

LAR-IQA：一种轻量、准确且鲁棒的无参考图像质量评价模型

LAR-IQA: A Lightweight, Accurate, and Robust No-Reference Image Quality Assessment Model

Nasim Jamshidi Avanaki¹, Abhijay Ghildyal¹, Nabajeet Barman², and
Saman Zadtootaghaj²

¹ Portland State University, OR, USA

² Sony Interactive Entertainment, UK/Germany

n.jamshidi.avanaki@gmail.com, abhijay@pdx.edu

{Nabajeet.Barman,Saman.Zadtootaghaj}@sony.com

摘要 近年来，基于深度学习的无参考图像质量评价（No-Reference Image Quality Assessment, NR-IQA）在多个开源数据集上取得了很高的性能。然而，这类模型通常体量庞大、结构复杂，难以在真实场景中部署，在资源和电量都受限的移动设备上尤其如此。为突破这一局限，我们提出一个结构紧凑的轻量 NR-IQA 模型：它在 ECCV AIM UHD-IQA 挑战赛的验证集和测试集上达到当前最优（state-of-the-art, SOTA）性能，同时比最快的 SOTA 模型还快近 5.7 倍。该模型采用双分支架构，两个分支分别在合成失真图像和真实失真图像上单独训练，从而增强模型对不同失真类型的泛化能力。为提升模型在多样的真实视觉条件下的鲁棒性，我们在训练过程中进一步引入多种色彩空间。我们还表明，用于最终质量回归时，新近提出的 Kolmogorov-Arnold 网络（KAN）比传统的多层感知机（MLP）更准确。在多个开源数据集上的评测凸显了所提轻量模型实用、高精度且鲁棒的表现。代码：<https://github.com/nasimjamshidi/LAR-IQA>。

Keywords: 质量评价 · IQA · 无参考 IQA · BIQA · 实时 · 轻量

1 引言

图像质量评价 (Image Quality Assessment, IQA) 任务要衡量人眼所感知的图像质量, 它至今仍是计算机视觉中最有意思也最具挑战性的方向之一。无参考 IQA (NR-IQA), 又称盲 IQA (BIQA), 关注的是在没有高质量参考图像可供比对的情况下, 估计退化图像的质量。图像中可能出现的失真种类繁多 (压缩、模糊、噪声等), 这让该任务尤为困难。NR-IQA 在摄影、监控、医疗、汽车、社交媒体、用户生成内容平台等诸多行业和应用中都举足轻重。每天都有数以百万计的用户生成内容 (user-generated content, UGC) 图像被上传, 并在 Instagram、X、Flickr 等大量社交平台上传播。在真实场景中, 用户拍摄的图像可能带有模糊、噪声 (来自相机传感器)、色彩失真、压缩伪影 (块效应), 或这些问题的组合。自动检测低质量或不合适的图像, 并据此指导必要的前处理与后处理 (质量增强、压缩系数选择、去模糊等), 对提升用户体验乃至这些公司的成败都至关重要。

NR-IQA 的一大挑战, 是设计出更小、更快、更高效、适合实时质量评价的方法。传统 NR-IQA 模型通常速度快、复杂度低, 但精度往往不足 [20, 27–30, 33, 44, 45, 47, 49]。另一方面, 传统的基于深度神经网络 (DNN) 的模型精度更高, 却普遍复杂度更高、计算开销大 [3, 5, 11, 35, 39, 46, 48]。近来基于大型多模态模型的 IQA 方法, 在视觉和文本编码器上都采用了 Transformer 等复杂架构 [43, 51], 导致模型体量和复杂度都很大, 难以胜任大多数实时评价任务。

一些更新的 NR-IQA 方法以 Transformer 作为骨干网络 [11, 46, 48], 因为已有研究表明 Transformer 能为 NR-IQA 提供更好的特征, 从而带来更准确、更鲁棒的结果。然而, 基于 Transformer 的模型参数量庞大, 训练和推理都需要可观的计算资源, 这限制了它们在低功耗设备上的部署。近来, 视觉语言模型 (Vision Language Model, VLM), 尤其是 CLIP [32], 在 NR-IQA 上取得了显著成功。经 NR-IQA 任务微调后 [2, 12, 39], CLIP 相比已有方法展现出良好的精度、鲁棒性和泛化能力, 能有效捕捉大规模数据集中形形色色的失真。不过, 我们并不采用基于 VLM 的预训练网络, 而是着力于打造一个更省算力的准确且鲁棒的模型——具体而言, 以输入图像尺寸为 3840×2160 像素时一次前向传播所需的乘累加 (Multiply-Accumulate, MAC) 运算次数来衡量算力。

1.1 贡献

本文围绕低复杂度、泛化能力和鲁棒性做出若干贡献, 下面逐一展开。

更低的复杂度。DNN 类方法能够捕捉复杂的模式和非线性失真，从而获得更高的精度和鲁棒性；有鉴于此，我们选用轻量的 MobileNetV3 [16] 作为基线网络。与其他基于 DNN 的 NR-IQA 方法 [2, 3, 35, 39] 相比，我们的网络参数更少，复杂度更低。

泛化能力。训练一个可泛化的 NR-IQA 模型，主要难点之一在于合成数据集缺乏真实场景中那种多样的失真。已有若干方法尝试用预训练、自监督和新型损失函数来弥合这一差距 [2, 3, 11, 39]。用对比损失做自监督学习、或在无标注数据上预训练 [3, 24]，对处理合成与真实两类退化都行之有效。而我们应对数据集不够全面这一问题的做法，是训练两个独立的模型——一个在合成失真数据集上训练，另一个在真实失真数据集上训练，再把两个单独训练好的模型组合起来，得到一个泛化性更强的模型。

这类多分支架构此前已有人提出。例如，[12] 使用了语义、美学、技术三个分支，再把它们的预测分数融合，得到最终质量分。相比之下，我们是首个提出双分支架构、专门应对合成与真实两类不同失真的工作。考虑到实时应用的需求，我们只采用 MobileNet [16, 22, 38, 50] 作为模型的骨干图像编码器，力求让 NR-IQA 模型既轻量又快速。

鲁棒性与精度。为提升轻量模型的鲁棒性和精度，我们综合采用几项策略：双分支分别训练（分别对应合成失真和真实失真）、以 KAN [21] 作为回归头、以及使用色彩空间损失函数。

总之，本文提出了一种面向低复杂度 NR-IQA 的新方法和新模型，同时利用真实与合成失真来实现鲁棒且准确的评价。我们的主要贡献如下：

- 我们提出将真实与合成两个 IQA 分支相结合，各自在不同的数据集上训练，以增强模型在各类失真上的鲁棒性和泛化能力。
- 我们比较了为合成分支和真实分支量身选择的的不同骨干架构，为最优设计选型提供依据。
- 我们做了一项消融实验，就质量回归模块对比 KAN 与 MLP。据我们所知，这是首个在该场景下考察并比较 KAN 与 MLP 效果的研究。
- 我们提出一种新的损失函数，针对不同色彩空间之间的差异，以提升 IQA 模型的鲁棒性。

我们的方法得到一个轻量模型，在 ECCV AIM UHD-IQA 挑战赛³ 的验证集和测试集上超越了当前最优性能 [13]，同时保持了实际应用所需的高效率。

³ <https://codalab.lisn.upsaclay.fr/competitions/19335>

本文其余部分组织如下：第 2 节讨论无参考图像质量评价、以及合成与真实数据融合的相关前期工作；第 3 节详细描述我们的模型架构，并对不同骨干、回归头和损失函数做对比分析；第 4 节给出消融实验结果，以及所提模型与 SOTA 的对比评测；最后，第 5 节总结全文，强调本研究发现的意义。

2 相关工作

2.1 无参考图像质量评价

多年来，学界提出了许多 NR-IQA 模型：从最初基于自然场景统计（Natural Scene Statistics, NSS）和手工特征的传统方法，到基于深度学习的方法，再到近来基于视觉语言模型（VLM）的模型。传统 NR-IQA 方法依赖于对特定失真的检测和度量，如块效应、模糊、条带、噪声等。这类方法利用手工设计的自然场景统计（NSS）特征来评价图像质量 [20, 27–30, 33, 40, 44, 45, 47, 49]。然而，这些针对特定失真、基于手工特征的模型，对含有多种失真类型的图像泛化能力不佳。

近来，受深度神经网络（DNN）在计算机视觉中成功的推动，学界提出了多种无需显式设计特征、泛化能力更好的 NR-IQA 方法。基于 DNN 的方法通过学习复杂模式并兼顾图像中的多种失真，在各类数据集上都取得了很高的精度 [3, 5, 46, 48]。然而，受限于可用数据集规模有限，它们往往在训练数据上过拟合，难以泛化到（训练数据中未出现的）新型失真；这一点在做跨数据集评价时表现得很明显，正如 [11] 所讨论的。

更近一些，视觉语言模型（VLM）在 IQA 领域也取得了成功，它结合视觉与文本信息来评价图像质量 [39]。由于 VLM 在融合视觉与语言两种模态的超大规模数据集上训练过，它们能借助上下文对图像内容有更细腻的理解。语义信息与文本描述/字幕相结合，使它们能更好地理解质量，其判断往往与人类评分相近。类似地，面向 IQA 任务也发展出了大型多模态模型（MLLM）[43, 51]，其精度高于传统的基于 DNN 的方法。不过，复杂的训练配置和庞大的规模，使基于 VLM 和 MLLM 的模型计算开销高昂。

2.2 跨多样数据集的泛化

训练大型视觉 Transformer（ViT）或 CNN 骨干网络，需要大量数据做预训练，才能保证对新数据集的泛化。有些方法用对比学习做预训练 [3, 24]，另一些则尝试把多个带有主观评分的图像数据集合并起来 [26]。然而，由于每个数据集内部

固有的主观偏差，合并数据集颇具挑战。正如 [26] 所述，这些偏差包括评分噪声、主观测试的顺序效应、各数据集质量分数分布的差异，以及主观实验中的长期依赖。此外还有一些偏差——尤其在使用众包时——源于不够受控的环境，例如显示器观看距离、显示尺寸和设置（如伽马和亮度）的差异。这些偏差使得直接合并来自不同测试的主观评分变得复杂。

为应对这些挑战，学界提出了各种替代方法，例如把主观偏差纳入损失函数 [26]，或改用排序损失代替均方误差（MSE）损失——其依据是假设主观偏差不会影响同一数据集内 MOS 的排序 [11]。而我们的做法是，用多任务训练的方式，让模型的每个分支在多个真实和合成数据集上训练。对每个分支而言，每个数据集都被当作一个带有独立回归头的单独任务。这样每个分支就能有效学到各个数据集各自的主观偏差，使图像编码器获得更好的泛化。在最终模型中，我们去掉这些独立的回归头，转而两个分支各配一个 KAN 头作为下采样器，再用一个更大的 KAN 头把得到的嵌入融合起来。

2.3 Kolmogorov-Arnold 网络 (KAN)

以往通常用全连接的 MLP 作为回归头 [7, 12]，而近来 KAN [21] 作为 MLP 的一种替代被提了出来。MLP 是把输入与权重相乘，KAN 则不同，它让输入经过可学习的 B 样条函数处理 [5]。在中间层，MLP 使用固定的激活函数，即 $\sigma(w \cdot x + b)$ 中的 σ ；KAN 则采用基于参数化 B 样条函数的可学习激活函数，形式为 $\phi(x) = \text{silu}(x) + \sum_i c_i B_i(x)$ ，其中 B 是基函数， c 是可训练的控制点参数。在这一组合中，不可训练的 $\text{silu}(x)$ 分量与可训练的样条分量相加，使 $\phi(x)$ 成为一个残差式激活函数。因此，KAN 由分布在网络各条边上的、基于样条的一元函数构成，这些函数被设计成可学习的激活函数。这些特性同时提升了网络的精度和可解释性。在我们的工作中，用 KAN 替换 MLP 头同样提升了模型精度，详见后文第 4.2 节。

3 所提方法

本文的目标是开发一个轻量模型，同时兼顾乘累加运算（MAC）次数和参数量。为此，我们在图像编码器中采用移动端架构 [22]。我们提出一个双分支架构，由真实分支和合成分支组成，如图 1 所示。每个分支都包含一个预处理模块和一个基于移动端的图像编码器。我们用一个 KAN 质量回归模块来融合两个分支的特征并预测最终质量分。起初，两个分支被当作独立的模型，分别在真实和合

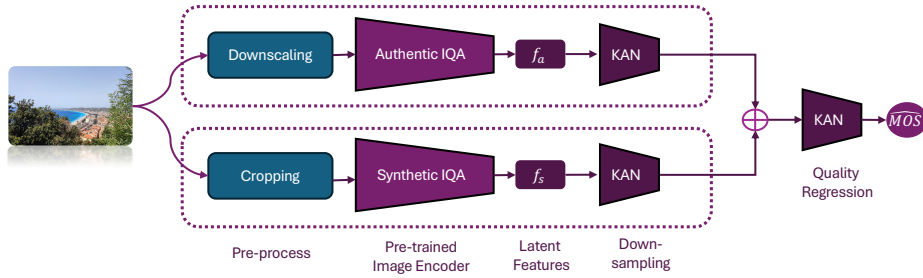


图 1: 所提模型架构。图像质量由两个分支评价: 真实 (Authentic) 分支和合成 (Synthetic) 分支。两个分支都以 MobileNetV3 [16] 作为轻量图像编码器。从两个分支提取的特征拼接后, 作为质量回归模块 KAN 的输入, 由其输出最终预测的图像质量分数。

表 1: 本工作所用 IQA 数据集概况。

数据库	源图像数量	失真类型	失真图像数量	失真类型数量	图像分辨率
KONIQ-10K [15]	10,073	真实	10,073	—	1024×768px
SPAQ [9]	11,125	真实	11,125	—	可变
BID [8]	—	真实	585	—	1280×960px to 2272×1704px
UHD-IQA [13]	6,073	真实	—	—	多为 3840×2160px
KADID-10k [19]	81	合成	10,125	25	512×384px
TID2013 [31]	25	合成	3,000	24	512×384px
PIPAL [18]	250	合成 + 算法	29,000	40	可变

成数据集上训练, 每个分支各用一个 KAN 回归头独立预测图像质量。在各自数据集上训练完成后, 真实分支和合成分支的图像编码器被整合进双分支模型。

3.1 数据集与评价指标

本工作使用了七个公开的 IQA 数据集: 两个含合成失真 (KADID-10K [19]、TID2013 [31]), 四个含真实失真 (KonIQ-10k [15]、SPAQ [9]、BID [8]、UHD-IQA [13]), 以及同时含合成与算法失真的 PIPAL [18]。实验所用数据集的概况见表 1。

为评价模型性能, 我们采用三种常用准则: 衡量预测单调性的斯皮尔曼秩相关系数 (SRCC), 以及衡量排序一致性的皮尔逊线性相关系数 (PLCC) 和肯德尔秩相关系数 (KRCC)。

3.2 数据预处理

在训练阶段，我们对模型的每个分支采用了不同的预处理方法。真实分支中，图像在预训练时被缩小到 224×224 像素、在 UHD-IQA 挑战赛中缩小到 384×384 ；而合成分支中，图像保持原始尺寸，只做裁剪操作。真实分支之所以缩小图像，是因为真实数据集的质量与其内容本身（如拍摄的物体、图像构图）密切相关，改变图像尺寸并不影响感知质量 [36, 42]。然而在合成分支中，缩小图像会造成高频细节丢失，进而损害合成失真图像的质量预测。因此，我们只在真实分支中缩小图像。对于合成分支，我们用 224 尺寸的裁剪图像来训练模型。

3.3 质量回归模块

对于质量回归头，我们同时考察 MLP 和 KAN。KAN 采用 [1] 提供的实现。为保证 MLP 与 KAN 的比较公平，我们做了两组对比。第一组中，两个模型采用完全相同的架构，由两个全连接（FC）层构成：第一层把 1000 维特征降到 128 维，第二层再把 128 维降到 1 维。考虑到 KAN 的参数量和 FLOPs 更高，我们还调整了 MLP 的复杂度以大致匹配 KAN 的 FLOPs——在 128 神经元层之前再加一层 1125 个神经元、采用 ReLU 激活的层。评价在两种场景下进行：数据集内评价和跨数据集评价，更多细节在第 4.2 节讨论。

3.4 图像编码器模块

对于图像编码器模块，我们考察了两种轻量的 CNN 架构和两种 ViT 架构。正如 [22] 所示，MobileNetV2 [34] 和 MobileNetV3 [16] 参数更少、计算复杂度低，是移动端 CNN 网络的绝佳选择。此外，我们还考虑了新近的移动端 ViT 模型 MobileCLIP-S2 [38]，以及以轻量著称的 MobileViT-S [25]。需要说明的是，尽管 MobileCLIP-S2 是一个 VLM，但在第 4.3 节中我们只使用它的视觉编码器架构，我们的模型并不采用 VLM 的训练方式。

3.5 损失函数

MSE 与 PLCC 损失 ($\mathcal{L}_{acc}$)。基于距离的损失（如 MAE 和 MSE）最小化预测分数与真值标签之间的绝对差异，保证分数预测的准确性。相比之下，基于相关性的损失（如 PLCC）保持图像质量的相对相关关系，与人类感知更契合 [6]。在本工作的所有实验中，我们把基于距离的（MSE）和基于相关性的（PLCC）两种目标组合成一个损失函数，以提升模型精度。

$$\mathcal{L}_{acc} = \alpha \cdot MSE + \beta \cdot PLCC \quad (1)$$

色彩空间鲁棒性 ($\mathcal{L}_{rob}$)。正如传统的基于 NSS 的模型所展示的 [4, 10, 23, 37], 在不同色彩空间中分析图像, 能更深入地揭示影响图像质量的感知属性, 从而提升视觉质量评价的准确性。多色彩空间方法能捕捉单一色彩空间可能覆盖不到的图像质量的多个方面。因此, 我们提出一种色彩空间损失, 在 RGB、YUV 和 LAB 色彩空间中评价图像质量, 从而获得更准确、更鲁棒的评价。我们的色彩空间损失 ($\mathcal{L}_{rob}$) 定义为

$$\widehat{MOS}_{\text{color-space}} = F_{\phi}(\mathbf{I}_{\text{color-space}}) \quad (2)$$

$$\sum_{\text{color-space}} \frac{1}{N} (\widehat{MOS}_{\text{color-space}} - MOS)^2 \quad (3)$$

其中 $\text{color-space} \in \{\text{RGB}, \text{YUV}, \text{LAB}\}$, $\mathbf{I}_{\text{color-space}}$ 表示指定色彩空间下的输入图像。我们用参数为 ϕ 的训练好的模型 F , 为 $\mathbf{I}_{\text{color-space}}$ 预测 MOS, 记作 $\widehat{MOS}_{\text{color-space}}$ 。计算损失时, 我们求它与真值 MOS 之间的均方误差。通过最小化这个色彩空间损失, 模型学会在各种色彩空间表示之间对齐其输出特征。这一做法使模型无论使用哪种色彩空间, 都能一致地预测图像质量, 提升了模型对不同失真的敏感度, 并确保评价与人类视觉感知高度吻合 (见第 4.4 节)。

4 评价方法与结果

第 4.1 节先讨论模型的实现细节。第 4.2 节给出合成模型和真实模型的初步结果, 并探索用 KAN 替换 MLP 回归头。第 4.3 节在固定回归头为 KAN 的前提下, 为合成模型和真实模型试验不同的骨干网络作为图像编码器。为增强鲁棒性, 第 4.4 节考察多种损失函数。最后, 第 4.5 节综合上述实验的结论, 讨论合成分支与真实分支的配置与整合。第 4.6 节则从精度指标 (SRCC、PLCC、KRCC、RMSE、MAE) 和计算复杂度 (MAC) 两方面, 将最终模型与当前最优方法作比较。

4.1 实现细节

模型的训练和测试在 PyTorch 上、单张 Nvidia A100 GPU 上完成。模型使用 AdamW 优化器训练 100 个 epoch, 初始学习率为 5×10^{-5} , 权重衰减为 1×10^{-4} 。我们使用自定义调度器, 先线性预热, 再进入余弦退火调度。输入图像尺寸因分支而异: 合成分支用原始尺寸, 真实分支缩放到 224×224 像素。

表 2: MLP 与 KAN 两种回归头的数据集内评价结果。合成模型在 KADID-10K 和 PIPAL 数据集上训练, 并用 10 折交叉验证评价。结果表明, 在两个数据集上 KAN 均优于 MLP。

数据集	KADID-10K			PIPAL		
	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC
MLP	0.961	0.946	0.825	0.868	0.845	0.680
KAN	0.965	0.941	0.857	0.887	0.860	0.692

表 3: MLP 与 KAN 作为回归头的跨数据集评价结果。合成模型在 KADID-10K 上训练、在 TID2013 上测试; 同样地, 真实模型在 KONIQ-10K 上训练、在 BID 上测试。结果表明 KAN 显著优于 MLP。

模型	回归头			PLCC	SRCC	KRCC
	MLP	KAN	FLOPs			
合成				KADID-10K / TID2013		
	✓		0.26M	0.64	0.59	0.45
	✓		2.60M	0.67	0.63	0.48
		✓	2.31M	0.71	0.66	0.50
真实				KONIQ-10K / BID		
	✓		0.26M	0.716	0.673	0.495
	✓		2.60M	0.726	0.673	0.501
		✓	2.31M	0.736	0.680	0.505

对于 UHD-IQA 数据集, 由于图像分辨率很高, 真实分支缩放到 384×384 , 合成分支则采用 1280×1280 的中心裁剪。最终所提模型的图像编码器模块采用 MobileNetV3, 两个分支保持一致。质量回归模块方面, 我们引入 KAN 模型, 分别用于降低特征维度和预测图像质量。

4.2 质量回归模块: MLP 对比 KAN

在本实验中, 我们用 MobileNetV2 训练, 并采用标准损失函数——即在一个 batch 内图像之间计算的 MSE 与 PLCC 损失的加权组合。学习率等其他设置保持不变。数据集内评价时, KADID-10K 和 PIPAL 各自单独用于训练和测试,

表 4: 不同图像编码器的性能比较。在所考察的编码器中，MobileNetV3 在所有指标上都稳定优于其他编码器。

模型	骨干	参数量	PLCC	SRCC	KRCC
KADID-10K / TID2013					
合成	MobileViT-S	5.6M	0.61	0.59	0.42
	MobileCLIP-S2 [38]	35.5M	0.69	0.65	0.46
	MobileNetV2 [34]	3.5M	0.71	0.66	0.50
	MobileNetV3 [16]	7.1M	0.72	0.68	0.50
KONIQ-10K / BID					
真实	MobileViT-S	5.6M	0.71	0.65	0.47
	MobileCLIP-S2	35.5M	0.78	0.75	0.57
	MobileNetV2	3.5M	0.74	0.68	0.51
	MobileNetV3	7.1M	0.79	0.78	0.60

采用 10 折交叉验证，性能指标按 10 折交叉验证结果的平均值报告。跨数据集评价时，我们使用两组训练/测试对：合成模型用 KADID-10K / TID2013，真实模型用 KONIQ-10K / BID。KADID-10K 和 PIPAL 上的数据集内评价结果见表 2。这些结果表明，用 KAN 作回归头比 MLP 有所提升。如前所述，我们用了两种 MLP 回归头：一种架构相同，另一种 FLOPs 相同。第二个实验中，我们列出了每个 MLP 模型的 FLOPs，并与 KAN 对比。表 3 给出跨数据集评价结果，显示 KAN 相比 MLP 有所提升，凸显了 KAN 更强的泛化能力。因此，在后续实验中我们只用 KAN 来完成质量回归任务。需要说明的是，公平比较 KAN 与 MLP 颇具难度：网络拓扑和激活函数选择的差异会显著影响 MLP 的性能；同样地，对 KAN 来说，样条的阶数、样条区间数等因素也会影响结果。尽管如此，本文表明，将 KAN 作为回归头的一部分，其效果可以与使用 MLP 头相当。

4.3 图像编码器的比较

以 KAN 作为回归头，我们现在比较作为模型骨干网络的不同图像编码器。与前面的实验类似，合成模型在 KADID-10K 上训练、在 TID2013 上测试，真实模型在 KONIQ-10K 上训练、在 BID 上测试。PLCC、SRCC、KRCC 等性能指标详见表 4。结果表明，两种基于 CNN 的模型都优于移动端 ViT 模型。

表 5: 在不同色彩空间下, 各种损失函数的性能比较。性能最佳的模型以粗体标出。

模型	损失		骨干	RGB			YUV			LAB		
	$\mathcal{L}_{acc}$	$\mathcal{L}_{rob}$		PLCC	SRCC	KRCC	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC
KADID-10K / TID2013												
合成	✓		Mobile-	0.72	0.68	0.50	0.41	0.34	0.25	0.47	0.39	0.28
	✓	✓	NetV3	0.73	0.70	0.53	0.49	0.47	0.35	0.55	0.52	0.37
合成	✓		Mobile-	0.69	0.65	0.46	0.44	0.38	0.29	0.46	0.40	0.29
	✓	✓	CLIP-S2	0.69	0.66	0.48	0.58	0.54	0.39	0.59	0.55	0.40
KONIQ-10K / BID												
真实	✓		Mobile-	0.79	0.78	0.60	0.59	0.58	0.42	0.59	0.57	0.42
	✓	✓	NetV3	0.82	0.79	0.61	0.69	0.68	0.51	0.76	0.74	0.56
真实	✓		Mobile-	0.78	0.74	0.56	0.65	0.62	0.46	0.67	0.62	0.46
	✓	✓	CLIP-S2	0.78	0.77	0.58	0.74	0.72	0.53	0.74	0.73	0.54

MobileCLIP-S2 性能下降可归因于训练数据有限, 因为基于 ViT 的模型通常需要大数据集才能有效学习其庞大的网络参数。不过, 正如下一节所报告的, 我们发现使用多种色彩空间时 MobileCLIP-S2 有助于获得更好的泛化。不出所料, MobileNetV3 的表现优于 MobileNetV2; 因此, 我们在后续模型设计中采用 MobileNetV3。

4.4 各种损失函数的比较

表 5 给出了在不同色彩空间 (RGB、YUV、LAB) 下各种损失函数的性能。结果表明, 同时使用 $MSE+PLCC$ 损失 ($\mathcal{L}_{acc}$) 和色彩空间损失 ($\mathcal{L}_{rob}$) 能提升模型性能。虽然在 RGB 色彩空间下提升不明显, 但在 YUV 和 LAB 色彩空间下, 各项性能指标都有大幅提升。采用 ViT 架构的 MobileCLIP-S2 [38] 更为鲁棒, 在不同色彩空间下都有更好的表现。这或许可归因于 ViT 模型的特性——它在处理过程中会对像素值做多次归一化。

图 2 展示了采用 MobileNetV3 的合成模型的 MOS 预测, 该模型在 KADID-10K 上训练、在 TID2013 数据集的样例图像上测试。尽管两个数据集在失真的尺度和类型上存在差异, 导致预测 MOS 与真值之间有出入, 但用色彩空间损失 ($\mathcal{L}_{rob}$) 训练能增强鲁棒性, 通过确保结果在不同色彩空间下保持一致, 帮助模型获得更好的泛化。

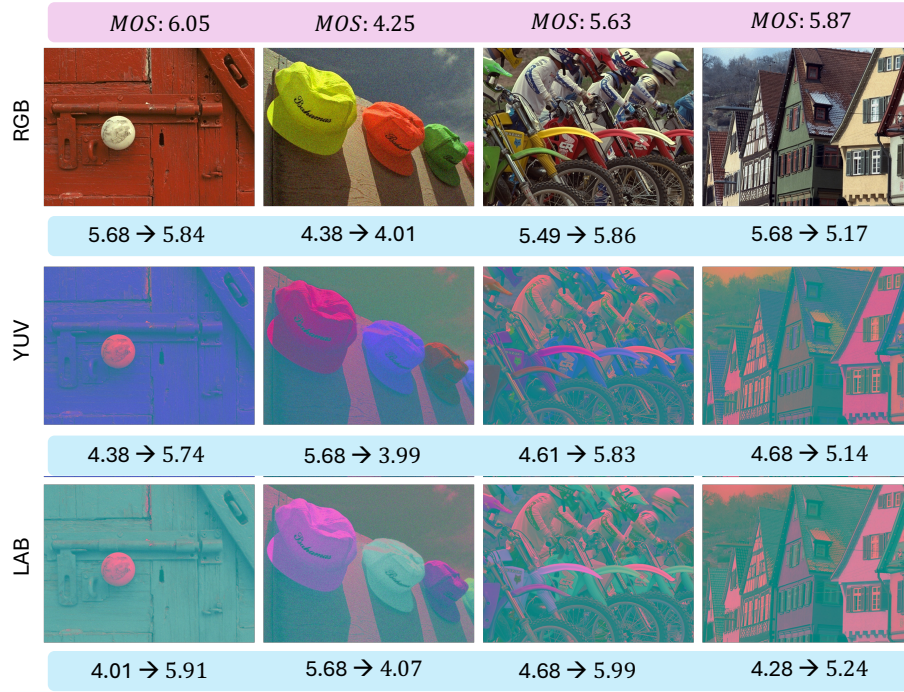


图2: 使用 MobileNetV3 的合成模型的 MOS 预测: 该模型在 KADID-10K 上训练, 并在 TID2013 数据集的样例图像上、跨不同色彩空间测试。真实 MOS 分数上方粉色条突出显示。在使用色彩空间损失 ($\mathcal{L}_{rob}$) 预训练前后、跨各色彩空间预测 MOS 分数的变化, 以蓝色条显示。可以注意到, 各图像在不同色彩空间下的预测 MOS 分数趋于收敛, 表明该指标对不同色彩空间具有鲁棒性。

4.5 整合合成分支与真实分支

上一节中, 我们概述了合成分支用 KADID-10K 数据集、真实分支用 KONIQ-10K 数据集的训练过程。然而, 这些数据集在图像数量和失真类型上都比较有限。为提升模型性能, 我们按照多任务训练的方式, 为每个分支扩充了多个数据集。具体来说, 合成分支在 KADID-10K、TID2013、PIPAL 数据集上训练, 真实分支在 KONIQ-10K、BID、SPAQ 数据集上训练。正如第 2.2 节所述, 我们通过整合两个在不同失真类型上训练的分支来缓解主观偏差, 每个分支针对各自的数据集各配一个独立的回归头。训练过程中, 我们从每个数据集随机留出 10% 作为验证集, 用于各分支的模型选择。关于 IQA 任务多任务训练过程的细节, 请读者参阅我们此前的论文 [17]。

表 6: LAR-IQA 模型在 UHD-IQA 数据集验证集上、不同训练设置的性能比较。

分支		微调	KAN 神 经元	PLCC	SRCC	KRCC
合成	真实					
✓			128	0.492	0.463	0.333
	✓		128	0.512	0.496	0.348
✓	✓		128	0.726	0.722	0.532
✓	✓		512	0.744	0.734	0.545
✓	✓	✓	512	0.809	0.803	0.611

预训练好的模型在去掉原回归头、加上新的 KAN 头之后被合并。如图 1 所示, 这些新头先对每个分支的嵌入下采样, 再把两个分支拼接起来, 最后接一个回归头。

在 UHD-IQA 数据集上训练: 我们分三步在 UHD-IQA 训练集上训练双分支模型。第一步, 冻结两个预训练分支 (真实和合成), 训练 KAN 头, 将输出从 1000 维降到 256 维 (每个分支 128 维), 再降到单个输出神经元。第二步, 改用一个更大的 KAN 头, 把 1000 维嵌入映射为每个分支 512 维 (共 1024 维), 再映射到一个神经元来完成回归任务。最后, 我们在新数据集上, 用真实分支和合成分支的预训练权重连同这个更大的 KAN 头一起微调模型。

表 6 给出了在 ECCV AIM 挑战赛⁴ 所用 UHD-IQA 数据集 [13] 上、各种训练策略的结果。模型用训练集训练、在验证集上测试。可以看到, 采用更大回归头 (512 维) 的模型性能更好。此外, 对模型做微调能显著改善结果, PLCC 达到约 0.81。总体而言, 这些结果很有前景。需要说明的是, 真实分支中我们把图像缩放到 $224 \times 224 \times 3$, 而合成分支中则使用全尺寸图像。

接下来, 我们给出所提模型与各基线在验证集和测试集上的对比评价。基线模型的结果取自原始文献 [13]。

4.6 与当前最优方法的比较

表 7 和表 8 分别给出了所提模型 (LAR-IQA) 与 SOTA 模型在 AIM UHD-IQA 验证集和测试集上的对比评价结果。可以看出, 无论在验证集还是测试集上, 我们的 LAR-IQA 在所有性能指标和复杂度指标上都优于现有 IQA 模型。所提模

⁴ <https://codalab.lisn.upsaclay.fr/competitions/19335>

表 7: 各方法在 UHD-IQA 验证集上的性能评价。↑ 表示越高越好，↓ 表示越低越好。最优和次优分数分别以粗体和下划线标出。

方法	PLCC↑	SRCC↑	KRCC↑	RMSE↓	MAE↓	参数量 ↓	MACs↓
HyperIQA [35]	0.182	0.524	0.359	0.087	0.055	27.3M	<u>211G</u>
Effnet-2C-MLSP [41]	0.627	0.615	0.445	0.060	0.050	-	345G
CONTRIQUE [24]	0.712	0.716	0.521	<u>0.049</u>	<u>0.038</u>	27.9M	855G
ARNIQA [3]	0.717	0.718	0.523	0.050	0.039	27.9M	855G
CLIP-IQA+ [39]	0.732	0.743	0.546	0.108	0.087	102M	895G
QualiCLIP [2]	<u>0.752</u>	<u>0.757</u>	<u>0.557</u>	0.079	0.064	102M	901G
LAR-IQA (MLP head)	<u>0.797</u>	<u>0.791</u>	<u>0.601</u>	<u>0.042</u>	<u>0.033</u>	21.2M	≤37G
LAR-IQA (KAN head)	0.809	0.803	0.611	0.040	0.031	21.1M	≤37G

型比最快的 SOTA 模型 HyperIQA 还快约 5.7 倍。此外，两个分支合计约 2100 万参数，适合在功耗和资源受限的移动设备上使用。

需要注意的是，两张表中的“MLP head”均指与 KAN 头 FLOPs 相近的 MLP 头。这样做是为了给读者提供在 UHD-IQA 数据集上 KAN 与 MLP 的性能对比。

此外，我们建议读者参阅 UHD-IQA 挑战赛 [14]，其中比较了为 UHD-IQA 任务提出的各种轻量模型。在挑战赛排名中，LAR-IQA 获得第二名。为保证比较公平，我们遵循了挑战赛的规则和数据集划分。

5 结论

针对缺乏兼具低复杂度、高精度和鲁棒性的 NR-IQA 指标这一问题，本工作提出了一个轻量 NR-IQA 模型，它在精度上超越当前 SOTA 模型，同时更适合在移动设备上部署。我们的模型采用双分支架构，分别处理真实失真和合成失真，因而相比只在单一失真类型上训练的模型，对各类失真更为鲁棒。此外，我们观察到，在多个数据集上分别训练有助于缓解各数据集固有的主观偏差，从而进一步提升模型的整体泛化能力。

为确保模型在各种色彩空间下的鲁棒性，我们在训练过程中引入了 RGB、YUV 和 LAB。此外，我们在质量回归模块中探索了 KAN，以替代常用的 MLP。实验结果表明，所提模型不仅在精度上超越 SOTA 的 NR-IQA 模型，复杂度也低得多，非常适合真实世界的应用。

表 8: 各方法在 UHD-IQA 测试集上的性能评价。↑ 表示越高越好, ↓ 表示越低越好。最优和次优分数分别以粗体和下划线标出。

方法	PLCC↑	SRCC↑	KRCC↑	RMSE↓	MAE↓	MACs(G)↓
HyperIQA [35]	0.103	0.553	0.389	0.118	0.070	<u>211</u>
Effnet-2C-MLSP [41]	0.641	0.675	0.491	0.074	0.059	345
CONTRIQUE [24]	0.678	0.732	0.532	0.073	0.052	855
ARNIQA [3]	0.694	0.739	0.544	0.074	0.052	855
CLIP-IQA+ [39]	0.709	0.747	0.551	0.111	0.089	895
QualiCLIP [2]	0.725	0.770	0.570	0.083	0.066	901
LAR-IQA (MLP head)	<u>0.774</u>	<u>0.809</u>	<u>0.616</u>	0.058	<u>0.042</u>	≤37
LAR-IQA (KAN head)	0.787	0.836	0.642	<u>0.061</u>	0.041	≤37

参考文献

1. Efficient-kan implementation [source code], <https://github.com/Blealtan/efficient-kan.git>
2. Agnolucci, L., Galteri, L., Bertini, M.: Quality-aware image-text alignment for real-world image quality assessment. arXiv:2403.11176 (2024)
3. Agnolucci, L., Galteri, L., Bertini, M., Del Bimbo, A.: Arniqa: Learning distortion manifold for image quality assessment. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 189–198 (2024)
4. Bampis, C.G., Gupta, P., Soundararajan, R., Bovik, A.C.: Speed-qa: Spatial efficient entropic differencing for image and video quality. IEEE signal processing letters **24**(9), 1333–1337 (2017)
5. Bodner, A.D., Tepsich, A.S., Spolski, J.N., Pourteau, S.: Convolutional kolmogorov-arnold networks. arXiv:2406.13155 (2024)
6. Chen, Z., Wang, J., Li, B., Yuan, C., Hu, W., Liu, J., Li, P., Wang, Y., Zhang, Y., Zhang, C.: Gmc-iqa: Exploiting global-correlation and mean-opinion consistency for no-reference image quality assessment. arXiv:2401.10511 (2024)
7. Cheon, M., Yoon, S.J., Kang, B., Lee, J.: Perceptual image quality assessment with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 433–442 (2021)
8. Ciancio, A., da Silva, E.A., Said, A., Samadani, R., Obrador, P., et al.: No-reference blur assessment of digital pictures based on multifeature classifiers. IEEE Transactions on image processing **20**(1), 64–75 (2010)

9. Fang, Y., Zhu, H., Zeng, Y., Ma, K., Wang, Z.: Perceptual quality assessment of smartphone photography. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3677–3686 (2020)
10. Ghadiyaram, D., Bovik, A.C.: Perceptual quality prediction on authentically distorted images using a bag of features approach. *Journal of vision* **17**(1), 32–32 (2017)
11. Golestaneh, S.A., Dadsetan, S., Kitani, K.M.: No-reference image quality assessment via transformers, relative ranking, and self-consistency. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1220–1230 (2022)
12. He, C., Zheng, Q., Zhu, R., Zeng, X., Fan, Y., Tu, Z.: Cover: A comprehensive video quality evaluator. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 5799–5809 (2024)
13. Hosu, V., Agnolucci, L., Wiedemann, O., Iso, D.: Uhd-iqa benchmark database: Pushing the boundaries of blind photo quality assessment. *arXiv preprint arXiv:2406.17472* (2024)
14. Hosu, V., Conde, M.V., Timofte, R., Agnolucci, L., Zadtootaghaj, S., Barman, N., et al.: AIM 2024 challenge on uhd blind photo quality assessment. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops (2024)
15. Hosu, V., Lin, H., Sziranyi, T., Saupe, D.: Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing* **29**, 4041–4056 (2020)
16. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al.: Searching for mobilenetv3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1314–1324 (2019)
17. Jamshidi Avanaki, N., Ghildiyal, A., Zadtootaghaj, S., Barman, N.: MSLIQA: Enhancing Learning Representations for Image Quality Assessment through Multi-Scale Learning. *arXiv* (2024)
18. Jinjin, G., Haoming, C., Haoyu, C., Xiaoxing, Y., Ren, J.S., Chao, D.: Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. pp. 633–651. Springer (2020)
19. Lin, H., Hosu, V., Saupe, D.: Kadid-10k: A large-scale artificially distorted iqa database. In: 2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX). pp. 1–3. IEEE (2019)
20. Liu, L., Dong, H., Huang, H., Bovik, A.C.: No-reference image quality assessment in curvelet domain. *Signal Processing: Image Communication* **29**(4), 494–505 (2014)

21. Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., Hou, T.Y., Tegmark, M.: Kan: Kolmogorov-arnold networks (2024)
22. Luo, C., He, X., Zhan, J., Wang, L., Gao, W., Dai, J.: Comparison and benchmarking of ai models and frameworks on mobile devices. *arXiv:2005.05085* (2020)
23. Madhusudana, P.C., Birkbeck, N., Wang, Y., Adsumilli, B., Bovik, A.C.: St-greed: Space-time generalized entropic differences for frame rate dependent video quality prediction. *IEEE Transactions on Image Processing* **30**, 7446–7457 (2021)
24. Madhusudana, P.C., Birkbeck, N., Wang, Y., Adsumilli, B., Bovik, A.C.: Image quality assessment using contrastive learning. *IEEE Transactions on Image Processing* **31**, 4149–4161 (2022)
25. Mehta, S., Rastegari, M.: Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178* (2021)
26. Mittag, G., Zadtootaghaj, S., Michael, T., Naderi, B., Möller, S.: Bias-aware loss for training image and speech quality prediction models from multiple datasets. In: 2021 13th International Conference on Quality of Multimedia Experience (QoMEX). pp. 97–102. IEEE (2021)
27. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing* **21**(12), 4695–4708 (2012)
28. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a “completely blind” image quality analyzer. *IEEE Signal processing letters* **20**(3), 209–212 (2012)
29. Moorthy, A.K., Bovik, A.C.: A two-step framework for constructing blind image quality indices. *IEEE Signal processing letters* **17**(5), 513–516 (2010)
30. Moorthy, A.K., Bovik, A.C.: Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE transactions on Image Processing* **20**(12), 3350–3364 (2011)
31. Ponomarenko, N., Jin, L., Ieremeiev, O., Lukin, V., Egiazarian, K., Astola, J., Vozel, B., Chehdi, K., Carli, M., Battisti, F., et al.: Image database tid2013: Peculiarities, results and perspectives. *Signal processing: Image communication* **30**, 57–77 (2015)
32. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763 (2021)
33. Saad, M.A., Bovik, A.C., Charrier, C.: Blind image quality assessment: A natural scene statistics approach in the dct domain. *IEEE transactions on Image Processing* **21**(8), 3339–3352 (2012)

34. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018)
35. Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., Zhang, Y.: Blindly assess image quality in the wild guided by a self-adaptive hyper network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3667–3676 (2020)
36. Sun, W., Min, X., Tu, D., Ma, S., Zhai, G.: Blind quality assessment for in-the-wild images via hierarchical feature fusion and iterative mixed database training. *IEEE Journal of Selected Topics in Signal Processing* **17**(6), 1178–1192 (2023)
37. Tu, Z., Yu, X., Wang, Y., Birkbeck, N., Adsumilli, B., Bovik, A.C.: Rapique: Rapid and accurate video quality prediction of user generated content. *IEEE Open Journal of Signal Processing* **2**, 425–440 (2021)
38. Vasu, P.K.A., Pouransari, H., Faghri, F., Vemulapalli, R., Tuzel, O.: Mobileclip: Fast image-text models through multi-modal reinforced training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15963–15974 (2024)
39. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2555–2563 (2023)
40. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing* **13**(4), 600–612 (2004)
41. Wiedemann, O., Hosu, V., Su, S., Saupe, D.: Konx: cross-resolution image quality assessment. *Quality and User Experience* **8**(1), 8 (2023)
42. Wu, H., Liao, L., Chaofeng, C., Hou, J., Wang, A., Sun, W., Yan, Q., Lin, W.: Disentangling aesthetic and technical effects for video quality assessment of user generated content. *arXiv:2211.04894* (2022)
43. Wu, H., Zhang, Z., Zhang, W., Chen, C., Liao, L., Li, C., Gao, Y., Wang, A., Zhang, E., Sun, W., et al.: Q-align: Teaching lmms for visual scoring via discrete text-defined levels. In: International Conference on Machine Learning
44. Xu, J., Ye, P., Li, Q., Du, H., Liu, Y., Doermann, D.: Blind image quality assessment based on high order statistics aggregation. *IEEE Transactions on Image Processing* **25**(9), 4444–4457 (2016)
45. Xue, W., Mou, X., Zhang, L., Bovik, A.C., Feng, X.: Blind image quality assessment using joint statistics of gradient magnitude and laplacian features. *IEEE Transactions on Image Processing* **23**(11), 4850–4862 (2014)

46. Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., Yang, Y.: Maniqa: Multi-dimension attention network for no-reference image quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1191–1200 (2022)
47. Ye, P., Kumar, J., Kang, L., Doermann, D.: Unsupervised feature learning framework for no-reference image quality assessment. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 1098–1105. IEEE (2012)
48. Yun, Y.K., Lin, W.: You Only Train Once: A Unified Framework for Both Full-Reference and No-Reference Image Quality Assessment. arXiv:2310.09560 (2024)
49. Zhang, L., Zhang, L., Bovik, A.C.: A feature-enriched completely blind image quality evaluator. IEEE Transactions on Image Processing **24**(8), 2579–2591 (2015)
50. Zhao, L., Wang, L.: A new lightweight network based on mobilenetv3. KSII Transactions on Internet and Information Systems (TIIS) **16**(1), 1–15 (2022)
51. Zhu, H., Wu, H., Li, Y., Zhang, Z., Chen, B., Zhu, L., Fang, Y., Zhai, G., Lin, W., Wang, S.: Adaptive image quality assessment via teaching large multimodal model to compare. arXiv:2405.19298 (2024)