肖像图像质量评价数据集”，载于 IEEE/CVF计算机视觉与模式识别会议论文集，2023年，第9968–9978页。[63] W. Sun、W. Zhang、Y. Jiang、H. Wu、Z. Zhang、J. Jia、Y. Zhou、Z. Ji、X. Min、W. Lin 等，“用于肖像图像质量评价的双分支网络”，arXiv预印本，arXiv:2405.08555，2024年。[64] J. Hernandez-Ortega、J. Galbally、J. Fi´errez 和 L. Beslay，“生物特征质量：综述及其在使用 FaceQnet 进行人脸识别中的应用”，arXiv预印本，arXiv:2006.03298，2020年。[65] F.-Z. Ou、X. Chen、R. Zhang、Y. Huang、S. Li、J. Li、Y. Li、L. Cao 和 Y.-G. Wang，“SDD-FIQA：基于相似度分布距离的无监督人脸图像质量评价”，载于 IEEE/CVF计算机视觉与模式识别会议论文集，2021年，第7670–7679页。[66] A. Radford、J. W. Kim、C. Hallacy、A. Ramesh、G. Goh、S. Agarwal、G. Sastry、A. Askell、P. Mishkin、J. Clark 等，“通过自然语言监督学习可迁移视觉模型”，载于国际机器学习会议，2021年，第8748–8763页。[67] P. J. Phillips、H. Moon、S. A. Rizvi 和 P. J. Rauss，“人脸识别算法的 FERET 评估方法”，IEEE《模式分析与机器智能汇刊》，第22卷，第10期，第1090–1104页，2000年。[68] G. B. Huang、M. Mattar、T. Berg 和 E. Learned-Miller，“野外标注人脸：用于研究非受限环境下人脸识别的数据库”，载于“真实场景”人脸图像研讨会：检测、对齐

，与识别 ，2008年。

[69] D. Yi、Z. Lei、S. Liao 和 S. Z. Li，《从头开始学习人脸表征》，*arXiv*预印本 *arXiv:1411.7923*，2014年。[70] Q. Cao、L. Shen、W. Xie、O. M. Parkhi 和 A. Zisserman，《Vggface2：一个用于识别不同姿态和年龄人脸的数据集》，载于*IEEE*自动人脸与手势识别国际会议，2018年，第67–74页。[71] P. ITU-T RECOMMENDATION，《多媒体应用的视频质量主观评估方法》，1999年。[72] R. BT，《电视图像质量主观评估方法》，国际电信联盟，第4卷，第19页，2002年。[73] ˚A. Bj¨orck，最小二乘问题的数值方法。SIAM，2024年。[74] J. C. Faure、C. Rieffe 和 J. C. Maltha，《不同面部组成部分对面部美学的影响》，《欧洲正畸学杂志》，第24卷，第1期，第1–7页，2002年。[75] Linzaer，《Ultra-light-fast-generic-face-detector-1mb》，https://github.com/Linzaer/Ultra-Light-Fast-Generic-Face-Detector-1MB，2019年。[76] K. Narayan、V. Vs 和 V. M. Patel，《Segface：长尾类别的人脸分割》，载于*AAAI*人工智能会议论文集，第39卷，第6期，2025年，第6182–6190页。[77] M. Maaz、A. Shaker、H. Cholakkal、S. Khan、S. W. Zamir、R. M. Anwer 和 F. Shahbaz Khan，《EdgeNext：面向移动视觉应用的高效融合CNN-变换器架构》，载于欧洲计算机视觉会议。施普林格，2022年，第3–20页。[78] M. D. Zeiler 和 R. Fergus，《卷积网络的可视化与理解》，载于欧洲计算机视觉会议，2014年，第818–833页。[79] R. Zhang、P. Isola、A. A. Efros、E. Shechtman 和 O. Wang，《深层特征作为感知度量的惊人效能》，载于*IEEE*计算机视觉与模式识别会议论文集，2018年，第586–595页。[80] S. Su、Q. Yan、Y. Zhu、C. Zhang、X. Ge、J. Sun 和 Y. Zhang，《在自适应超网络的引导下对真实场景图像质量进行盲评估》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2020年，第3667–3676页。[81] S. A. Golestaneh、S. Dadsetan 和 K. M. Kitani，《通过变换器、相对排序和自一致性进行无参考图像质量评估》，载于*IEEE/CVF*冬季计算机视觉应用会议论文集，2022年，第3209–3218页。[82] J. Deng、W. Dong、R. Socher、L.-J. Li、K. Li 和 L. Fei-Fei，《ImageNet：大型层次化图像数据库》，载于*2009*年*IEEE*计算机视觉与模式识别会议。IEEE，2009年，第248–255页。[83] D. Qin、C. Leichner、M. Delakis、M. Fornoni、S. Luo、F. Yang、W. Wang、C. Banbury、C. Ye、B. Akin 等，《MobileNetV4：移动生态系统的通用模型》，载于欧洲计算机视觉会议，2024年，第78–96页。[84] Y. Li、G. Yuan、Y. Wen、J. Hu、G. Evangelidis、S. Tulyakov、Y. Wang 和 J. Ren，《EfficientFormer：以MobileNet的速度运行的视觉变换器》，神经信息处理系统进展，第35卷，第12 934–12 949页，2022年。[85] A. Wang、H. Chen、Z. Lin、J. Han 和 G. Ding，《RepViT：从ViT视角重新审视移动端CNN》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2024年，第15 909–15 920页。



Yanwei Jiang 于2022年获得上海交通大学工学学士学位，目前正在图像通信与信息处理研究所攻读博士学位。他的研究兴趣包括图像质量评估和多媒体信号处理。



Wei Sun 于2016年获得华东理工大学电子与信息工程工学学士学位，于2023年获得中国上海上海交通大学博士学位。2023年至2025年，他在上海交通大学从事博士后研究。目前，他是华东师范大学的研究科学家。他的研究兴趣包括图像质量评估、感知信号处理和移动视频处理。



Yingjie Zhou 于2023年获得中国矿业大学电子与信息工程工学学士学位。目前，他正在中国上海交通大学图像通信与网络工程研究所攻读博士学位。他目前的研究兴趣包括数字人质量评估和情感分析。



Xiangyang Zhu 于2017年获得中国湖北华中科技大学工学学士学位，并于2024年获得中国香港特别行政区香港城市大学博士学位。自2024年起，他在上海人工智能实验室担任博士后。他的研究兴趣包括多模态表示学习以及大语言模型/多模态大语言模型评估。



Yuqin Cao 于2021年获得中国上海交通大学工学学士学位。她目前正在中国上海交通大学图像通信与网络工程研究所攻读博士学位。她目前的研究兴趣是音视频质量评估。



Jun Jia 于2018年获得中国长沙湖南大学计算机科学与技术专业学士学位，并于2024年获得中国上海交通大学电子工程系博士学位。他的研究兴趣包括计算机视觉和图像处理。



Yunhao Li 于2019年获得北京航空航天大学（中国北京）工学学士学位。他目前正在上海交通大学图像通信与网络工程研究所（中国上海）攻读博士学位。他的研究兴趣包括质量评估、以人为中心的理解与生成。



Sijing Wu 于2020年获得西安交通大学（中国陕西）工学学士学位。她目前正在上海交通大学图像通信与网络工程研究所（中国上海）攻读博士学位。她的研究兴趣包括以人为中心的计算机视觉、三维头部生成与动画。



Dandan Zhu （IEEE会员）于2019年获得中国上海同济大学博士学位。2019年至2021年，她在上海交通大学教育部人工智能重点实验室担任博士后研究员。目前，她是华东师范大学计算机科学与技术学院副教授。她的研究兴趣包括多媒体信号处理、计算机视觉和视觉注意力建模。



Xiongkuo Min 于2013年获得中国武汉大学工学学士学位，并于2018年获得中国上海交通大学博士学位。他目前是上海交通大学图像通信与网络工程研究所的长聘轨副教授。他的研究兴趣包括图像/视频/音频质量评估、体验质量、视觉注意力建模、扩展现实和多模态信号处理。



Guangtao Zhai（IEEE会士，2024年）分别于2001年和2004年获得山东大学（中国山东）工学学士学位和工学硕士学位，并于2009年获得上海交通大学（中国上海）博士学位，目前担任该校图像通信与信息处理研究所研究教授。2008年至2009年，他曾是麦克马斯特大学（加拿大安大略省汉密尔顿）电气与计算机工程系的访问学生，并于该校担任博士后研究员，自2010年至2012年。2012年至2013年，他在德国埃尔朗根-纽伦堡弗里德里希·亚历山大大学多媒体通信与信号处理研究所担任洪堡研究学者。他的研究兴趣包括多媒体信号处理和感知信号处理。