arXiv:2601.00225v1 [cs.CV] 1 Jan 2026

迈向合成到真实图像质量评估：重塑合成数据分布的新视角

Aobo Li Jinjian Wu* **YongxuLiu Leida Li Weisheng Dong**西安电子科技大学 abli@stu.xidian.edu.cn，{jinjian.wu, wsdong}@mail.xidian.edu.cn，{yongxu.liu, ldli}@xidian.edu.cn

摘要

盲图像质量评估（BIQA）借助深度学习取得了显著进展，但大规模带标签数据集的匮乏仍然是一项挑战。尽管合成数据提供了一种颇具前景的解决方案，但基于现有合成数据集训练的模型往往泛化能力有限。本文发现，从合成数据集中学习到的表征通常呈现离散的簇状模式，从而阻碍回归性能：高质量图像的特征聚集在参考图像周围，而低质量图像的特征则按照失真类型聚集。我们的分析表明，这一问题源于合成数据的分布，而非模型架构。因此，我们提出了一种新颖的框架 SynDR-IQA，通过重塑合成数据分布来增强盲图像质量评估的泛化能力。基于样本多样性和冗余对泛化误差影响的理论推导，SynDR-IQA 采用两种策略：分布感知的多样化内容上采样，在保留内容分布的同时提升视觉多样性；以及密度感知冗余簇下采样，通过减少密集聚集区域的样本来平衡样本分布。在三种跨数据集设置（合成到真实、合成到算法生成、合成到合成）下开展的大量实验，验证了我们方法的有效性。代码可在 https://github.com/Li-aobo/SynDR-IQA 获取。

1 引言

无参考图像质量评估（盲图像质量评估，BIQA）旨在不依赖参考图像的情况下，自动且准确地评估图像质量 [1, 2]。它在提升多媒体应用中的用户体验、增强下游图像处理算法的鲁棒性，以及指导图像增强方法的优化方面发挥着至关重要的作用。然而，由于任务本身复杂，且与人类感知高度相关，BIQA 任务颇具挑战。

近年来，得益于深度学习模型强大的表征能力，主流 BIQA 方法已大幅超越传统方法。然而，深度学习的成功在很大程度上依赖于大规模标注数据集。获取主观质量标签的高昂成本限制了现有数据集的扩展。合成失真中参考图像的可用性以及质量劣化的可控性表明，通过人工合成数据进行低成本数据增强似乎是一种可行的解决方案。实践中，直接使用现有合成失真数据集进行训练，得到的质量表征往往不理想，泛化能力也有限。

我们观察到一个关键现象：在合成数据上训练的模型往往会产生离散且呈簇状的特征分布。具体而言，如图 1 所示，高质量图像特征会形成

*通讯作者。

第 39 届神经信息处理系统大会（NeurIPS 2025）。

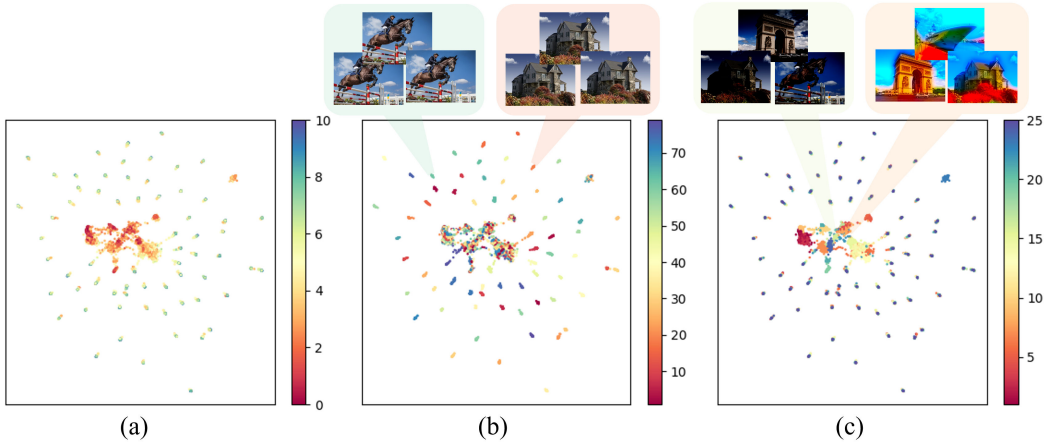


图 1：(a)、(b) 和 (c) 展示了基线模型从合成失真数据集 KADID-10k [4] 中学习到的相同表征的 UMAP [3] 可视化结果。三种可视化仅在颜色映射上有所不同：(a) 颜色表示平均意见分（MOS）值（数值越高表示质量越好）；(b) 颜色对应参考图像；(c) 颜色表示失真类型。(b) 和 (c) 中分别添加了具有代表性的样本，以说明高质量簇和低质量簇中的冗余。附录 A 中提供了不同骨干网络和数据集上的更多可视化结果，以证明这一现象的普遍性。

基于参考图像的不同簇，而低质量图像特征则会根据失真类型聚集。中等质量图像特征缺乏平滑过渡，且往往附着于高/低质量簇。这种不连续的表征不利于盲图像质量评估等回归任务的性能 [6, 7]。我们认为，这一现象主要由合成失真数据集中的两个核心问题导致：

- 1) **内容多样性不足**，这是由合成失真数据集中的参考图像数量有限造成的。这导致模型倾向于过拟合，阻碍了全局一致的质量表征的形成。

- 2)**冗余样本过多**，这是由于合成失真数据集中的失真图像均由参考图像、失真类型和失真强度的固定组合构成。这使模型过度关注这些重复模式，而忽略更广泛的信息，从而加剧了过拟合问题。

为深入理解这些问题，我们从理论上推导了样本多样性和冗余对泛化误差的影响。在此理论基础上，我们从全新的视角设计了一个名为 SynDR-IQA 的框架，通过重塑合成数据分布来提升盲图像质量评估的泛化能力。具体而言，**针对内容多样性不足的问题**，我们提出了分布感知多样内容上采样（DDCUp）策略。我们根据现有训练集的内容分布，从无标签候选参考集中采样参考图像以生成失真图像，从而增加视觉实例的多样性，帮助模型学习跨不同内容的一致表征。为了为新生成的失真图像赋予标签，我们采用了一个关键假设：在相同失真条件下，相似内容应产生相似的质量劣化。基于这一假设，我们参考训练集中与相似参考图像对应的给定标注数据，为新生成的图像生成伪标签，从而确保所生成伪标签的合理性和一致性。**针对冗余样本过多的问题**，我们设计了密度感知冗余簇下采样（DRCDown）策略。该策略识别训练数据集中的高密度冗余簇，并在保留低密度区域样本的同时，有选择地移除这些簇中的样本。这能够减轻冗余样本的负面影响，同时缓解数据分布不平衡，从而帮助模型学习更具泛化能力的表征。我们的贡献总结如下：

• 我们观察到一个关键现象：在合成数据上训练的模型呈现出离散且聚簇的特征分布，并对其潜在原因进行了深入分析。通过理论推导，我们证明了样本多样性和冗余对模型泛化误差的影响。

• 从重塑合成数据分布这一新颖视角出发，我们提出了 SynDR-IQA 框架，其中包括 DDCUp策略和 DRCDown策略，以增强盲图像质量评估模型的泛化能力。

• 在多种跨数据集设置下进行了大量实验，包括合成到真实、合成到合成以及合成到算法生成，验证了 SynDR-IQA 框架的有效性。此外，作为一种基于数据的方法，SynDR-IQA 可以与现有的基于模型的方法集成，而不会增加推理成本。

2 相关工作

基于深度学习的盲图像质量评估方法。 深度学习彻底改变了盲图像质量评估领域，在准确性和鲁棒性方面取得了显著进展 [8, 9]。近期研究探索了多种创新方法，以应对该领域的挑战。Zhu等人 [10] 提出了 MetaIQA，利用元学习提升模型在多种失真类型上的泛化能力。Su等人 [11] 提出了一种自适应超网络架构，用于真实场景下的自适应质量估计。Ke等人 [12] 开发了 MUSIQ，这是一种多尺度图像质量变换器，可处理尺寸各异的原生分辨率图像。Saha等人 [13] 提出了 Re-IQA，这是一种无监督混合专家方法，可联合学习互补的内容特征与质量特征，用于感知质量预测。Shin等人 [14] 提出了 QCN，利用比较变换器和分数枢轴提升跨数据集泛化能力。Xu等人 [15] 展示了将局部失真特征注入大型预训练视觉变换器对图像质量评估任务的有效性。尽管取得了这些进展，基于深度学习的盲图像质量评估方法仍严重依赖大规模标注数据集。获取真实世界图像主观质量标签的成本高昂且耗时，这显著限制了现有数据集的扩充。受此限制，研究人员开始探索利用合成失真来泛化至真实世界失真的潜力。

盲图像质量评估中的合成到真实泛化。由于合成失真与真实世界失真之间存在显著的域差异，在合成失真数据上训练的模型面对真实世界图像时往往表现不佳。为弥合这一差距，一些研究探索了无监督域适应（UDA）技术。Chen等人 [16] 提出了一种课程式无监督域适应方法，用于视频质量评估，使模型逐步从源域适应到目标域。Lu等人 [17] 提出了 StyleAM，在特征风格空间中对齐源域和目标域，而该空间与图像质量的关联更为紧密。Li等人 [18] 开发了 FreqAlign，从频域视角挖掘感知导向的可迁移性，并选择最佳频率分量进行对齐。最近，Li等人 [19] 提出了 DGQA，这是一种失真引导的无监督域适应框架，利用自适应多域选择匹配源域和目标域的数据分布，从而减少离群源域带来的负迁移。然而，以往的工作忽视了合成失真数据集的分布问题。在本研究中，我们提出了一种新颖的框架 SynDR-IQA，通过重塑合成数据的分布，增强盲图像质量评估的合成到真实泛化能力。

3 方法论

本节介绍 SynDR-IQA 框架，该框架旨在通过重塑合成数据分布，提升盲图像质量评估模型的泛化能力。我们首先对盲图像质量评估进行问题形式化，为本文奠定基础。随后，我们开展理论分析，探讨样本多样性和冗余对泛化误差的影响。在此理论基础上，我们详细介绍 SynDR-IQA 的两个核心组件：分布感知多样内容上采样（DDCUp），用于解决合成数据集多样性有限的问题；以及密度感知冗余簇下采样（DRCDown），用于缓解数据冗余问题。

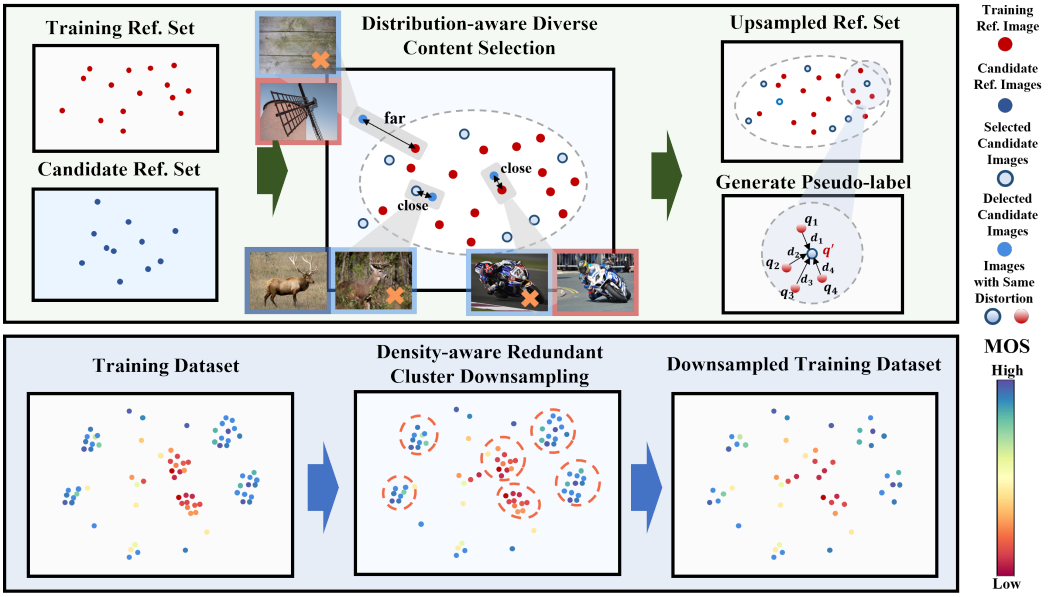


图 2：该图展示了 SynDR-IQA 中两种策略的核心概念。DDCUp策略（上方）从候选参考集中选择与训练参考集在分布上相似、但内容多样的图像，用于合成失真样本。这些样本的伪标签取决于其参考图像的最近邻。DRCDown策略（下方）识别训练数据集中的高密度簇，并有选择地移除这些簇中的样本，同时保留低密度区域中的样本。

3.1 问题形式化

盲图像质量评估旨在无需参考图像即可预测图像的感知质量。令 $\mathcal{X}$ 表示所有可能图像构成的空间，$\mathcal{Y} = [0, 1]$ 表示质量分数的取值范围（为简便起见）。盲图像质量评估任务可形式化为学习函数 f：$\mathcal{X} \to \mathcal{Y}$，将输入图像映射为其质量分数。最优函数 f^* 通过最小化期望风险来定义：$f^* = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}(f(x), y)]$，其中，$\mathcal{F}$ 为假设空间，$\mathcal{D}$ 为图像及其质量分数的真实分布，$\mathcal{L}$ 为损失函数。由于无法获取真实数据分布，实践中我们转而最小化训练数据集上的经验误差 $\hat{\mathcal{D}} = \{(x_i, y_i)\}_{i=1}^n$：$\hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \mathcal{L}(f(x_i), y_i)$。

3.2 理论分析

合成失真数据集的构建遵循系统化流程：对一组参考图像 [4] 施加预定义的失真类型和不同强度级别。这种生成机制具有两个关键特征：第一，低强度失真生成的图像与其参考图像几乎相同，而高强度失真生成的图像则主要呈现失真特定模式（代表性视觉示例见附录 B）。第二，数据集生成通常会对参考图像、失真类型和强度级别进行均匀采样。这两个特征共同导致一个根本问题：样本来自不同的局部分布，而非从真实分布中进行相同采样，因此在分布空间中形成离散的簇状结构。

为了理解这种聚类分布如何影响模型泛化，我们需要扩展经典的泛化误差分析，以考虑样本来自不同的局部分布，而非底层的真实分布。为形式化这一扩展，我们将合成数据集 $\hat{\mathcal{D}} = \{(x_i, y_i)\}_{i=1}^n$ 建模为包含 m 个簇，其中每个簇由一个来自真实分布 $\mathcal{D}$ 的 独立同分布 样本，以及与之相关的 $k_i - 1$ 个从相应局部分布 $\mathcal{D}_i \subset \mathcal{D}$ 中抽取的样本组成。这一形式化描述导出了以下泛化界：

定理 3.1（聚类合成数据的泛化界）。设 $\mathcal{F}$ 为函数的假设类 $f: \mathcal{X} \to \mathcal{Y}$，且 $\hat{\mathcal{D}} = \{(x_i, y_i)\}_{i=1}^{n}$ 为一个数据集，其中包含从真实分布 $\mathcal{D}$中独立同分布抽取的 m 个样本，以及分别从相应局部分布 $\mathcal{D}_i \subset \mathcal{D}$中抽取的 $k_i - 1$ 个冗余样本。以至少 $1-\delta$的概率，对所有 $f \in \mathcal{F}$，有：

$$|R(f) - R_{\text{emp}}(f)| \leq 2\,\text{Rad}_m(\mathcal{F}) + \sqrt{\frac{2\log(2/\delta)}{m}} + \sqrt{\frac{\eta\log(2/\delta)}{8m}} + \frac{2\log(2/\delta)}{3m}. \tag{1}$$

其中，$R(f)$是真实风险，$R_{emp}(f)$是经验风险，$\text{Rad}_m(\mathcal{F})$是拉德马赫复杂度，而 $\eta = \frac{1}{m}\sum_{i=1}^{m} k_i^{-1}$ 定义为冗余异质性，用于量化冗余样本分布不均衡的程度。

根据定理 3.1，我们可以观察到，泛化误差上界受 Rademacher复杂度 $\text{Rad}_m(\mathcal{F})$、独立同分布样本的数量 m（也称为多样样本）以及冗余异质性 η的影响。排除模型相关因素（由 $\text{Rad}_m(\mathcal{F})$ 刻画）后，该界表明，不同样本的数量 m 在决定上界方面起主要作用。增加多样样本的采样量可以有效降低这一上界。通过平衡样本以减少冗余异质性 η，也能有效降低泛化误差上界。然而，盲目增加样本可能会导致来自局部分布的冗余样本增多，进而造成更高的冗余异质性 η，放大第三项，从而降低整体泛化性能。

这些理论洞见促使我们从**增加内容多样性**（增加 m）和**平衡样本密度**（减少 η）的角度重塑样本分布，从而提升盲图像质量评估的泛化性能。

3.3 SynDR-IQA 框架

基于这些理论洞见，我们提出了如图 2 所示的 SynDR-IQA 框架，该框架包含两种主要策略：分布感知的多样内容上采样和密度感知冗余簇下采样。这些策略共同旨在重塑合成数据分布，以获得具有更强泛化能力的盲图像质量评估模型。

3.3.1 分布感知多样内容上采样

我们的 DDCUp策略旨在通过引入多样化的视觉内容来丰富训练数据，同时保持原始内容分布。

算法 1 分布感知的多样内容上采样策略 要求：训练数据集 $\mathcal{D}$、训练参考集 $\mathcal{D}_r$、候选参考集 $\mathcal{D}_c$、特征提取器$f(\cdot)$、距离度量 Dist$(\cdot)$1 $\mathcal{D}_r^{new} \leftarrow \emptyset$：初始化：$DistT \leftarrow \{$ Dist$(f(x_r^i)\ f(x_r^j))|x_r^i$、$x_r^j \in \mathcal{D}_r$、$i \neq j\}$3：[2]对于每个 $x_c \in \mathcal{D}_c$ 执行 4：$DistC \leftarrow$Dist$(f(x_c)\ f(\mathcal{D}_r))$ 5：如果 Min$(DistC)$ $>$Median$(DistT)$ 且 Max$(DistC)$ $<$Max$(DistT)$，则 6：$DistN \leftarrow$Dist$(f(x_c)\ f(\mathcal{D}_r^{new}))$ 7：如果 $\mathcal{D}_r^{new}$为 $\emptyset$或 Min$(DistN)$ $>$Median$(DistT)$，则 8：$\mathcal{D}_r^{new} \leftarrow \mathcal{D}_r^{new} \cup \{x_c\}$9：结束 if 10：结束 if 11：结束 for 12：$\mathcal{D}' \leftarrow \mathcal{D} \cup$GenSyn$(\mathcal{D}, \mathcal{D}_r, \mathcal{D}_r^{new}, ...)$ 13：返回上采样数据集 $\mathcal{D}'$

算法 1详细介绍了DDCUp策略。我们从 KADIS-700k[4] 中选择额外的参考图像，以构建候选参考集 $\mathcal{D}_c$。为避免引入过多噪声，我们将其规模限制为

与训练参考集相同。使用在 ImageNet 上预训练的特征提取器 $f(\cdot)$，提取参考图像的特征。我们采用距离度量 Dist(·)（在我们的实现中为余弦距离）来指导选择过程。我们的算法确保所选图像不会过度偏离原始内容分布，同时彼此之间保持足够的差异（第 5 行和第 7 行），从而促进多样性并防止冗余。附录 D 中的定性分析通过可视化直观地展示了这些效果。

算法 2合成数据生成

输入： 训练数据集 $\mathcal{D}$、训练参考集 $\mathcal{D}_r$、新参考集 $\mathcal{D}_r^{new}$、特征提取器 $f(\cdot)$、距离度量 Dist(·)、最近邻数量 k、参考特征距离阈值 T_{rf}、失真算法 GenDist(·)、失真类型数量 T、失真级别数量 L

1: 初始化 $\mathcal{D}_{GenSyn} \leftarrow \emptyset$
2: **对于**每个 $x_{new} \in \mathcal{D}_r^{new}$ **do**
3: $nns \leftarrow$kNN($f(x_{new})$ $f(\mathcal{D}_r)$ k,Dist) {k 近邻 of x_{new}}
4: $nns \leftarrow \{nn \mid$分布($f(x_r^{nns[0]})$ $f(x_{new})$) $-$分布($f(x_r^{nn})$ $f(x_{new})$) $< T_{rf}, nn \in nns\}$
5: $w \leftarrow$Softmax({Dist($f(x_r^{nn})$ $f(x_{new})$) $\mid nn \in nns\}$)
6: **对于**每个 $t \in T$, $l \in L$**执行**
7: $x_{new}^{(t,l)} \leftarrow$ GenDist(x_{new}, t, l) $y_{new}^{(t,l)} \leftarrow \sum_{nn \in nns} w_{nn} y_{nn}^{(t,l)}$
8: $\mathcal{D}_{GenSyn} \leftarrow \mathcal{D}_{GenSyn} \cup \{(x_{new}^{(t,l)}, \ y_{new}^{(t,l)})\}$
9: **结束循环**
10: **结束循环**
11: **返回**生成的合成数据集 $\mathcal{D}_{GenSyn}$

选出多样化的参考图像后，我们生成相应的失真图像和伪标签，以扩充训练数据集。算法 2 详细说明了这一过程。对于每张选出的参考图像，我们采用与创建训练数据集时相同的失真生成过程。值得注意的是，为减少冗余样本的增加并防止引入额外的标签噪声，我们仅采用等级为 1、3 和 5 的失真强度。为这些新生成的失真图像生成可靠的伪标签，我们利用以下假设：在相同失真条件下，相似内容应产生相似的质量劣化。对于每张新的参考图像 x_{new}，我们根据特征距离，在原始训练集的参考图像中找出其 k 个最近邻（kNN）。为进一步提高伪标签的可靠性，我们对这些最近邻进行筛选，仅保留特征距离 x_{new} 不超过特定阈值 T_{rf}（设为 0.05）的邻居（第 4 行）。随后，将 x_{new} 的每张失真图像的伪标签计算为其最近邻对应失真图像标签的加权平均值，其中权重由特征距离的 Softmax 值确定。

3.3.2 密度感知的冗余簇下采样

冗余样本过多可能会使模型产生偏差，阻碍其对未见数据的泛化能力。这也会加剧冗余异质性（定理 3.1 中的 η），进一步增大泛化误差。因此，我们提出 DRCDown策略，以减轻冗余样本带来的负面影响并降低 η，从而进一步提升模型的泛化性能。

算法 3 详细介绍了 DRCDown策略。与 DDCUp 不同，该策略会在每个训练轮次开始前提取特征，以确保减少失真级冗余，同时高效利用训练数据。我们根据距离度量 Dist(·) 和标签距离（在我们的实现中为 L1）识别相似样本对（第 3 行）。这样既能消除冗余，又不会丢弃特征差异很小但质量差异很大的困难样本。失真特征距离阈值 T_{df} 和真实值距离阈值 T_g 分别设为 0。1 和 1（针对 $[0, 10]$ 中的 MOS 值）。通过同时考虑特征相似性和标签相似性，我们旨在有针对性地降低对冗余异质性影响显著的高密度簇的密度。

在识别出相似样本对后，我们采用并查集（DSU）数据结构将这些样本对分组为簇（第 7 行）。对于大小大于 $2T_u$ 的每个簇，我们将其随机下采样为最多包含 $\lfloor N_u/2 \rfloor$，T_u 个样本，其中 N_u 是原始簇大小（第 10 行）。这

算法 3 密度感知冗余簇下采样策略 要求：训练数据集 $\mathcal{D}$、特征提取器 $f(\cdot)$、距离度量 $Dist(\cdot)$、失真特征距离阈值 T_{df}、真实值距离阈值 T_g、最小并集大小阈值T_u1：初始化 $\mathcal{D}_{down} \leftarrow \emptyset$、 $SimPairs \leftarrow \emptyset$2：对于每个 (x_i)、$(x_j, y_j) \in \mathcal{D}$ 执行 3：如果 Dist($f(x_i)$ $f(x_j)$) $< T_{df}$且 $|y_i - y_j| < T_g$则 4：$SimPairs \leftarrow SimPairs \cup \{(x_i, x_j)\}$5：结束 if 6：结束 for 7：$Unions \leftarrow$DSU($SimPairs$) {合并相似样本对的不相交集合} 8：对于每个 $u \in Unions$ 执行 9：如果 Length(u)$/2 > T_u$ 则 10：从并集中随机选择 Max($\lfloor$Length(u)$/2 \rfloor, T_u$) 个样本} 11：结束 if 12：$\mathcal{D}_{down} \leftarrow \mathcal{D}_{down} \cup u$13：结束 for 14：返回下采样数据集 $\mathcal{D}_{down}$

阈值 T_u 可防止过度下采样，确保下采样数据集保留足够的信息以进行有效训练。通过有选择地移除过度集中的区域中的样本，DRCDown策略有效减少了冗余，并促进数据分布更加均衡，直接解决了冗余异质性高的问题，从而有助于提升泛化性能。这种冗余的减少有助于降低定理 3.1 所确立的泛化误差界，进而得到更稳健、泛化能力更强的IQA模型。

4 实验

4.1 实验设置

数据集与协议。 我们在八个IQA数据集上开展实验：四个合成失真数据集LIVE [20], CSIQ [21], TID2013 [22],KADID-10k [4],、三个真实失真数据集LIVEC [23], KonIQ-10k [24], BID [25], , 以及同时包含合成失真和算法失真的数据集PIPAL [26] 。我们采用斯皮尔曼等级相关系数（SRCC）和皮尔逊线性相关系数（PLCC）评估模型性能。两个系数的取值范围均为-1至1，数值越接近1，表示性能越好。

实现细节。 在我们的实验中，我们使用与 DGQA [19] 相同的模型架构（ResNet-50 [5]）和损失函数（L1损失）。对于合成到真实和合成到算法生成设置，模型使用 DGQA 选出的失真类型进行训练；对于合成到合成设置，则使用所有失真类型。遵循标准的 IQA协议 [27, 28], ，我们按参考图像以 80/20 的比例划分数据集，用于数据集内实验，并重复十次，报告 SRCC/PLCC 的中位数。对于跨数据库实验，模型在 KADID-10k 上训练，并在其他数据集上测试。训练期间，从每张图像中随机采样一个 224×224 图像块，并随机进行水平翻转以实现数据增强。小批量大小设为 32，学习率为 2×10^{-5}。我们使用权重衰减为 5×10^{-4} 的 Adam优化器对模型进行 24 个周期的优化。测试期间，对每张图像的五个图像块的预测结果取平均值，作为最终输出。所有实验均使用 PyTorch 实现，并在一台配备 2.10GHz Intel Xeon(R) CPU E5-2620 v4 处理器和四块 NVIDIA GTX 1080 Ti GPU 的服务器上运行。

4.2 性能评估

在从合成到真实场景下的表现。 我们首先评估 SynDR-IQA 从合成失真迁移到真实失真时的泛化能力。SynDR-IQA 与最先进方法的比较结果汇总于表1。SynDR-IQA 在全部三个真实数据集上均取得领先表现，仅在 FreqAlign（KADID-10k→KonIQ-10k，SRCC）和 Q-Align（KADID-10k→LIVEC，PLCC）上略逊一筹。与次优方法 DGQA 相比，我们的方法在三个数据集上的平均 SRCC 和 PLCC 分别提升了

表1：从合成到真实场景下的性能比较（KADID-10k→LIVEC、KonIQ-10k 和 BID）。

方法	LIVEC		KonIQ-10k		BID		平均值	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
RankIQA [29]	0.491	0.495	0.603	0.551	0.510	0.367	0.535	0.471
DBCNN [30]	0.572	0.589	0.639	0.618	0.620	0.609	0.613	0.606
HyperIQA [11]	0.490	0.487	0.545	0.556	0.379	0.282	0.472	0.442
MUSIQ [12]	0.517	0.524	0.554	0.573	0.575	0.600	0.549	0.566
VCRNet [31]	0.561	0.548	0.517	0.525	0.542	0.545	0.540	0.540
KGANet [32]	0.575	-	0.528	-	-	-	-	-
CLIPIQA+ [33]	0.512	0.543	0.511	0.515	0.474	0.442	0.499	0.500
Q-Align [34]	0.702	**0.744**	0.668	0.665	-	-	-	-
DANN [35]	0.499	0.484	0.638	0.636	0.586	0.510	0.574	0.543
UCDA [16]	0.382	0.358	0.496	0.501	0.348	0.391	0.408	0.417
RankDA [36]	0.451	0.455	0.638	0.623	0.535	0.582	0.542	0.553
StyleAM [17]	0.584	0.561	0.700	0.673	0.637	0.567	0.640	0.600
FreqAlign [18]	0.618	0.588	**0.748**	0.721	0.674	0.708	0.680	0.673
DGQA [19]	0.696	0.690	0.681	0.687	0.770	0.753	0.716	0.710
SynDR-IQA	**0.713**	0.714	0.727	**0.735**	**0.788**	**0.764**	**0.743**	**0.737**

表2：合成到算法生成（KADID-10k→在PIPAL上的算法失真）设置下的性能比较。

失真类型	DGQA		SynDR-IQA	
	SRCC	PLCC	SRCC	PLCC
传统超分辨率	0.4897	0.4567	$\mathbf{0.5247}_{+3.50\%}$	$\mathbf{0.4808}_{+2.41\%}$
由PSNR引发的超分辨率	0.5419	0.5404	$\mathbf{0.5845}_{+4.26\%}$	$\mathbf{0.5680}_{+2.76\%}$
存在核失配的超分辨率	0.5810	0.5956	$\mathbf{0.6263}_{+4.53\%}$	$\mathbf{0.6342}_{+3.86\%}$
基于生成对抗网络的超分辨率	0.1629	0.1353	$\mathbf{0.1998}_{+3.69\%}$	$\mathbf{0.1629}_{+2.76\%}$
去噪	0.5393	0.5279	$\mathbf{0.5749}_{+3.56\%}$	$\mathbf{0.5552}_{+2.73\%}$
超分辨率和联合去噪	**0.5588**	**0.5193**	$0.5557_{-0.31\%}$	$0.5023_{-1.70\%}$
平均值	0.4789	0.4625	$\mathbf{0.5110}_{+3.21\%}$	$\mathbf{0.4839}_{+2.14\%}$

2.7% 和 2.7%。这一显著提升表明 SynDR-IQA 具有出色的合成到真实泛化能力。

合成到算法失真设置下的性能。 我们进一步评估了 SynDR-IQA 在合成到算法失真设置下的表现。表2比较了我们的方法与 DGQA 在不同类型算法失真上的表现。对于大多数失真类型，SynDR-IQA 的表现始终优于 DGQA。值得注意的是，在由PSNR引发的超分辨率和存在核失配的超分辨率任务中，其 SRCC 提高了超过 4%。仅在超分辨率和联合去噪任务中观察到轻微下降。总体而言，这些结果表明 SynDR-IQA 能够有效泛化至算法失真，展现出处理复杂且未见失真类型的鲁棒性。

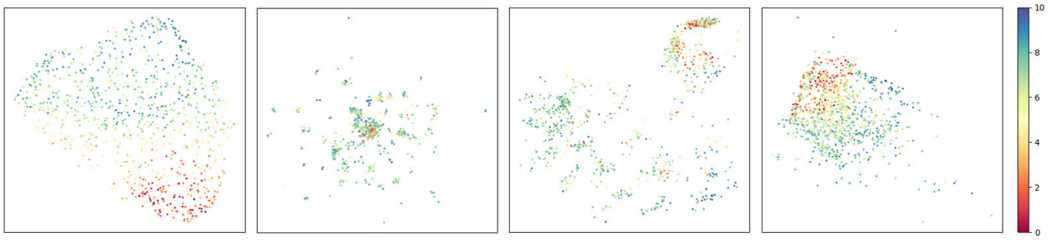
合成到合成设置下的性能。 我们还评估了 SynDR-IQA 在合成到合成设置下的表现。SRCC 结果汇总于表3。无论是在数据集内还是跨数据集评估中，SynDR-IQA 的表现均优于基线。在 KADID-10k 上，我们的方法取得了 0.8922 的 SRCC，表明其对训练数据的拟合效果更好。在跨数据集测试中，SynDR-IQA 仍保持较高的性能，展现出对其他合成失真数据集更强的泛化能力。这些结果证实，我们的方法能够使模型学习到更鲁棒、泛化性更强的特征表示。

4.3 消融研究

为了了解 SynDR-IQA 中各组件的作用，我们在从合成到真实场景下进行了消融实验（SRCC 见表4）。此处，*CD*表示训练时加入完整的候选数据集，*CD+SEL* 表示 DDCUp，而 *DOWN* 表示 DRCDown。

表3：合成到合成设置下的性能比较（在 KADID-10k、KADID-10k→LIVE、CSIQ 和 TID2013 上进行单数据库评估）。

方法	KADID-10k		LIVE		CSIQ		TID2013	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
基线	0.8528	0.8526	0.9173	0.8988	0.7965	0.8017	0.7077	0.7220
SynDR-IQA	**0.8922**	**0.8974**	**0.9258**	**0.9014**	**0.8069**	**0.8092**	**0.7147**	**0.7328**



(a) (b) (c) (d)

图 3：使用同一模型在四种不同训练流程下，从 LIVEC 提取的特征的 UMAP 可视化：(a) 直接在LIVEC上训练，(b) 直接在 KADID-10k 上训练，(c) 基于 DGQA 在 KADID-10k 上训练，(d) 基于 SynDR-IQA 在 KADID-10k 上训练。

候选数据集的影响。 比较**a)** 和**b)**可见，KonIQ-10k 上的性能有所提升，但在 LIVEC 和 BID 上略有下降。这表明，盲目增加数据以丰富内容多样性未必有益，因为这也可能引入与训练数据集分布差异显著的样本以及新的冗余样本，从而阻碍模型泛化。

DDCUp 的效果。 比较 **a)**、**b)**和**e)**可见，DDCUp 显著提升了性能。DDCUp 在保持数据分布的同时有选择地提高多样性，从而有效平衡内容多样性与分布一致性，提升模型在不同数据集上的泛化能力。

DRCDown 的效果。 比较 **a)** 和**c)**可见，仅使用 DRCDown 就已展现出相较于基线的稳定提升。这表明，控制样本密度能够解决数据冗余和不平衡问题，从而获得更精确、泛化能力更强的特征表示。

所有组件组合而成的 SynDR-IQA 整体性能最佳，其平均 SRCC 较基线提升了 2.71%。这些结果表明，SynDR-IQA 中提出的每个组件都有助于构建更稳健、泛化能力更强的模型。

4.4 可视化分析

我们采用四种不同的训练流程实现相同的模型：（a）直接在 LIVEC 上训练；（b）直接在 KADID-10k 上训练；（c）基于 DGQA 在 KADID-10k 上训练；（d）基于 SynDR-IQA 在 KADID-10k 上训练。随后，我们使用这些模型从 LIVEC 中提取特征并进行 UMAP 可视化，如图 3 所示。可视化结果清楚地表明，直接在 KADID-10k 上训练的模型所得到的表征形成了明显且分散的簇，说明其对真实失真的泛化能力较差。DGQA 得到的表征有所改善，但仍较为分散。相比之下，SynDR-IQA 生成了连续平滑的特征模式，与直接在 LIVEC 上训练的模型所得到的结果更加接近。这从视觉上验证了 SynDR-IQA 在弥合合成域与真实域之间的差距方面的有效性。

5 结论

在本工作中，我们旨在解决在合成数据集上训练的盲图像质量评估模型泛化能力有限这一关键挑战。我们的研究揭示了一个重要规律：从合成数据集中学习到的表征往往形成离散且聚簇的分布，其中高质量图像特征

表4：SynDR-IQA 各组件在从合成到真实场景下的消融研究。符号 ✓和 表示包含某个组件。

索引	*CD*	*SEL*	*DOWN*	LIVEC	KonIQ-10k	BID	平均值
a)				0.6958	0.6810	0.7696	0.7155
b)	✓			0.6901	0.7105	0.7677	0.7228
c)			✓	0.6962	0.6887	0.7793	0.7214
d)	✓		✓	0.7095	0.7107	0.7775	0.7326
e)	✓	✓		**0.7219**	0.7119	0.7822	0.7387
f)	✓	✓	✓	0.7127	**0.7268**	**0.7884**	**0.7426**

围绕参考图像聚类，而低质量特征则根据失真类型聚类，这严重阻碍了模型对真实失真或未见失真的泛化。受此启发，我们从理论上分析了样本多样性和冗余对泛化误差的影响。我们的理论洞见构成了 SynDR-IQA 的基础。这是一种旨在重塑合成数据分布、提升盲图像质量评估模型泛化能力的新型框架。SynDR-IQA 采用两项关键策略：1）DDCUp，在保持训练数据集内容分布的同时增强内容多样性；2）DRCDown，通过降低密集簇的密度来优化样本分布。在三种跨数据集设置下开展的全面实验一致表明，采用 SynDR-IQA 框架训练的模型具备更强的泛化能力。

致谢与资金披露

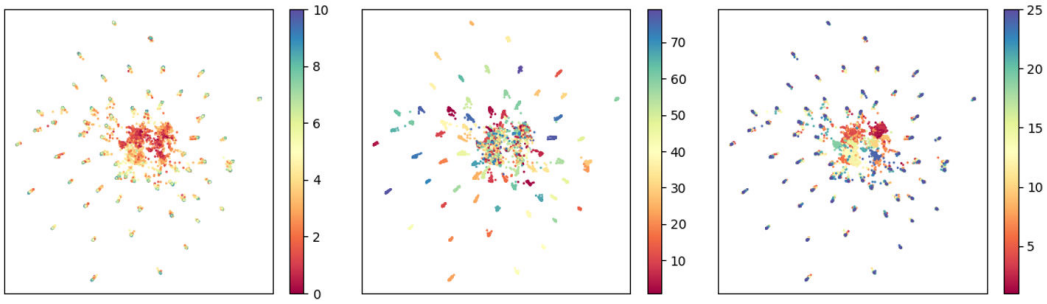
本研究部分受到国家重点研发计划（2023YFA1008500）、国家自然科学基金（62401420）以及浙江省自然科学基金杭州联合基金（LHZSZ25F010003）的资助。

参考文献

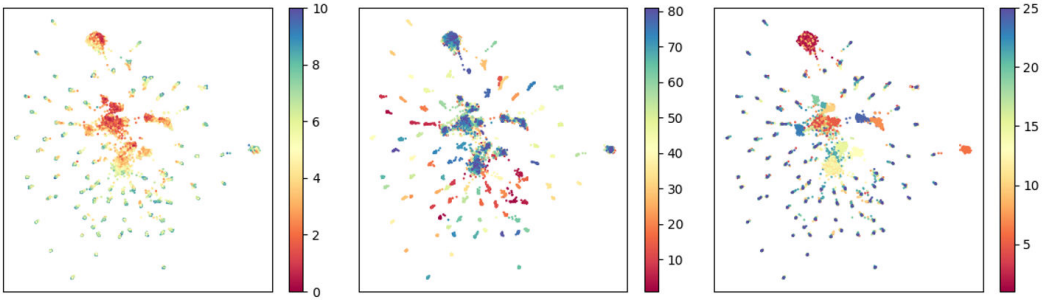
[1] Qiaohong Li、Weisi Lin、Jingtao Xu 和 Yuming Fang。利用统计结构与亮度特征进行无参考图像质量评估。*IEEE*多媒体汇刊，18(12):2457–2469，2016年。[2] Bo Yan、Bahetiyaer Bare 和 Weimin Tan。自然度感知的深度无参考图像质量评估。*IEEE*多媒体汇刊，21(10):2603–2615，2019年。[3] Leland McInnes、John Healy 和 James Melville。UMAP：用于降维的统一流形逼近与投影。*arXiv*预印本 *arXiv:1802.03426*，2018年。[4] Hanhe Lin、Vlad Hosu 和 Dietmar Saupe。KADID-10k：大规模人工失真图像质量评估数据库。载于*2019*年第十届国际多媒体体验质量会议，第1–3页。IEEE，2019年。[5] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。用于图像识别的深度残差学习。载于*IEEE*计算机视觉与模式识别会议论文集，第770–778页，2016年。[6] Kaiwen Zha、Peng Cao、Yuzhe Yang 和 Dina Katabi。监督式对比回归。*arXiv*预印本 *arXiv:2210.01189*，2022年。[7] Aobo Li、Jinjian Wu、Yongxu Liu、Leida Li、Weisheng Dong 和 Guangming Shi。基于感知比较的无参考图像质量评估。*IEEE*多媒体汇刊，2024年。[8] Le Kang、Peng Ye、Yi Li 和 David Doermann。用于无参考图像质量评估的卷积神经网络。载于*IEEE*计算机视觉与模式识别会议论文集，第1733–1740页，2014年。[9] Jongyoo Kim 和 Sanghoon Lee。完全深度无参考图像质量预测器。*IEEE*信号处理专题选刊，11(1):206–220，2016年。

[10] Hancheng Zhu、Leida Li、Jinjian Wu、Weisheng Dong 和 Guangming Shi。MetaIQA：用于无参考图像质量评估的深度元学习。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第14143–14152页，2020年。[11] Shaolin Su、Qingsen Yan、Yu Zhu、Cheng Zhang、Xin Ge、Jinqiu Sun 和 Yanning Zhang。在自适应超网络引导下对真实场景中的图像质量进行无参考评估。载于*IEEE*计算机视觉与模式识别会议论文集，第3667–3676页，2020年。[12] Junjie Ke、Qifei Wang、Yilin Wang、Peyman Milanfar 和 Feng 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，第5148–5157页，2021年。[13] Avinab Saha、Sandeep Mishra 和 Alan C. Bovik。Re-IQA：用于真实场景图像质量评估的无监督学习。载于*2023*年*IEEE/CVF*计算机视觉与模式识别会议，第5846–5855页，2023年。[14] Nyeong-Ho Shin、Seon-Ho Lee 和 Chang-Su Kim。基于几何序学习的无参考图像质量评估。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第12799–12808页，2024年。[15] Kangmin Xu、Liang Liao、Jing Xiao、Chaofeng Chen、Haoning Wu、Qiong Yan 和 Weisi Lin。通过具有局部特征增强的高效变换器适配提升图像质量评估。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第2662–2672页，2024年。[16] Pengfei Chen、Leida Li、Jinjian Wu、Weisheng Dong 和 Guangming Shi。用于无参考视频质量评估的无监督课程域适应。载于*IEEE/CVF*国际计算机视觉会议，第5178–5187页，2021年。[17] Yiting Lu、Xin Li、Jianzhao Liu 和 Zhibo Chen。StyleAM：面向感知的无监督域适应，用于非参考图像质量评估。*arXiv*预印本 *arXiv:2207.14489*，2022年。[18] Xin Li、Yiting Lu 和 Zhibo Chen。FreqAlign：从频域视角挖掘无参考图像质量评估中的面向感知的可迁移性。*IEEE*多媒体汇刊，2023年。[19] Aobo Li、Jinjian Wu、Yongxu Liu 和 Leida Li。弥合合成到真实的差距：用于无参考图像质量评估的失真引导的无监督域适应。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第28422–28431页，2024年。[20] Hamid R Sheikh、Muhammad F Sabir 和 Alan C Bovik。对近期全参考图像质量评估算法的统计评估。*IEEE*图像处理汇刊，15(11):3440–3451，2006年。[21] Larson Damon M. Chandler。最明显失真：全参考图像质量评估与策略的作用。《电子成像杂志》，19(1):011006，2010年。[22] Nikolay Ponomarenko、Oleg Ieremeiev、Vladimir Lukin、Karen Egiazarian、Lina Jin、Jaakko Astola、Benoit Vozel、Kacem Chehdi、Marco Carli 以及 Federica 和 Battisti。TID2013彩色图像数据库：特性与初步结果。载于欧洲视觉信息处理研讨会，第106–111页，2013年。[23] Deepti Ghadiyaram 和 Alan C Bovik。关于主观与客观图像质量的大规模在线众包研究。*IEEE*图像处理汇刊，25(1):372–387，2015年。[24] Vlad Hosu、Hanhe Lin、Tamas Sziranyi 和 Dietmar Saupe。KonIQ-10k：用于深度学习无参考图像质量评估的生态效度数据库。*IEEE*图像处理汇刊，29:4041–4056，2020年。[25] Alexandre Ciancio、Eduardo AB da Silva、Amir Said、Ramin Samadani、Pere Obrador 等。基于多特征分类器的数字图像无参考模糊评估。*IEEE*图像处理汇刊，20(1):64–75，2010年。

[26] Jinjin Gu、Haoming Cai、Haoyu Chen、Xiaoxing Ye、Jimmy Ren Ren 和 Chao Dong。PIPAL：用于感知图像复原的大规模图像质量评估数据集。发表于欧洲计算机视觉会议，第 633–651 页，Cham，2020年。Springer International Publishing。[27] Xudong Li、Timin Gao、Runze Hu、Yan Zhang、Shengchuan Zhang、Xiawu Zheng、Jingyuan Zheng、Yunhang Shen、Ke Li、Yutao Liu、Pingyang Dai 和 Rongrong Ji。通过缓解语义噪声敏感性，实现无参考图像质量评估的自适应特征选择。发表于第*41*届国际机器学习会议论文集，ICML'24。JMLR.org，2024年。[28] Yan Zhong、Xingyu Wu、Li Zhang、Chenxi Yang 和 Tingting Jiang。Causal-iqa：基于因果推断，迈向图像质量评估的泛化。发表于第*41*届国际机器学习会议，2024年。[29] Xialei Liu、Joost van de Weijer 和 Andrew D Bagdanov。RankIQA：通过排序学习进行无参考图像质量评估。发表于*IEEE*计算机视觉与模式识别会议论文集，第 1040–1049 页，2017年。[30] Weixia Zhang、Kede Ma、Jia Yan、Dexiang Deng 和 Zhou Wang。利用深度双线性卷积神经网络进行无参考图像质量评估。*IEEE*视频技术电路与系统汇刊，30(1)：36–47，2018年。[31] Zhaoqing Pan、Feng Yuan、Jianjun Lei、Yuming Fang、Xiao Shao 和 Sam Kwong。VCRNet：用于无参考图像质量评估的视觉补偿复原网络。*IEEE*图像处理汇刊，31：1613–1627，2022年。[32] Tianwei Zhou、Songbai Tan、Baoquan Zhao 和 Guanghui Yue。用于无参考图像质量评估的知识引导注意力多任务深度神经网络。*IEEE*视频技术电路与系统汇刊，2024年。[33] Jianyi Wang、Kelvin CK Chan 和 Chen Change Loy。探索 CLIP 对图像观感的评估能力。发表于*AAAI*人工智能会议论文集，第 37 卷，第 2555–2563 页，2023年。[34] Haoning Wu、Zicheng Zhang、Weixia Zhang、Chaofeng Chen、Liang Liao、Chunyi Li、Yixuan Gao、Annan Wang、Erli Zhang、Wenxiu Sun、Qiong Yan、Xiongkuo Min、Guangtao Zhai 和 Weisi Lin。Q-Align：通过离散文本定义等级，教授大语言模型进行视觉评分。发表于第*41*届国际机器学习会议论文集，第 54015–54029 页，2024年。[35] Hana Ajakan、Pascal Germain、Hugo Larochelle、François Laviolette 和 Mario Marchand。领域对抗神经网络。*arXiv*预印本 *arXiv:1412.4446*，2014年。[36] Baoliang Chen、Haoliang Li、Hongfei Fan 和 Shiqi Wang。利用无监督域适应进行无参考屏幕内容图像质量评估。*IEEE*图像处理汇刊，30：5463–5476，2021年。[37] Karen Simonyan 和 Andrew Zisserman。用于大规模图像识别的超深卷积网络。*arXiv*预印本 *arXiv:1409.1556*，2014年。[38] Ze Liu、Yutong Lin、Yue Cao、Han Hu、Yixuan Wei、Zheng Zhang、Stephen Lin 和 Baining Guo。Swin Transformer：采用移位窗口的层次化视觉 Transformer。发表于*IEEE/CVF*国际计算机视觉会议论文集，第 10012–10022 页，2021年。[39] Alexey Dosovitskiy。一张图像胜过 16×16 个词：面向大规模图像识别的 Transformer。*arXiv*预印本 *arXiv:2010.11929*，2020年。[40] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。通过自然语言监督学习可迁移的视觉模型。发表于国际机器学习会议，第 8748–8763 页。PMLR，2021年。



(a) VGG-16



(b) Swin-T

图 4：VGG-16 [37] 和 Swin-T [38] 在 KADID-10k [4]上学习到的特征的 UMAP [3] 可视化结果。两者在合成失真数据上均呈现离散的聚类模式。

不同骨干网络和数据集上的 UMAP 可视化结果

为进一步验证图1中观察到并在引言中讨论的现象是否具有普遍性，我们对不同骨干网络和数据集上学习到的特征分布进行了可视化。图4–6展示了在各种实验条件下学习到的特征的 UMAP [3] 投影。

图 4 展示了在 KADID-10k [4]上训练的 VGG-16 [37] 和 Swin Transformer Tiny（Swin-T） [38] 的结果。两种模型在合成失真上均呈现离散且聚集的特征结构：高质量样本按参考图像聚类，低质量样本则按失真类型聚类。图 5 展示了在 TID2013 [22], 上训练的 ResNet-50 [5] 呈现的类似模式，证实这一现象在不同合成数据集上具有一致性。相比之下，图 6 展示了在真实图像质量评价数据集（LIVEC [23], 、BID [25], 和 KonIQ-10k [24]）上训练的 ResNet-50 [5] 的 UMAP 结果，其中特征分布更加连续、平滑。

这些差异源于数据分布本身。受内容多样性有限、失真组合冗余等因素的制约，合成失真数据集往往会使模型的特征空间过度聚类且不连续。真实数据集的数据分布则更加均衡且具有内在多样性，使模型能够学习到更加连续、语义连贯且具有泛化能力的表征。

B KADID-10k 的视觉示例

为进一步说明正文中讨论的分布特征，我们展示了 KADID-10k 数据集中的若干代表性视觉示例 [4]。

图7展示了部分低失真图像，它们均源自同一参考图像（I01.png）。这些图像具有极高的视觉相似性，与原始参考图像几乎无法区分。这表明，低强度失真对图像视觉外观的影响极小，导致样本之间存在大量冗余。

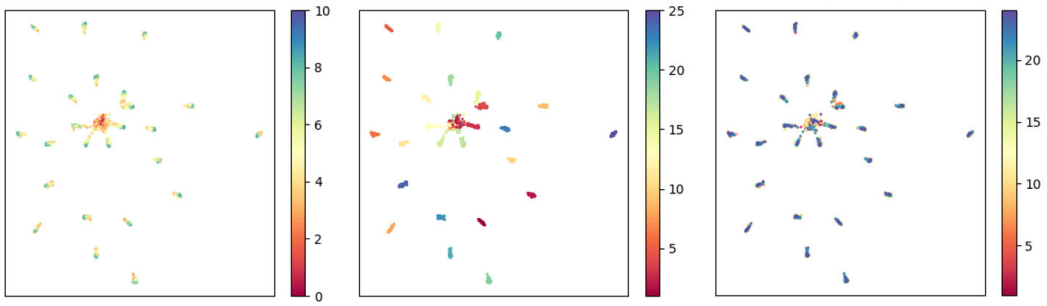


图5：在 TID2013 [22]，上训练的 ResNet-50 [5] 的 UMAP [3] 可视化结果，展示了合成数据集中的一致聚类趋势。

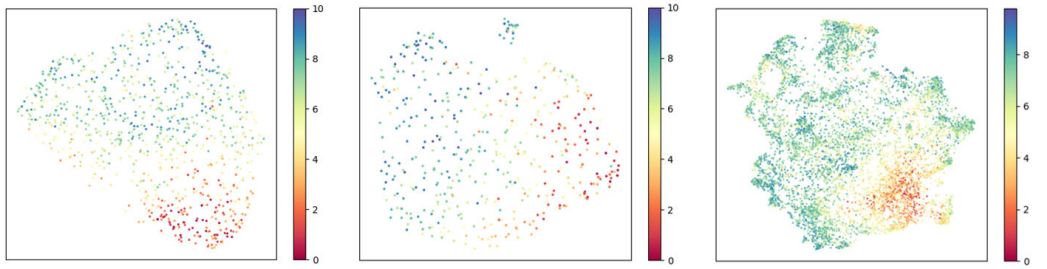


图6：在真实数据集（LIVEC [23]、BID [25]，和 KonIQ-10k [24]）上训练的 ResNet-50 [5] 的 UMAP [3] 可视化结果。特征更加连续且分布均匀，反映出更加均衡、多样的数据结构。

图8展示了 KADID-10k 中的部分高失真图像。这些图像主要呈现强烈且重复的失真特定模式。可以观察到，具有相同失真类型的样本呈现几乎相同的伪影结构；在某些情况下，即使失真类型不同，也可能产生相似的整体视觉模式。这些模式揭示了合成失真过程造成的强烈聚类效应。

这些示例直观地印证了正文中的观察结果：低失真样本在视觉上存在冗余，而高失真样本往往具有高度相似且依赖失真的伪影。这进一步表明，合成失真数据集中的数据分布具有离散和聚类的特征。

C 定理3.1的证明

证明。分析聚类数据的关键挑战在于，样本不满足标准泛化理论所依赖的独立同分布（*i.i.d.*）假设。为应对这一挑战，我们首先应用三角不等式，将泛化误差分解为两个部分：

$$|R(f) - R_{\text{emp}}(f)| \leq |R(f) - R_{\text{m-iid}}(f)| + |R_{\text{m-iid}}(f) - R_{\text{emp}}(f)|, \tag{2}$$

其中，$R_{\text{m-iid}}(f)$ 表示基于从真实分布 $\mathcal{D}$ 中抽取的 m 独立同分布 样本计算得到的经验风险。现在我们对各项分别进行界定。该分解将误差分为：1）标准独立同分布泛化误差 $|R(f) - R_{\text{m-iid}}(f)|$；2）聚类偏差 $|R_{\text{m-iid}}(f) - R_{\text{emp}}(f)|$。

我们分别分析各个组成部分，然后合并结果。

步骤 1：界定标准独立同分布泛化误差。

我们将真实风险 $R(f)$ 与基于 m 独立同分布 样本 $R_{\text{m-iid}}(f)$ 计算得到的经验风险之间的绝对差的上确界定义为：

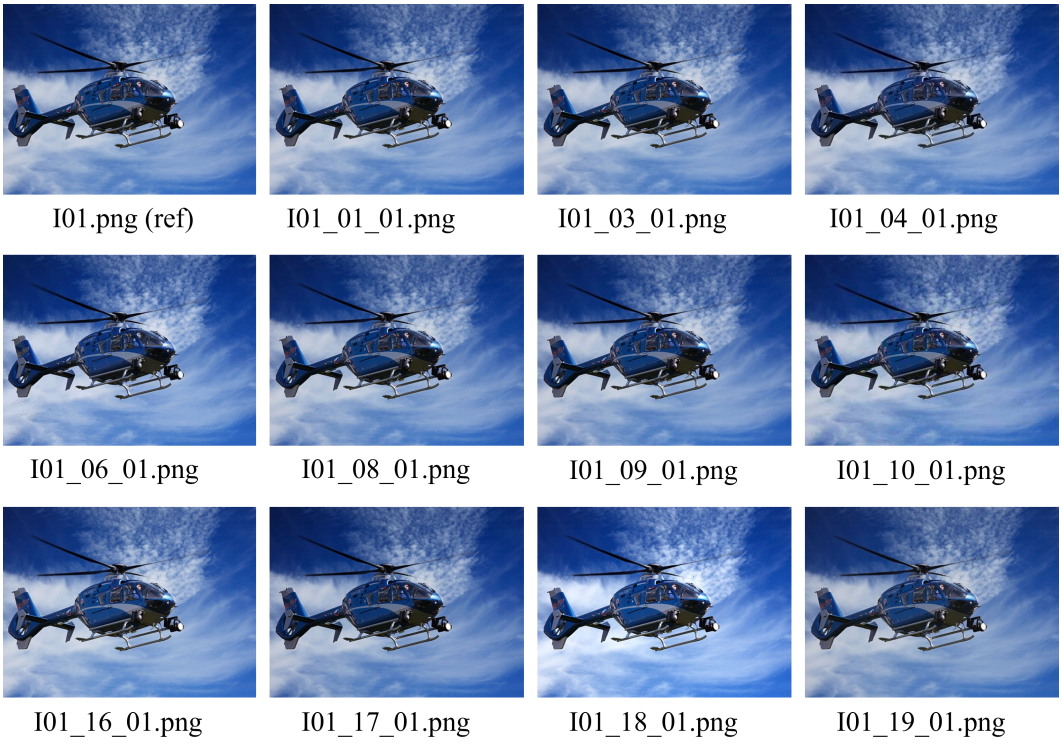


图7：KADID-10k 数据集中部分低失真图像 [38]，以 I01.png 为参考。这些图像具有极高的视觉相似性。

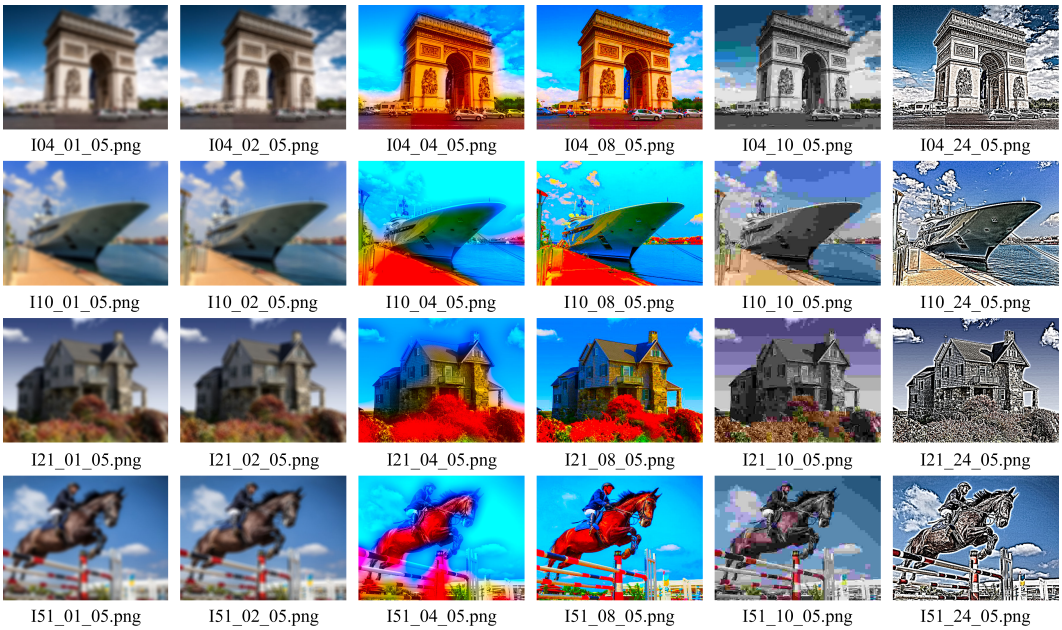


图8：KADID-10k 数据集中部分高失真图像 [38]。相同的失真呈现出非常一致的模式，甚至不同的失真也可能呈现出非常相似的模式。

$$\Phi(X) = \sup_{f \in \mathcal{F}} |R(f) - R_{\text{m-iid}}(f)| .$$

通过应用 McDiarmid不等式，以至少 $1 - \delta/2$ 的概率：

$$\Phi(X) \leq \mathbb{E}[\Phi(X)] + \sqrt{\frac{2\log(2/\delta)}{m}},$$

然后，我们使用 Rademacher复杂度对期望进行上界估计。因此，以至少 $1 - \delta/2$ 的概率：

$$\mathbb{E}[\Phi(X)] \leq 2 \ \text{Rad}_m(\mathcal{F}),$$

其中，$\text{Rad}_m(\mathcal{F})$ 是基于 m *i.i.d* 样本的经验Rademacher复杂度。

步骤2：界定聚类偏差 我们将数据集划分为 m 个聚类中心$\{x_1, x_2, \ldots, x_m\}$，每个聚类中心均从局部分布 $\mathcal{D}_i \subset \mathcal{D}$ 中独立采样。为简单起见，我们假设这 m *i.i.d.* 个聚类中心样本无噪声，即对于每个 x_i，标签由 $\mathbb{E}[Y_i]$给出。

对于每个中心 x_i，我们从其局部分布 $\mathcal{D}_i$ 中生成 k_i 个样本 $\{(x_{i1}, y_{i1}), (x_{i2}, y_{i2}) \ldots, (x_{ik_i}, y_{ik_i})\}$，并将随机变量 Y_i 定义为该聚类上的平均损失：

$$Y_i = \frac{1}{k_i} \sum_{j=1}^{k_i} \ell(f(x_i j), y_{ij}).$$

由于 $\mathcal{Y} = [0, 1]$ 且损失函数 $\ell(\cdot)$ 是 L1，我们有 $l(f(x_i j)\ y_{ij}) \in [0, 1]$，并且可以对 Y_i 的方差给出上界：

$$\text{Var}(Y_i) \leq \frac{1}{4k_i}.$$

由于簇内样本之间具有较高的相关性，簇间误差近似为$R_{\text{emp}}(f)$。此时，应用上述方差界的 Bernstein不等式，以至少 $1 - \delta/2$ 的概率，

$$|R_{m\text{-}iid}(f) - R_{\text{emp}}(f)| \approx \frac{1}{m} \sum_{i=1}^{m} (Y_i - \mathbb{E}[Y_i]) \leq \sqrt{\frac{\eta \log(2/\delta)}{8m}} + \frac{2\log(2/\delta)}{3m}, \tag{3}$$

其中 $\eta = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{k_i}$。

步骤3：合并各项。

对这两个分量应用并集界，以至少 $1 - \delta$ 的概率，

$$|R(f) - R_{\text{emp}}(f)| \leq 2\,\text{Rad}_m(\mathcal{F}) + \sqrt{\frac{2\log(2/\delta)}{m}} + \sqrt{\frac{\eta \log(2/\delta)}{8m}} + \frac{2\log(2/\delta)}{3m}. \tag{4}$$

□

D 算法 1 的定性分析

为了直观地展示算法 1 的效果，我们使用在 ImageNet 上预训练的网络（即算法 1 中使用的特征提取器 $f(\cdot)$）从上采样参考图像集中提取特征，并进行了 UMAP 可视化，如图9所示。

从整体分布来看，可以观察到，新添加的参考样本主要分布在原始训练参考样本之间的边界区域或间隙区域。这

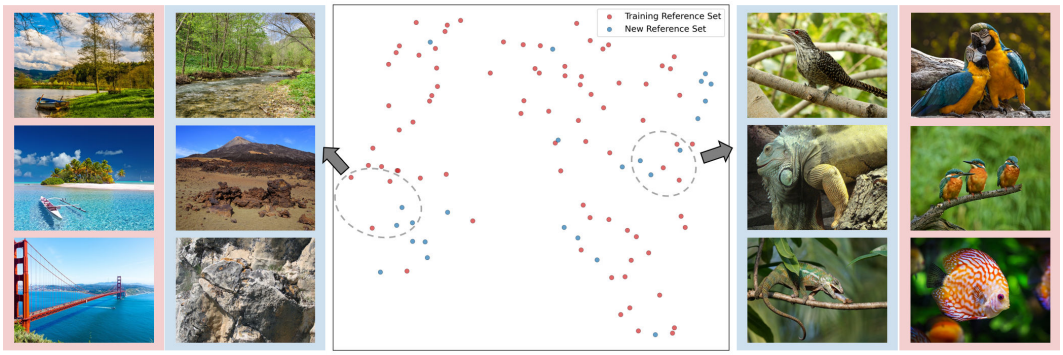


图9：通过算法 1进行内容上采样后参考图像集的 UMAP 可视化，特征由 ImageNet 预训练网络 $f(\cdot)$提取。红点表示原始训练参考图像样本，蓝点表示由算法 1生成的新选样本。左右面板展示了对应于两个局部区域的图像示例。

表明算法 1 能够有效扩大样本空间的覆盖范围，同时不破坏原始数据分布，从而生成更多样且更加均衡的参考集。

图9左侧和右侧的两个局部区域进一步说明，新添加的样本在内容和视觉特征方面与原始样本存在合理差异。这表明，算法 1在保持分布一致性的同时成功提升了内容多样性，并避免引入冗余样本。

E 超参数消融分析

为了全面评估所提出框架中关键超参数的影响，我们在从合成到真实场景下（KADID-10k →LIVEC、KonIQ-10k 和 BID）进行了一系列消融实验。

DDCUp 的超参数分析。 如算法 2所定义，DDCUp 涉及两个关键参数：最近邻数量 k，其决定伪标签参考的多样性；以及 T_{rf}，其控制用于选择可靠邻居的特征相似度阈值。如表 5所示，当 k 过小时，由于邻居信息不足，伪标签会变得不稳定；而当 k 过大时，不可靠的邻居会引入噪声，导致性能略有下降。同样地，将 T_{rf} 从 0.05 放宽至 0.1 会造成性能小幅下降，因为此时会纳入可靠性较低的邻居。当 $k = 5$ 和 $T_{rf} = 0.05$时，可获得最佳结果，在多样性和可靠性之间实现了良好平衡。

表 5：DDCUp 中 k 和 T_{rf} 的消融分析。

SRCC	k	T_{rf}	LIVEC	KonIQ-10k	BID
Ours	5	0.05	**0.713**	**0.727**	0.788
1	1	–	0.703	0.710	0.782
2	3	0.05	0.704	0.710	0.783
3	5	0.1	0.704	0.713	0.779
4	10	0.1	0.712	0.723	0.776
5	ALL	–	0.708	0.723	**0.789**

DRCDown 的超参数分析。 如算法 3 所述，DRCDown 包含三个关键参数：质量分数阈值 T_g、特征差异阈值 T_{df} 和最小保留样本数 T_u。如表 6 所示， T_g 和 T_{df} 共同控制冗余检测过程。如果 T_g 或 T_{df} 过大，模型会过度剪除相似样本，并可能误删多样化样本。相反，如果 T_g 或 T_{df} 过小，则无法有效检测冗余样本。在 $T_g = 1$

和 $T_{df} = 0.1$ 的设置下取得了最佳结果，在有效减少冗余与保留样本多样性之间实现了良好平衡。

表 6：DRCDown 中 T_g 和 T_{df} 的消融分析。

SRCC	T_g	T_{df}	LIVEC	KonIQ-10k	BID
Ours	1	0.1	**0.713**	**0.727**	0.788
1	0.5	0.1	0.700	0.703	0.792
2	1.5	0.1	0.702	0.698	**0.796**
3	1	0.05	0.709	0.716	0.792
4	1	0.15	0.709	0.715	0.790
5	0.5	0.05	0.701	0.695	0.789
6	1.5	0.15	0.710	0.704	0.792

表7报告了不同 T_u 取值下的结果。较小的 T_u 会导致过度下采样，丢失关键信息；而较大的 T_u 则会保留大量冗余，削弱 DRCDown 的效果。当 $T_u = 20$时，模型表现最佳，此时冗余簇足够紧凑，同时仍能保持充分的表征多样性。

表7：DRCDown 中 T_u 的消融分析。

SRCC	T_u	LIVEC	KonIQ-10k	BID
Ours	20	**0.713**	**0.727**	0.788
1	10	0.703	0.709	0.777
2	30	0.701	0.694	**0.791**

从这些结果中可以看出，DRCDown 模块在相当大的参数值范围内也表现稳定。我们的最终配置（$T_g = 1$、 $T_{df} = 0.1$、 $T_u = 20$）始终在冗余抑制与数据集代表性之间实现最佳权衡。

F 模型架构与预训练消融

为全面验证 SynDR-IQA 的普适性和鲁棒性，我们从两个互补的角度开展消融研究：（1）不同骨干架构的影响；（2）不同预训练策略（即 ImageNet 和 CLIP）的效果。这些实验旨在共同评估 SynDR-IQA 的有效性是否依赖于特定的模型架构或特征初始化方案。

F.1 骨干架构的影响

为考察架构的普适性，我们使用两种具有代表性的基于Transformer的骨干网络评估 SynDR-IQA：视觉Transformer（ViT-B/16） [39] 和 Swin Transformer Tiny（ Swin-T） [38]。两种骨干网络均遵循与主要实验相同的训练方案。基线对应 DQGA。

如表 8 所示，SynDR-IQA 在所有真实数据集上均持续提升基线性能，使用 Swin-T 时 SRCC 平均提升 7.2%，使用 ViT-B/16 时提升 2.0%。这些结果表明，SynDR-IQA 能够有效泛化至不同的网络架构，证实了我们框架的稳健性和多功能性。

表 8：从合成到真实场景下，使用不同 ImageNet预训练骨干网络时基线与 SynDR-IQA 的性能比较（仅 SRCC）。

骨干网络	方法	LIVEC	KonIQ-10k	BID	平均值
Swin-T	基线	0.620	0.600	0.729	0.649
	SynDR-IQA	**0.670**	**0.719**	**0.777**	**0.721**
ViT-B/16	基线	0.714	0.694	0.740	0.716
	SynDR-IQA	**0.729**	**0.717**	**0.761**	**0.736**

F.2 预训练策略的影响

除了网络架构差异之外，我们还进一步研究了不同预训练策略对模型性能的影响。具体而言，我们比较了 ImageNet预训练与 CLIP 预训练[40], ，后者通过大规模视觉-语言预训练展现出强大的语义层面特征学习能力。

我们首先探究，完全用 CLIP 预训练模型替换骨干网络是否能够提升性能。表9在相同的架构和训练配置下，对比了以 CLIP 预训练的 ViT-B/16 和以 ImageNet 预训练的 ViT-B/16 作为骨干网络的效果。

Tab表9：将 ImageNet 和 CLIP 预训练模型用作骨干网络时的比较（仅 SRCC） .

骨干网络	方法	LIVEC	KonIQ-10k	BID	平均值
ViT-B/16（ImageNet）	基线	0.714	0.694	0.740	0.716
	SynDR-IQA	**0.729**	0.717	0.761	**0.736**
ViT-B/16（CLIP）	基线	0.653	0.683	0.706	0.681
	SynDR-IQA	0.692	**0.744**	**0.766**	0.734

如表9所示，将骨干网络替换为CLIP，并未带来相较于ImageNet预训练的明显性能优势。我们认为这是因为CLIP的多模态预训练目标侧重于视觉和语言模态之间的高层次语义对齐，却牺牲了对感知质量评估任务至关重要的低层次视觉细节。

然而，受CLIP卓越语义理解能力的启发，我们进一步探索了一种更精细的策略：仅在DDCUp模块中使用CLIP（如 $f(\cdot)$所示），同时保留经ImageNet预训练的ViT-B/16作为主要的IQA骨干网络。这种设计既利用了CLIP处理多样化内容时卓越的语义理解能力，又保留了主网络中进行准确质量预测所必需的低层次视觉表征。

表10：在DDCUp模块中使用不同编码器时的性能比较。主IQA骨干网络仍为ViT-B/16（经ImageNet预训练）（仅报告SRCC）。

DDCUp 编码器	LIVEC	KonIQ-10k	BID	平均值
ViT-B/16（ImageNet）	0.729	0.717	0.761	0.736
ViT-B/16（CLIP）	**0.739**	**0.751**	**0.778**	**0.756**

如表10所示，在 DDCUp模块中选择性地集成 CLIP 可带来显著的性能提升，并在所有测试数据集上取得新的最先进结果。我们认为，这些提升源于 CLIP 增强的语义理解能力，使模型能够在上采样过程期间更准确地进行内容感知参考选择，并进一步提升模型整体的训练效果。

G 定性结果

为了定性展示我们方法的有效性，我们在图10中展示了来自 LIVEC 的若干代表性示例。这些示例涵盖了多种场景，具有不同的质量分数、失真、场景和内容。值得注意的是，我们的模型仅使用合成失真数据集 KADID-10k 进行训练，却能够在这些真实世界图像上生成与人工标注的真实分数高度一致的预测结果，从而验证了我们方法有效的合成到真实泛化能力。此外，与最先进的方法 CLIP-IQA [33], 相比，我们的方法与人类感知的对齐程度显著更高。

H 局限性

尽管 SynDR-IQA 在合成到算法的泛化场景中取得了显著提升，但在实际可用性方面仍存在明显的性能差距。我们认为，主要挑战源于现有合成失真模式与算法失真特征之间的根本差异。当前的合成失真数据集主要关注

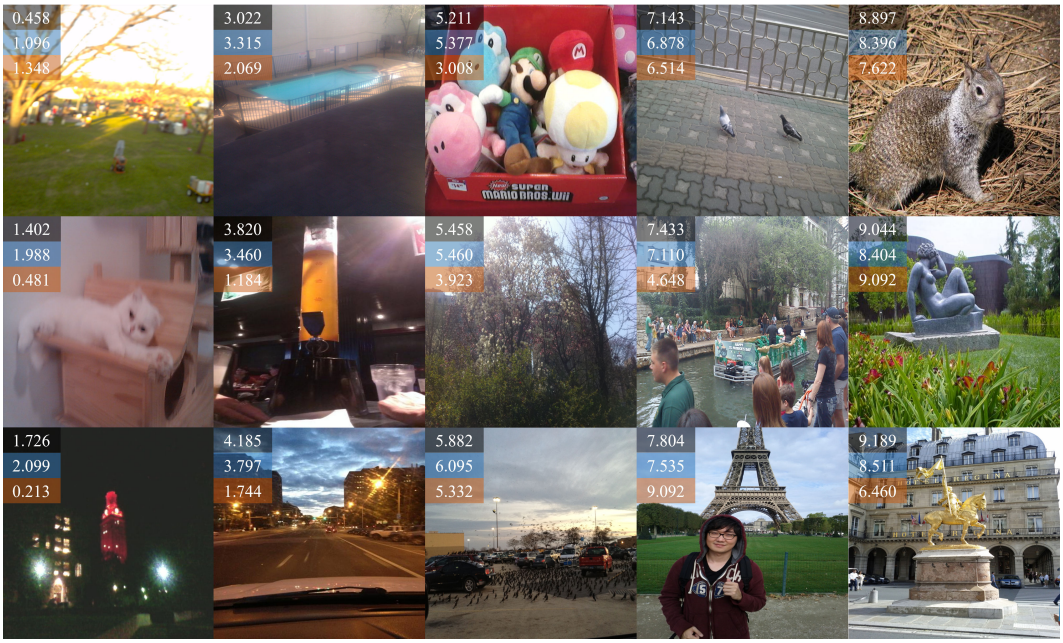


图 10：SynDR-IQA 在 LIVEC 上的定性结果。对于每张图像，顶部数字表示人工标注的真实分数，并归一化到范围 [0, 10]；中间数字表示我们模型的预测质量分数；底部数字表示 CLIP-IQA 预测的质量分数 [33]。这些图像的真实分数从左到右、从上到下逐渐增加。

传统退化类型（例如模糊、噪声、压缩），而这些类型与现代图像处理算法引入的复杂模式存在显著差异，尤其是涉及基于深度学习的方法时。

这一局限性凸显了对合成失真生成方法的需求：这类方法能够生成具有算法失真及其他复杂特征的合成失真，同时保持可控的图像质量退化。我们认为，通过未来的研究解决这一局限性，对于进一步提升盲图像质量评估模型在不同失真场景下的泛化能力至关重要。