场景感知图像感知得分 (SPIPS)：结合全局和局部感知进行图像质量评估

Zhiqiang Lao, Heather Yu
Futurewei Technologies Inc
Basking Ridge, New Jersey, USA
{ zlao,hyu } @futurewei.com

***Abstract*—人工智能的快速进步和智能手机的广泛使用导致了图像数据（包括真实的相机捕捉图像和虚拟的 AI 生成图像）的指数级增长。这一激增强调了对能够准确反映人类视觉感知的强大图像质量评估（IQA）方法的关键需求。传统的 IQA 技术主要依赖于空间特征——例如信噪比、局部结构失真和纹理不一致性——来识别伪影。虽然这些方法对于未经处理或传统方式修改的图像是有效的，但在由深度神经网络（DNN）驱动的现代图像后期处理的背景下则显得不足。基于 DNN 的图像生成、增强和修复模型的兴起极大地提高了视觉质量，但同时也使得准确评估变得愈加复杂。为了解决这个问题，我们提出了一种新颖的 IQA 方法，在深度学习方法与人类感知之间架起桥梁。我们的模型将深度特征解构为高级语义信息和低级感知细节，并分别处理每个流。这些特征随后与传统的 IQA 度量结合，提供了一个更全面的评估框架。这种混合设计使得模型能够同时评估全局上下文和复杂的图像细节，更好地反映了人类视觉过程，后者首先解释整体结构然后再关注细微元素。最终阶段采用了多层感知机（MLP）将综合特征映射到一个简明的质量评分。实验结果表明，与现有的 IQA 模型相比，我们的方法在与人类感知判断的一致性方面取得了更好的效果。**

***Index Terms*—computer vision, artificial intelligence, image quality assessment, visual perception, deep learning**

I. 介绍

计算机视觉在现实世界应用中的有效性取决于目标函数与人类视觉系统的契合程度。去噪、超分辨率以及有损压缩等任务的端到端解决方案 [1], [2] 依赖于能够准确反映人类对视觉变化的感知的可微分相似性度量。然而，通常使用的像素级差异度量标准，如 PSNR 和 MSE，虽然是可微分的，但却无法很好地与人类视觉感知对齐。

像素级度量无法准确捕捉人类感知的局限性，促使了补丁级相似度度量的发展，这些度量受到人类心理学中心理物理学分支的影响。目前最有效的是多尺度结构相似度度量（MS-SSIM）[32], [54] ，它考虑了亮度和对比度感知。然而，尽管有这些进步，人类视觉系统的复杂性仍然难以手动建模。这在 MS-SSIM 在标准化图像质量评估（IQA）实验 [5] 中预测人类偏好的不足之处中尤为明显。

为了超越人工设计的相似性度量，研究人员采用从大型预训练神经网络中提取的深度特征。例如，学习感知图像块相似性（LPIPS）[5] 度量依赖于这些深度特征之间的 L2 距离来近似人类感知。在同一研究中，作者引入了伯克利 Adobe 感知块相似性 (BAPPS) 数据集，该数据集已成为广泛认可的相似性度量感知一致性评估的基准。LPIPS 利用深度特征作为输入，通过一个较小的神经网络进行训练，该神经网络使用人类注释的 BAPPS 数据捕捉人类对图像相似性的偏好 [6] 。

978-1-xxxx-xxxx-x/25/ $ 31.00 © 2025 IEEE

人类视觉系统表现出强大的层次感知能力，通常采用自上而下的方法——先把握整体场景，然后再关注细节。相比之下，由于像摄像机这样的视觉传感器的局限性，图像在计算机中以像素阵列的形式存储，导致计算机视觉采用自下而上的方法，先注重细节再形成整体理解。因此，计算机视觉中的图像质量评估往往强调精细细节，而不是更广泛的图像背景。基于学习的图像质量评估方法通常利用诸如 AlexNet 或 VGG 这样的主干网络，这些网络提供高层次的感知能力。然而，这些模型最初是为图像分类、目标检测和分割等任务开发的——这些任务优先识别显著内容而非评估视觉质量。因此，它们提取的特征本质上不适合图像质量评估的目标。依赖为分类优化的特征来评估图像质量引入了一种不匹配，可能会影响性能。这突显了当前基于学习的方法的一个主要局限：缺乏针对任务的适应性，分类驱动的特征可能会忽略感知质量的关键方面。鉴于这些限制，本文通过整合深度和传统特征来增强全局特征在图像质量评估中的作用。所提出的度量依据这一原则实现了与人类视觉感知更紧密的对齐。

我们的贡献可以概括为：

- 高层次语义特征和低层次感知图像特征被区别对待，独立处理，然后进行融合。这种差异化处理使模型在评估图像质量时能够同时考虑细粒度细节和整体场景理解，从而实现更全面的评估，更接近人类视觉感知。
- 将传统图像质量评估技术与深度学习模型相结合，有助于减少它们评估结果中的常见不一致性。

在本节中，我们回顾了与无数据和学习到的（包括无监督和有监督）全参考图像质量评估（FR-IQA）密切相关的文献。

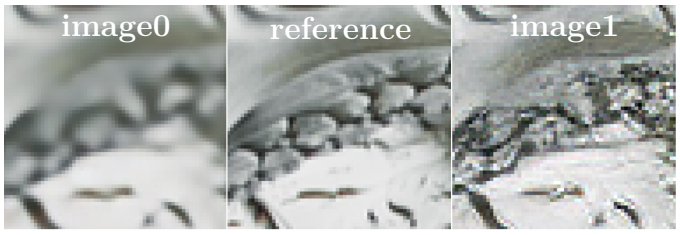
A. 无数据 FR-IQA

在像素级别工作的无数据失真度量，例如均方误差 (MSE)，通常用于有损压缩应用，但是长久以来已知与人类感知的相关性较差。块级别度量已被证明在心理物理任务上与人类判断的相关性更好。最著名的是结构相似性指数（SSIM [32] 以及其多尺度变量 MS-SSIM [54] ），通过比较高层块特征如亮度和对比度来定义图像之间的距离 [32] 。SSIM [32] 广泛用于商用电视应用，而 MS-SSIM [54] 是评估许多计算机视觉任务性能的标准度量。本工作中提出的方法也是无数据的，并在基准数据集上优于 MS-SSIM。

PSNR 计算参考图像和生成图像之间的平方差异，提供了一种直观的图像质量下降测量方法。视觉信息保真度（VIF）是另一种度量标准，它根据自然场景统计和人类视觉系统的特性量化与参考图像相比保留了多少视觉信息。

		PSNR	SSIM	VIF	LPIPS	DISTS	SPIPS	Human
0	>						✓	✓
1								
1	>	✓	✓	✓	✓	✓		
0								

(a) Human Preference : $image0 > reference > image1$



		PSNR	SSIM	VIF	LPIPS	DISTS	SPIPS	Human
0	>	✓	✓	✓	✓	✓		
1								
1	>						✓	✓
0								

(b) Human Preference : $image0 < reference < image1$

Fig. 1: 针对 BAPPS 数据集的人类偏好对不同指标进行定性比较。SPIPS 在不同的失真类型中始终与人类判断保持一致。

SSIMPLUS 由 Rehman 等人提出，通过整合人类视觉、显示特性和观看条件等因素，推动了 SSIM 的发展，使实时感知质量预测成为可能。有一些基于信息理论模型的成功全参考（FR）方法，例如信息保真度标准（IFC）。

B. 学习的全参考图像质量评估

在基于模型的方面，FID 通过 Inception 模型捕获的真实图像和生成图像的特征向量之间的统计距离，评估视觉保真度的指标 [15] 。Inception 得分（IS）由 Salimans 等人开发 [16] ，通过分类器在像 ImageNet 这样的多样化数据集上训练的预测熵来量化图像的多样性和清晰度。核 Inception 距离（KID）由 Binkowski 等人提出 [17] ，通过使用多项式核计算特征分布之间的平方最大平均差异改进了 FID。此方法避免对激活的分布形式做出假设，更好地适应了深度网络中 ReLU 激活的非负性质。

许多学习型全参考图像质量评估（FR-IQA）方法设计上借鉴了 [5] 的学习感知图像块相似性（LPIPS）方法和 [19] 的深度图像结构和纹理相似性（DISTS），其中神经网络在某个辅助任务上进行训练，并将中间层作为输入图像的感知表示。定义图像之间的无监督距离为其表示之间差异的 L2 范数。监督距离则将这些表示作为输入给第二个模型，该模型在关于输入图像感知质量的人为标注数据上进行训练（例如，第二节中讨论的 2-AFC 数据集的标签）。从在其辅助任务上表现优异的神经网络中提取表示并不能保证在感知任务上取得良好表现 [6] ，因此难以决定哪些现有模型将产生与感知相关的距离函数。在 [20] 中，使用结构-纹理分解（STD）来测量结构、纹理和高频相似性。Zhang 等人 [21] 提出了一种感知驱动的相似性-清晰度权衡，以平衡参考相似性与无参考清晰度之间的质量分数，并很好地处理由基于 GAN 的超分辨率算法生成的视觉上令人愉悦的假纹理。

自监督方法被 Madhusudana 等人 [10], [23] 和 Wei 等人 [11] 用于无监督和监督的 FR-IQA。图像通过预定义的失真函数进行破坏，然后使用对比性成对损失训练神经网络，以预测失真的类型和程度。无监督距离的定义如前所述，并使用岭回归学习监督距离函数。这种方法需要训练数据，而我们的方法则完全不需要训练。

得益于强大的特征表达能力，基于深度学习的通用方法，包括基于 CNN 的模型 [22]–[25] 和基于 Transformer 的模型 [26]–[29] 在处理图像失真方面也表现出色。它们在有限的训练样本规模和复杂的失真条件方面投入了大量精力。在 [24] 中，特征提取网络以失真感知的方式训练，以应对各种失真。Zhou 等人 [25] 设计了一种自监督架构，以增强内容和失真的表示能力。

II. 拟议方法

图 2 展示了我们的 SPIPS 图像质量评估模型的整体框架，该模型由三个关键模块组成：

传统图像质量评估（IQA）模块：给定一对输入图像，该模块计算诸如 PSNR、SSIM 和 MS-SSIM 等传统图像质量指标。与输出单一数值的传统方法不同，该模块生成与输入图像大小相同的评估图像，提供空间感知的质量评估。

基于深度特征的 IQA 模块：利用预训练的深度学习模型（例如，AlexNet, VGG），该模块从多个层提取特征。由于不同层捕捉图像的不同方面，这些特征被分为两组：低级图像感知特征和高级语义特征。例如，在 AlexNet 中，前三层的特征被分类为感知特征，而后两层被指定为语义特征。对于给定的一对输入图像，计算它们各自的感知和语义特征，对应层之间的均方误差（MSE）作为质量评估结果，量化评估图像与参考图像之间的差异。

图像质量特征提取模块：此模块从前两个模块的输出中提取与质量相关的特征。这些提取出的特征与图像质量直接相关，然后融合以生成最终的图像质量得分。

A. 传统 IQA 模块

传统的图像质量评价模块使用行业标准公式（包括 PSNR、SSIM 和 MS-SSIM [30]–[32], [54] ）来评估待评估图像的质量。为了与深度图像质量评估模块保持一致并区别于传统使用方式，进行了两个关键修改：

- 这个模块不是生成一个单一的数值分数，而是输出每个像素的质量评估值，从而生成一个质量评估图，而不是一个平均值。
- 通常，PSNR、SSIM 和 MS-SSIM 的定义是，数值越高表示图像质量越好。然而，在基于深度特征的模块中，图像质量是通过测量被评估的图像和参考图像之

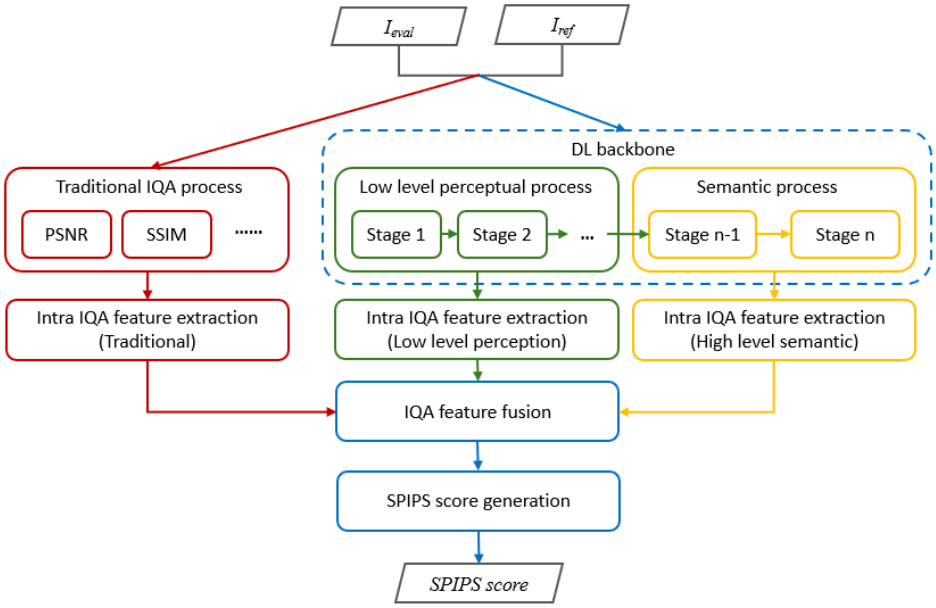


Fig. 2: 我们的 SPIPS 模型的整体框架分为三个模块：传统图像质量评估模块（红色）、低级图像感知特征评估模块（绿色）和高级图像语义特征评估模块（黄色）。SPIPS 需要两个输入图像，I_{eval} （待评估的图像）和 I_{ref} （参考或真实值图像）。传统图像质量评估模块（红色）使用 PSNR、SSIM 和 MS-SSIM 等标准行业指标分析 I_{eval} 和 I_{ref} 对应区域之间的质量差异，生成质量评估图。然后将 I_{eval} 和 I_{ref} 输入预训练的深度学习主干网络（例如，AlexNet、VGG、SqueezeNet，用浅蓝色虚线表示），从每一层（不包括全连接层前）提取特征图。这些特征图根据其侧重点进行分类：低级图像感知特征（从除最后两个层外的所有层获得的）和高级语义特征（从特征堆栈的最后两个层获得的）。然后计算 I_{eval} 和 I_{ref} 的深度特征表示之间的均方误差（MSE），以生成深度评估图。每个评估图在组合成 I_{eval} 的最终图像质量得分之前，都会进行独立的质量增强，通过加权平均过程完成。与 LPIPS 模型 [5] 的优化方法类似，将该得分与人类视觉感知得分进行比较以计算损失值。通过类似于 LPIPS 模型的训练策略，SPIPS 模型参数通过反向传播进行迭代优化。

间特征表示的差异来评估的，其中差异越小表示质量越高。为了一致性，首先将 PSNR、SSIM 和 MS-SSIM 评估图规范化到范围 $[0,1]$ ，然后从 1 中减去每个像素的值。此转换确保在最终评估图上，较低的数值对应于较高的图像质量。

给定一对图像，$I_{eval} \in \mathbb{R}^{3\times H\times W}$ （待评估的图像）和 $I_{ref} \in \mathbb{R}^{3\times H\times W}$ （真实图像），使用 PSNR、SSIM 和 MS-SSIM 计算来评估它们的质量差异。

$$Q_{psnr} = 1 - \mathcal{N}(\mathcal{P}(I_{eval}, I_{ref})) \tag{1}$$

$$Q_{ssim} = 1 - \mathcal{N}(\mathcal{S}(I_{eval}, I_{ref})) \tag{2}$$

$$Q_{msssim} = 1 - \mathcal{N}(\mathcal{S}_{\mathrm{MS}}(I_{eval}, I_{ref})) \tag{3}$$

其中，$Q_{psnr} \in \mathbb{R}^{3\times H\times W}$ 、$Q_{ssim} \in \mathbb{R}^{3\times H\times W}$ 、$Q_{msssim} \in \mathbb{R}^{C\times H\times W}$ 、C 是刻度的数量，$\mathcal{P}(\cdot)$ 是 PSNR 操作，$\mathcal{S}(\cdot)$ 是 SSIM 操作，$\mathcal{S}_{\mathrm{MS}}(\cdot)$ 是 MS-SSIM 操作，$\mathcal{N}(\cdot)$ 是对 $[0,1]$ 操作的归一化。

B. 基于深度特征的 IQA 模块

基于卷积神经网络（CNN）的深度学习模型具有强大的特征提取能力，并已广泛应用于计算机视觉，包括图像质量评估。我们的 SPIPS 模型同样利用了这些强大的图像特征。基于 CNN 的模型通常在不同分辨率下提取特征，其中高分辨率特征强调图像细节，而低分辨率特征则捕捉更广泛的语义特征。

与类似的方法一样，SPIPS 利用如 AlexNet、VGG 和 SqueezeNet 等预训练的 CNN 模型进行特征提取。然而，与之前的方法不同的是，SPIPS 更加重视语义特征。除了提取语义信息外，它还捕获与图像质量特定相关的特征。这些特征随后被与图像细节的感知特征和传统的图像质量指标整合在一起，从而形成一种更全面的视觉评估，更加贴近人类的视觉感知。

让我们考虑一对图像，$I_{eval} \in \mathbb{R}^{3\times H\times W}$ （要评估的图像）和 $I_{ref} \in \mathbb{R}^{3\times H\times W}$ （真实值），基于 CNN 的特征提取如下。

$$F_{eval} = \left\{ \Phi^{(l)}_{\mathrm{CNN}}(I_{eval}) \mid l = 1, \ldots, L \right\} \tag{4}$$

$$F_{ref} = \left\{ \Phi^{(l)}_{\mathrm{CNN}}(I_{ref}) \mid l = 1, \ldots, L \right\} \tag{5}$$

，其中 $F^{(l)}_{eval}$ ，$F^{(l)}_{ref} \in \mathbb{R}^{C^{(l)}\times H^{(l)}\times W^{(l)}}$ ，$C^{(l)}, H^{(l)}, W^{(l)}$ 分别是层 l 处的特征通道数、高度和宽度，CNN 是预训练的主干网络（AlexNet, VGG 等），L 是 CNN 中特征提取层的总数。

可以通过如下方式计算 CNN 特征的图像质量图。

$$Q_{\mathrm{CNN}} = \left\{ (F^{(l)}_{eval} - F^{(l)}_{ref}) \odot (F^{(l)}_{eval} - F^{(l)}_{ref}) \mid l = 1, \ldots, L \right\} \tag{6}$$

，其中 $Q_{\text{CNN}}^{(l)} \in \mathbb{R}^{\text{C}^{(l)} \times \text{H}^{(l)} \times \text{W}^{(l)}}$ ，$\odot$ 表示 Hadamard 积 [34]（元素逐次相乘），该操作计算 $F_{eval}^{(l)}$ 和 $F_{ref}^{(l)}$ 对应元素的平方差。

基于在不同层生成的 Q_{CNN} 的质量评估重点，Q_{CNN} 被分为图像质量感知组和图像质量语义组，定义如下：以 AlexNet 为例，它由五个特征提取层（$L = 5$）组成，将方程中的 CNN 替换为 AlexNet 可以得到：$Q_{percept} = \left\{ Q_{\text{AlexNet}}^{(l)} \mid l = 1, 2, 3 \right\}$ 和 $Q_{semantic} = \left\{ Q_{\text{AlexNet}}^{(l)} \mid l = 4, 5 \right\}$

在这个阶段，我们已经获得了三组图像质量评估图：$Q_{tradition} = \{Q_{psnr}, Q_{ssim}, Q_{msssim}\}$ 、$Q_{percept}$ 和 $Q_{semantic}$ 。接下来，我们需要从这些质量评估图中提取与图像质量直接相关的特征。类似于图像特征提取，提取图像质量特征的过程也使用卷积操作。每一组图像质量集合经过不同的卷积处理，以保证提取的图像质量特征能够有效捕捉每个组的独特特性。

$$F_{tradition} = \mathcal{R}(\mathcal{C}_{tradition}(Q_{tradition})) \tag{7}$$

$$F_{percept} = \mathcal{R}(\mathcal{C}_{percept}(Q_{percept})) \tag{8}$$

$$F_{semantic} = \mathcal{R}(\mathcal{C}_{semantic}(Q_{semantic})) \tag{9}$$

，其中 $\mathcal{C}_{tradition}(\cdot)$ 、$\mathcal{C}_{percept}(\cdot)$ 和 $\mathcal{C}_{semantic}(\cdot)$ 是分别应用于 $Q_{tradition}$ 、$Q_{percept}$ 和 $Q_{semantic}$ 的卷积操作，$\mathcal{R}(\cdot)$ 是 ReLU。

C. 图像质量评估特征融合和评分计算

最终的图像质量评估得分是通过结合提取的图像质量特征以实现更全面的评估来获得的，表示如下：

$$score_{spips} = \lambda_1 \bar{F}_{tradition} + \lambda_2 \bar{F}_{percept} + \lambda_3 \bar{F}_{semantic} \tag{10}$$

，其中 λ_1 、λ_2 和 λ_3 分别是传统、感知和语义的加权项，用于确定它们在最终 $score_{spips}$ 计算中的贡献，及 $\lambda_1 + \lambda_2 + \lambda_3 = 1$ 。$\bar{F}_{tradition}$ 是 $F_{tradition}$ 的平均值，$\bar{F}_{percept}$ 和 $\bar{F}_{semantic}$ 也是如此。

$score_{spips}$ 代表对 I_{eval} 的最终图像质量评估，其中较低的值表示更好的质量。

III. 实验

A. 数据集和实验设置

BAPPS（Berkeley Adobe 感知补丁相似性）数据集 [5] 用于训练和验证 SPIPS 模型。该数据集专为分析计算机视觉模型进行的图像质量评估与人类视觉系统感知之间的差异而设计。它包括了由于自然图像处理、传统算法（例如，图像压缩）和基于 CNN 的模型产生的多种失真。这些失真被分为六组：传统失真（Trad）、基于 CNN 的失真（CNN）、视频去模糊（Deblur）、帧插值（Interp）、超分辨率（SR）和上色（Color）。表 I 总结了 BAPPS 数据集的关键统计数据。基于该数据集构建的 LPIPS 模型是最广泛使用的基于深度特征的图像质量评估模型之一。数据集中的图像分辨率为 256 × 256。

BAPPS 数据集使用两种方法来评估图像质量。第一种方法称为 2AFC（两种选择强迫选择），每个样本会显示三个图像：$image0$ 、$reference$ 和 $image1$ 。任务是决定两个图像（$image0$ 或 $image1$ ）中哪个的质量更接近参考图像。第二种方法称为 JND（刚好可察觉的差异），它为每个样本显示一对图像，并询问它们之间是否存在明显差异。

Properties	Value
# of Ref. Images	187.7k
# of Test Images	375.4k
Distortion / Enhancement Type	Simulated / DNN-based
Image resolution	256 × 256
# of Human Annotations	484.3k

TABLE I BAPPS 数据集的关键统计数据。

通常使用的相关性指标，包括 Spearman 秩相关系数（SRCC）[42] 、Pearson 线性相关系数（PLCC）[41] 和 Kendall 秩相关系数（KRCC）[43] ，用于评估基于计算机的图像质量评估与人类感知之间的一致性。这些指标中的每一个都从不同的角度对预测性能进行评价。

SRCC 是一种基于排名的非参数度量，用于测量预测与人工标注的图像质量分数之间的单调关系，使其特别适用于评估相对排名一致性而不是绝对差异。另一方面，PLCC 评估预测分数与真实分数之间的线性相关性，反映出数值预测的准确性以及其与人工判断的一致程度。最后，KRCC 量化预测值的排名与人工提供标签的排名之间的相似性，提供了另一种预测可靠性视角。

通过同时利用这三个相关指标，可以对模型逼近人类对图像质量感知的能力进行综合评估。

我们的模型整合了传统和深度学习组件。因此，在我们的实验比较中，我们评估了来自这两类的代表性方法——特征工程和深度学习方法。具体来说，我们比较了 PSNR [30] 、SSIM [32] 、VIF [60] 、DISTS [19] 和 LPIPS [5] 。

B. 结果评估

图 1 展示了 2AFC 子集的测试拆分中的两个例子。如前所述，每个例子包括三个图像：$image0$ 、$image1$ 和一个 $reference$ 图像。每个算法或模型（我们的方法和基准线）会针对 $image0$ 和 $image1$ 相对于参考图像赋予质量评分。然后将这些相对分数与人类判断（真实值）进行比较，以评估其正确性。

在图 1 (a) 中显示的例子中，人类视觉系统明显将图像排序为 $image0 > reference > image1$ ，这表明 $image0$ 的质量高于 $image1$ 。图 1(a) 下方的对号代表每个算法或模型对 $image0$ 和 $image1$ 相对质量的决策。所有基线方法都错误地把 $image1$ 排名为比 $image0$ 质量更高，这与人类的感知相反。而我们的模型则与人类视觉判断一致。

类似地，在图 1 (b) 中，人类视觉系统可以轻易辨别出 $image1$ 的质量显著优于 $image0$ ，正确的排名应该是 $image1 > reference > image0$ 。然而，所有基线算法和模型都错误地将 $image0$ 排在 $image1$ 之上，认为其具有更高的质量，这与人类的感知相悖。相比之下，我们的模型的判断与人类视觉评估一致，证明在这个例子中，SPIPS 更接近于人类视觉系统的反映。

C. 消融研究

为了评估在我们的模型中新引入组件的影响，我们进行了一个消融研究。具体来说，我们设计了两种消融场景：

消融实验 1：移除语义模块

消融研究 2：排除传统的图像质量评价指标如 PSNR 和 SSIM

Fig. 3: 对完整的 SPIPS 模型及其没有语义模块的消融变体进行定性比较。SPIPS0 和 SPIPS1 分别表示 SPIPS 模型对 *image*0 和 *image*1 的评估分数。图像 *image*0 和 *image*1 是正在比较的两个候选图像，*reference* 作为真实情况。*ablation*0 和 *ablation*1 表示不包括语义模块的 SPIPS 模型消融版本对 *image*0 和 *image*1 的评估分数。人类偏好：*image*0 < *image*1

Fig. 4: 完整 SPIPS 模型与其去除传统图像质量评估指标（如 PSNR 和 SSIM）的变体进行定性比较。SPIPS0、SPIPS1、*image*0 、*reference* 和 *image*1 的作用与图 3 中的一致。同样，*ablation*0 和 *ablation*1 代表由 SPIPS 模型的去除变体生成的 *image*0 和 *image*1 的评估得分。然而，与图 3 不同的是，此处使用的变体排除了传统的图像质量评估指标。人类偏好：*image*0 > *image*1

对于每次消融实验，我们训练了一个对应的模型并在与完整模型相同的数据集上进行评估。

图 3 展示了 SPIPS 模型与其消融变体消融 1 的图像质量评估对比。由于消融 1 模型不包含语义特征模块，所以它缺乏 SPIPS 模型评估整体图像结构的能力，转而更侧重于局部细节。在这个例子中，*image*0 看起来更平滑或分辨率更低，而 *image*1 则具有更高的分辨率但也有更多的噪声。因为消融 1 在评估图像质量时更重视噪声，所以它认为 *image*1 的质量比 *image*0 更低。

图 3 中的特征图阐述了 SPIPS 和消融 1 模型在评估 *image*0 和 *image*1 质量上的差异（蓝色表示差异小，红色/黄色表示差异大）。*image*0 、*reference* 和 *image*1 的下半部分主要包含均质内容（例如，水面）。在 SPIPS0 特征图中，这一区域在 *image*0 和 *reference* 之间表现出极小的差异。在 SPIPS1 图中，同一区域 *image*1 与 *reference* 之间的差异稍大——反映为蓝绿色调——但仍然相对较小。然而，消融 1 特征图在下区域显示绿色-黄色色调，表明消融 1 模型认为 *image*1 和 *reference* 之间存在显著差异。这与人类感知相反，人类会认为 *image*1 和 *reference* 的下半部分在视觉上类似。差异的原因在于 *image*1 包含更多的噪声和细节，消融 1 模型将其解释为不相似，而人类观察者往往忽略这种细小的变化并认为 *image*1 和 *reference* 更加相似。

图 4 展示了 SPIPS 模型与消融 2 变体之间的比较。与消融 1 模型不同，消融 2 优先考虑语义和感知特征，而弱化细节图像信息。因此，它对 *image*1 的质量评级高于 *image*0 。这体现在其对 *image*1 和 *reference* 塔区域结构差异的增强，而在天空等更均匀区域则淡化了细微的细节差异。

统计结果列在表格 II 、III 和 IV 中，这些表格比较了去除组件后的两个模型和完整模型的性能。

正如预期的那样，两个消融模型在所有指标上都表现得比完整模型差。这些结果验证了新增组件的有效性，并加强了它们对所提模型整体性能的贡献。

2AFC	PLCC ↑ (Pearson's Linear Correlation Coefficient)					
	CNN	Color	Deblur	Interp	SR	Trad
SPIPS-abla1	0.78	0.45	0.36	0.39	0.52	0.67
SPIPS-abla2	0.79	0.47	0.39	0.42	0.54	0.69
SPIPS	0.81	0.51	0.41	0.45	0.59	0.71

TABLE II SPIPS 模型和消融实验的 PLCC 比较（2AFC 数据集）。

2AFC	SRCC ↑ (Spearman's Rank Correlation Coefficient)					
	CNN	Color	Deblur	Interp	SR	Trad
SPIPS-abla1	0.77	0.45	0.36	0.38	0.52	0.67
SPIPS-abla2	0.78	0.47	0.39	0.41	0.54	0.69
SPIPS	0.80	0.51	0.41	0.45	0.59	0.71

TABLE III SPIPS 模型和消融方法的 SRCC 比较（2AFC 数据集）。

2AFC	KRCC ↑ (Kendall's Rank Correlation Coefficient)					
	CNN	Color	Deblur	Interp	SR	Trad
SPIPS-abla1	0.69	0.40	0.32	0.34	0.46	0.60
SPIPS-abla2	0.70	0.41	0.35	0.36	0.47	0.61
SPIPS	0.71	0.46	0.36	0.40	0.52	0.63

TABLE IV SPIPS 模型和消融的 KRCC 比较（2AFC 数据集）。

D. 与以前工作的比较

图 1 展示了两个示例，每个示例包含三张图像：*image*0 、*image*1 和 *reference* 图像。在 B 节中，我们突出了我

们的模型与基线方法在视觉判断上的区别。基于这种直观的视觉比较，我们在本节中提供了 SPIPS 和其他模型的详细定量分析。

正如在第 III-A 节中介绍的，BAPPS 数据集将图像质量分为六类：CNN、Color、Deblur、Interp、SR 和 Trad。我们不是提供单一的整体评分，而是分别报告每个类别的表现，以提供更全面的评估。

表 V 显示了六个不同模型/算法输出与六个类别中的人类判断之间的皮尔逊相关系数。较高的值表示与人类感知的更强一致性。如图所示，我们的模型在所有类别中均始终实现最高相关性，表明 SPIPS 比竞争方法更接近反映人类视觉评估。

表 VI 和 VII 使用不同的相关性指标提供了类似的定量比较。这些结果与表 V 中的结果一致，进一步验证了我们模型的优越性能。

此外，表格 VIII 展示了 BAPPS 数据集的 JND 子集的结果。由于 JND 任务专注于识别两幅图像之间是否存在可感知的差异，而无需精确的数值判断，传统的测量方法如 PLCC 则不太有效。相反，基于排序的相关性测量，如 SRCC 和 KRCC 更为合适。如表格 VIII 所示，我们的模型在 SRCC 和 KRCC 方面表现出色，超过所有基线，尽管在 CNN 子数据集上的 PLCC 略低于 LPIPS 和 DISTS，但这证明了其在捕捉细粒度感知差异方面的优势。

2AFC	PLCC ↑ (Pearson's Linear Correlation Coefficient)					
	CNN	Color	Deblur	Interp	SR	Trad
PSNR[30]	0.72	0.40	0.32	0.14	0.41	0.19
SSIM[32]	0.71	0.38	0.28	0.16	0.35	0.34
VIF[60]	0.62	0.04	0.32	0.32	0.44	0.16
DISTS[19]	0.75	0.37	0.35	0.42	0.56	0.63
LPIPS[5]	0.78	0.38	0.35	0.42	0.58	0.61
SPIPS (ours)	0.81	0.51	0.41	0.45	0.59	0.71

TABLE V PLCC 比较：SPIPS 与先前的 FR-IQA 指标（2AFC 数据集）

2AFC	SRCC ↑ (Spearman's Rank Correlation Coefficient)					
	CNN	Color	Deblur	Interp	SR	Trad
PSNR[30]	0.72	0.40	0.31	0.14	0.41	0.19
SSIM[32]	0.71	0.38	0.27	0.16	0.35	0.34
VIF[60]	0.61	0.04	0.31	0.32	0.44	0.17
DISTS[19]	0.75	0.37	0.34	0.41	0.56	0.63
LPIPS[5]	0.77	0.38	0.35	0.41	0.58	0.61
SPIPS (ours)	0.80	0.51	0.41	0.45	0.59	0.71

TABLE VI SRCC 比较：SPIPS 与以往的 FR-IQA 指标（2AFC 数据集）

2AFC	KRCC ↑ (Kendall's Rank Correlation Coefficient)					
	CNN	Color	Deblur	Interp	SR	Trad
PSNR[30]	0.65	0.35	0.28	0.12	0.37	0.17
SSIM[32]	0.64	0.34	0.24	0.14	0.31	0.30
VIF[60]	0.55	0.03	0.28	0.28	0.39	0.15
DISTS[19]	0.67	0.33	0.31	0.36	0.50	0.56
LPIPS[5]	0.69	0.34	0.31	0.37	0.52	0.54
SPIPS (ours)	0.71	0.46	0.36	0.40	0.52	0.63

TABLE VII KRCC 比较：SPIPS 与先前的 FR-IQA 指标（2AFC 数据集）。

JND	CNN			Trad		
	PLCC ↑	SRCC ↑	KRCC ↑	PLCC ↑	SRCC ↑	KRCC ↑
PSNR[30]	0.61	0.63	0.50	0.16	0.10	0.07
SSIM[32]	0.46	0.59	0.46	0.27	0.28	0.22
VIF[60]	0.53	0.56	0.44	0.13	0.07	0.05
DISTS[19]	0.63	0.67	0.53	0.01	0.05	0.03
LPIPS[5]	0.63	0.71	0.56	0.55	0.57	0.45
SPIPS (ours)	0.60	0.73	0.58	0.55	0.65	0.52

TABLE VIII PLCC、SRCC 和 KRCC 比较：SPIPS 与先前 FR-IQA 指标 (JND 数据集)

IV. 结论

受近年来越来越多应用于深度学习的图像质量评估 (IQA) 方法的推动，我们还观察到了不同 IQA 指标之间的矛盾以及计算机视觉算法与人类视觉系统做出的评估之间的明显差异。这些观察促使我们探索一种更统一且与人类视觉系统一致的图像质量评估方法。

在这项工作中，我们提出了一个统一的框架，整合了传统和基于深度学习的图像质量评估 (IQA) 方法。我们的方法分别提取和处理语义特征，以捕捉高层次的图像内容，然后将其与低层次的感知特征和传统的 IQA 指标结合。此整合旨在提供一种更全面和客观的图像质量评估，与人类视觉感知更加一致。

初步结果令人鼓舞。展望未来，我们的工作将主要集中在两个方向。首先，我们计划用更多样化和感知上有意义的样本来扩展我们的数据集，以增强模型对图像质量的理解。其次，我们旨在引入视觉变换器结构以进一步优化我们的模型。通过在不同特征图上生成补丁的 tokens 并建模 Token 之间的关系，我们希望实现对图像质量的更深刻且认知上更贴近人类感知习惯的理解。

我们对这项研究的持续进展持乐观态度，并期待在不久的将来分享更多的进展。

References

[1] Ballé, Johannes, Valero Laparra, and Eero P. Simoncelli. "End-to-end optimized image compression." arXiv preprint arXiv:1611.01704 2016.

[2] Ballé, Johannes, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. "Variational image compression with a scale hyperprior." arXiv preprint arXiv:1802.01436 2018.

[3] Wang, Zhou, Eero P. Simoncelli, and Alan C. Bovik. "Multiscale structural similarity for image quality assessment." The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003. Vol. 2. Ieee, 2003.

[4] Wang, Zhou, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. "Image quality assessment: from error visibility to structural similarity." IEEE transactions on image processing 13, no. 4 (2004): 600-612.

[5] Zhang, Richard, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. "The unreasonable effectiveness of deep features as a perceptual metric." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 586-595. 2018.

[6] Kumar, Manoj, Neil Houlsby, Nal Kalchbrenner, and Ekin D. Cubuk. "Do better ImageNet classifiers assess perceptual similarity better?.", Transactions of Machine Le

[7] Cover, Thomas M. Elements of information theory. John Wiley & Sons, 1999.

[8] Girod, Bernd. "What's wrong with mean-squared error?." In Digital images and human vision, pp. 207-220. 1993.

[9] Madhusudana, Pavan C., Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C. Bovik. "Image quality assessment using contrastive learning." IEEE Transactions on Image Processing 31 (2022): 4149-4161.

[10] Madhusudana, Pavan C., Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C. Bovik. "Conviqt: Contrastive video quality estimator." IEEE Transactions on Image Processing 32 (2023): 5138-5152.

[11] Wei, Xuekai, Jing Li, Mingliang Zhou, and Xianmin Wang. "Contrastive distortion-level learning-based no-reference image-quality assessment." International Journal of Intelligent Systems 37, no. 11 (2022): 8730-8746.

[12] Talebi, Hossein, and Peyman Milanfar. "Learned perceptual image enhancement." In 2018 IEEE international conference on computational photography (ICCP), pp. 1-13. IEEE, 2018.

[13] Sheikh, Hamid R., and Alan C. Bovik. "Image information and visual quality." IEEE Transactions on image processing 15, no. 2 (2006): 430-444.

[14] Rehman, Abdul, Kai Zeng, and Zhou Wang. "Display device-adapted video quality-of-experience assessment." In Human vision and electronic imaging XX, vol. 9394, pp. 27-37. SPIE, 2015.

[15] Heusel, Martin, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. "Gans trained by a two time-scale update rule converge to a local nash equilibrium." Advances in neural information processing systems 30 2017.

[16] Salimans, Tim, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. "Improved techniques for training gans." Advances in neural information processing systems 29 (2016).

[17] Bińkowski, Mikołaj, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. "Demystifying mmd gans." arXiv preprint arXiv:1801.01401 (2018).

[18] H. R. Sheikh, A. C. Bovik and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," in IEEE Transactions on Image Processing, vol. 14, no. 12, pp. 2117-2128, Dec. 2005

[19] Ding, Keyan, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. "Image quality assessment: Unifying structure and texture similarity." IEEE transactions on pattern analysis and machine intelligence 44, no. 5 (2020): 2567-2581.

[20] Zhou, Fei, Rongguo Yao, Bozhi Liu, and Guoping Qiu. "Visual quality assessment for super-resolved images: Database and method." IEEE Transactions on Image Processing 28, no. 7 (2019): 3528-3541.

[21] Zhang, Keke, Tiesong Zhao, Weiling Chen, Yuzhen Niu, Jinsong Hu, and Weisi Lin. "Perception-Driven Similarity-Clarity Tradeoff for Image Super-Resolution Quality Assessment." IEEE Transactions on Circuits and Systems for Video Technology 34, no. 7 (2023): 5897-5907.

[22] Kang, Le, Peng Ye, Yi Li, and David Doermann. "Convolutional neural networks for no-reference image quality assessment." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1733-1740. 2014.

[23] Madhusudana, Pavan C., Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C. Bovik. "Image quality assessment using contrastive learning." IEEE Transactions on Image Processing 31 (2022): 4149-4161.

[24] Pan, Zhaoqing, Hao Zhang, Jianjun Lei, Yuming Fang, Xiao Shao, Nam Ling, and Sam Kwong. "DACNN: Blind image quality assessment via a distortion-aware convolutional neural network." IEEE Transactions on Circuits and Systems for Video Technology 32, no. 11 (2022): 7518-7531.

[25] Zhou, Zehong, Fei Zhou, and Guoping Qiu. "Blind image quality assessment based on separate representations and adaptive interaction of content and distortion." IEEE Transactions on Circuits and Systems for Video Technology 34, no. 4 (2023): 2484-2497.

[26] Golestaneh, S. Alireza, Saba Dadsetan, and Kris M. Kitani. "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.

[27] Yang, Sidi, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. "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.

[28] Qin, Guanyi, Runze Hu, Yutao Liu, Xiawu Zheng, Haotian Liu, Xiu Li, and Yan Zhang. "Data-efficient image quality assessment with attention-panel decoder." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 2, pp. 2091-2100. 2023.

[29] Sun, Wei, Xiongkuo Min, Danyang Tu, Siwei Ma, and Guangtao Zhai. "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, no. 6 (2023): 1178-1192.

[30] Gonzalez, Rafael C. Digital image processing. Pearson education india, 2009.

[31] Hore, Alain, and Djemel Ziou. "Image quality metrics: PSNR vs. SSIM." In 2010 20th international conference on pattern recognition, pp. 2366-2369. IEEE, 2010.

[32] Wang, Zhou, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. "Image quality assessment: from error visibility to structural similarity." IEEE transactions on image processing 13, no. 4 (2004): 600-612.

[33] Wang, Zhou, Eero P. Simoncelli, and Alan C. Bovik. "Multiscale structural similarity for image quality assessment." In The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, vol. 2, pp. 1398-1402. Ieee, 2003.

[34] Styan, George PH. "Hadamard products and multivariate statistical analysis." Linear algebra and its applications 6 (1973): 217-240.

[35] Lao, Zhiqiang, Yu Guo, Xiyun Song, Yubin Zhou, Zongfang Lin, Heather Yu, and Liang Peng. "High-Fidelity 4x Neural Reconstruction of Real-time Path Traced Images." In Proceedings of the Winter Conference on Applications of Computer Vision, pp. 157-166. 2025.

[36] Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. "Microsoft coco: Common objects in context." In Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13, pp. 740-755. Springer International Publishing, 2014.

[37] Lee, Seongmin, Benjamin Hoover, Hendrik Strobelt, Zijie J. Wang, ShengYun Peng, Austin Wright, Kevin Li, Haekyu Park, Haoyang Yang, and Duen Horng Polo Chau. "Diffusion explainer: Visual explanation for text-to-image stable diffusion." In 2024 IEEE Visualization and Visual Analytics (VIS), pp. 96-100. IEEE, 2024.

[38] Marcus, Gary, Ernest Davis, and Scott Aaronson. "A very preliminary analysis of DALL-E 2." arXiv preprint arXiv:2204.13807 (2022).

[39] Nichol, Alex, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. "Glide: Towards photorealistic image generation and editing with text-guided diffusion models." arXiv preprint arXiv:2112.10741 (2021).

[40] Betker, James, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang et al. "Improving image gen-

eration with better captions." Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf 2, no. 3 (2023): 8.
[41] Sedgwick, Philip. "Pearson' s correlation coefficient." Bmj 345 (2012).
[42] Sedgwick, Philip. "Spearman' s rank correlation coefficient." Bmj 349 (2014).
[43] Abdi, Hervé. "The Kendall rank correlation coefficient." Encyclopedia of measurement and statistics 2 (2007): 508-510.
[44] Saad, Michele A., Alan C. Bovik, and Christophe Charrier. "Blind image quality assessment: A natural scene statistics approach in the DCT domain." IEEE transactions on Image Processing 21, no. 8 (2012): 3339-3352.
[45] Mittal, Anish, Anush Krishna Moorthy, and Alan Conrad Bovik. "No-reference image quality assessment in the spatial domain." IEEE Transactions on image processing 21, no. 12 (2012): 4695-4708.
[46] Ghadiyaram, Deepti, and Alan C. Bovik. "Perceptual quality prediction on authentically distorted images using a bag of features approach." Journal of vision 17, no. 1 (2017): 32-32.
[47] Mittal, Anish, Rajiv Soundararajan, and Alan C. Bovik. "Making a "completely blind" image quality analyzer." IEEE Signal processing letters 20, no. 3 (2012): 209-212.
[48] Zhang, Lin, Lei Zhang, and Alan C. Bovik. "A feature-enriched completely blind image quality evaluator." IEEE Transactions on Image Processing 24, no. 8 (2015): 2579-2591.
[49] Wang, Zhou, Eero P. Simoncelli, and Alan C. Bovik. "Multiscale structural similarity for image quality assessment." In The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, vol. 2, pp. 1398-1402. Ieee, 2003.
[50] Chen, Chaofeng, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. "Topiq: A top-down approach from semantics to distortions for image quality assessment." IEEE Transactions on Image Processing (2024).
[51] Lao, Shanshan, Yuan Gong, Shuwei Shi, Sidi Yang, Tianhe Wu, Jiahao Wang, Weihao Xia, and Yujiu Yang. "Attentions help cnns see better: Attention-based hybrid image quality assessment network." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1140-1149. 2022.
[52] Sheikh, Hamid R., and Alan C. Bovik. "Image information and visual quality." IEEE Transactions on image processing 15, no. 2 (2006): 430-444.
[53] Wang, Shiqi, Kede Ma, Hojatollah Yeganeh, Zhou Wang, and Weisi Lin. "A patch-structure representation method for quality assessment of contrast changed images." IEEE Signal Processing Letters 22, no. 12 (2015): 2387-2390.
[54] Wang, Zhou, Eero P. Simoncelli, and Alan C. Bovik. "Multiscale structural similarity for image quality assessment." In The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, vol. 2, pp. 1398-1402. Ieee, 2003.
[55] Zhang, Lin, Lei Zhang, Xuanqin Mou, and David Zhang. "FSIM: A feature similarity index for image quality assessment." IEEE transactions on Image Processing 20, no. 8 (2011): 2378-2386.
[56] Zhang, Lin, Ying Shen, and Hongyu Li. "VSI: A visual saliency-induced index for perceptual image quality assessment." IEEE Transactions on Image processing 23, no. 10 (2014): 4270-4281.
[57] Zheng, Heliang, Huan Yang, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. "Learning conditional knowledge distillation for degraded-reference image quality assessment." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10242-10251. 2021.
[58] Zhou, Wei, and Zhou Wang. "Quality assessment of image super-resolution: Balancing deterministic and statistical fidelity." In Proceedings of the 30th ACM international conference on multimedia, pp. 934-942. 2022.
[59] Zhang, Lin, Lei Zhang, and Alan C. Bovik. "A feature-enriched completely blind image quality evaluator." IEEE Transactions on Image Processing 24, no. 8 (2015): 2579-2591.
[60] Sheikh, Hamid R., and Alan C. Bovik. "A visual information fidelity approach to video quality assessment." In The first international workshop on video processing and quality metrics for consumer electronics, vol. 7, no. 2, pp. 2117-2128. sn, 2005.