Blind Image Quality Assessment via Vision-Language Correspondence: A Multitask Learning Perspective

Weixia Zhang[1], Guangtao Zhai[1], Ying Wei[2], Xiaokang Yang[1], Kede Ma[2,3*]
[1] MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[2] Department of Computer Science, City University of Hong Kong
[3] Shenzhen Research Institute, City University of Hong Kong
{zwx8981, zhaiguangtao, xkyang}@sjtu.edu.cn
{yingwei, kede.ma}@cityu.edu.hk

Abstract

We aim at advancing blind image quality assessment (BIQA), which predicts the human perception of image quality without any reference information. We develop a general and automated multitask learning scheme for BIQA to exploit auxiliary knowledge from other tasks, in a way that the model parameter sharing and the loss weighting are determined automatically. Specifically, we first describe all candidate label combinations (from multiple tasks) using a textual template, and compute the joint probability from the cosine similarities of the visual-textual embeddings. Predictions of each task can be inferred from the joint distribution, and optimized by carefully designed loss functions. Through comprehensive experiments on learning three tasks - BIQA, scene classification, and distortion type identification, we verify that the proposed BIQA method 1) benefits from the scene classification and distortion type identification tasks and outperforms the state-of-the-art on multiple IQA datasets, 2) is more robust in the group maximum differentiation competition, and 3) realigns the quality annotations from different IQA datasets more effectively. The source code is available at https://github.com/zwx8981/LIQE.

1. Introduction

As a fundamental computational vision task, blind image quality assessment (BIQA) [63] aims to predict the visual quality of a digital image with no access to the underlying pristine-quality counterpart (if any). In the age of deep learning, the development of BIQA can be mainly characterized by strategies to mitigate the conflict between the large number of trainable parameters and the small number of human quality annotations in the form of mean opinion scores (MOSs). When synthetic distortions (*e.g*., Gaussian noise and JPEG compression) are of primary concern, patchwise training [4], quality-aware pre-training [32, 37, 76], and learning from noisy pseudo-labels [2, 38, 67] are practical training tricks with less (or no) reliance on MOSs. Here the underlying assumptions are that 1) the pristine-quality images exist and are accessible, 2) the visual distortions can be simulated efficiently and automatically, and 3) full-reference IQA models [64] are applicable and provide adequate quality approximations. However, all these assumptions do not hold when it comes to realistic camera distortion (*e.g*., sensor noise, motion blurring or a combination of both). A different set of training tricks have been explored, including transfer learning [17, 76], meta learning [80], and contrastive learning [40]. Emerging techniques that combine multiple datasets for joint training [78] and that identify informative samples for active fine-tuning [66] can also be seen as ways to confront the data challenge in BIQA.

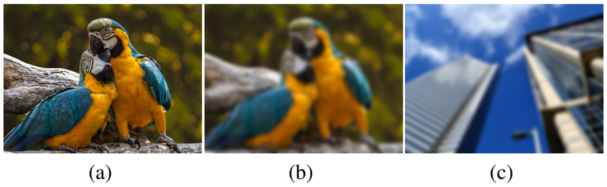

Figure 1. **(a)** A “parrots” image of pristine quality. **(b)** A distorted version of (a) by global Gaussian blurring. **(c)** A distorted “cityscape” image by the same level of Gaussian blurring. Humans are able to “see through” the Gaussian blur, and recognize the two parrots in (b) with no effort, suggesting the internal representations for the task of visual recognition should be *distortion-insensitive*. This makes it conceptually conflicting to BIQA, which relies on *distortion-sensitive* representations for quality prediction.

In this paper, we aim to accomplish something in the

[*] Corresponding author.

same spirit, but from a different multitask learning perspective. We ask the key question:

> *Can BIQA benefit from auxiliary knowledge provided by other tasks in a multitask learning setting?*

This question is of particular interest because many high-level computer vision tasks (*e.g*., object recognition [8] and scene classification [5]), with easier-to-obtain ground-truth labels, seem to be conceptually conflicting to BIQA. This is clearly illustrated in Fig. 1. Humans are able to "see through" the Gaussian blur, and recognize effortlessly the two parrots in (b). That is, if we would like to develop computational methods for the same purpose, they should rely on *distortion-insensitive* features, and thus be robust to such corruptions. This is also manifested by the common practice in visual recognition that treats synthetic distortions as forms of data augmentation [16]. In stark contrast, BIQA relies preferentially on *distortion-sensitive* features to quantify the perceptual quality of images of various semantic content. Ma *et al*. [37] proposed a cascaded multitask learning scheme for BIQA, but did not investigate the relationships between BIQA and high-level vision tasks. Fang *et al*. [9] included scene classification as one task, but required manually specifying the parameters (*i.e*., computations) to share across tasks, which is difficult and bound to be suboptimal.

Taking inspiration from recent work on vision-language pre-training [49], we propose a general and automated multitask learning scheme for BIQA, with an attempt to answer the above-highlighted question. Here, "automated" means that the model parameter sharing for all tasks and the loss weighting assigned to each task are determined automatically. We consider two additional tasks, scene classification and distortion type identification, the former of which is conceptually conflicting to BIQA, while the latter is closely related. We first summarize the scene category, distortion type, and quality level of an input image using a textual template. For example, Fig. 1 (c) may be described as "a photo of a *cityscape* with *Gaussian blur* artifacts, which is of *bad* quality." We then employ the contrastive language-image pre-training (CLIP) [49], a joint vision and language model trained with massive image-text pairs, to obtain the visual and textual embeddings. The joint probability over the three tasks can be computed from the cosine similarities between the image embedding and all candidate textual embeddings[1]. We marginalize the joint distribution to obtain the marginal probability for each task, and further convert the discretized quality levels to a continuous quality score using the marginal distribution as the weighting.

We supplement existing IQA datasets [7, 11, 17, 22, 27, 54] with scene category and distortion type labels, and jointly optimize the entire method on a combination of them by minimizing a weighted sum of three fidelity losses [58], where the loss weightings are adjusted automatically based on the training dynamics [31]. From extensive experimental results, we arrive at a positive answer to the highlighted question: BIQA can indeed benefit from both scene classification and distortion type identification. The resulting model, which we name **L**anguage-**I**mage **Q**uality **E**valuator (LIQE), not only outperforms state-of-the-art BIQA methods [13,55,76,78] in terms of prediction accuracy on multiple IQA datasets, but also exhibits improved generalizability in the group maximum differentiation (gMAD) competition [35]. In addition, we provide quantitative evidence that LIQE better realigns MOSs from different IQA datasets in a common perceptual scale [48].

2. Related Work

In this section, we give an overview of recent progress in BIQA, with an emphasis on new paradigms. We then review CLIP [49] and its applications, and multitask learning in machine learning.

2.1. BIQA

Conventional BIQA either relied on hand-engineered features in the form of natural scene statistics (NSS) [12, 44–46] or shallow feature learning [18, 69, 71] in the form of codebooks. Deep learning takes advantage of the end-to-end optimization of feature extraction and quality regression, and has significantly advanced the field of BIQA. Over the years, BIQA methods have explored different *computational structures*: generalized divisive normalization [37], multi-level feature aggregation [25], adaptive convolution [55], and self-attention [13, 19, 70]; and *objective functions*: ℓ_p-norm induced metric, fidelity-based ranking losses [58], Pearson linear correlation coefficient (PLCC), and differentiable approximations to Spearman rank correlation coefficient (SRCC) [3], and generalized correlation loss for accelerated convergence [24].

New paradigms for BIQA flourish in recent years, aiming at exploring promising directions of next-generation BIQA. Representative work includes: patch-to-picture learning for local quality prediction [72], active learning for worthy sample identification [65, 66], unified optimization for cross-distortion scenarios [78], meta-learning for fast adaptation [80], continual learning for streaming distortions [29, 39, 74, 77], and perceptual attacks for robustness evaluation [75]. In this paper, we leverage multitask learning to facilitate auxiliary knowledge transfer.

2.2. CLIP Applications

CLIP has shown great promise to assist a broad scope of vision tasks. Originally, Radford *et al*. [49] leveraged 400

[1] We specify nine scene categories, eleven distortion types, and five quality levels, giving rise to 495 textual descriptions/embeddings in total.

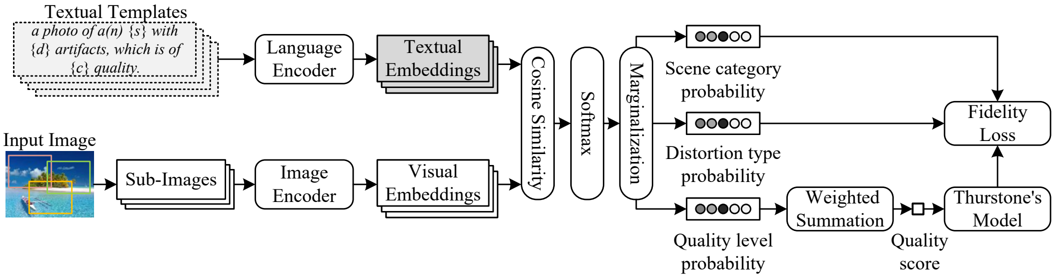

Figure 2. System diagram of the proposed LIQE.

million image-text pairs to pre-train a family of CLIP models, which present remarkable zero-shot transfer ability to a wide range of downstream vision tasks. Zhou *et al.* [79] suggested prompt tuning to improve transfer effectiveness, at the cost of language interpretability. Shortly after its inception, CLIP has found its way to (open-vocabulary) semantic segmentation [23, 68] and object detection [15, 26]. Vinker *et al.* [60] found a novel use of CLIP for object sketching with excellent understanding of object semantics. Closest to ours, Wang *et al.* [62] assumed CLIP models to be inherently quality-aware, and adopted them to assess image quality and aesthetics via prompt engineering. Our method differs significantly from theirs [62] both conceptually and computationally. We exploit CLIP in the multitask learning setting to aid BIQA by auxiliary knowledge transfer. Moreover, we fine-tune the CLIP model instead of prompt tuning for much better quality prediction performance without sacrificing language interpretability.

2.3. Multitask Learning

Multitask learning as a specific case of multi-objective optimization [42] typically improves task accuracy, memory cost, and inference time through shared computation and information across tasks. Existing methods differ mainly in two design choices: model parameter sharing and loss weighting. Instead of manually specifying which parameters to share [1, 21, 43] or learning to determine task specific parameters with a combinatorial complexity [41, 51, 56, 59, 61], we assume all parameters in the image encoder of LIQE are shared, whose capacity is dynamically allocated to each task during end-to-end optimization. For loss weighting, Sener *et al.* [52] and Lin *et al.* [28] cast multitask learning as multi-objective optimization, and solved for the optimal loss weighting according to the Karush-Kuhn-Tucher conditions. Liu *et al.* [30] implemented an automated weighting scheme based on model agnostic meta learning [10], which allows specification of primary (*i.e.*, the most important) tasks. Other effective heuristics for loss weighting include learning task uncertainty [20], gradient normalization [6], and loss descending rate [31]. In this paper, we adopt the method in [31] due to its conceptual simplicity, computational convenience, and high efficiency.

3. LIQE from Multitask Learning

In this section, we first present some preliminaries of the problem formulation. We then describe a general and automated multitask learning scheme for BIQA based on vision-language correspondence, followed by specifications of loss functions to drive the end-to-end optimization. The system diagram of LIQE is shown in Fig. 2.

3.1. Preliminaries

Given an image $\boldsymbol{x} \in \mathbb{R}^N$ that may undergo several stages of degradations, the goal of a BIQA model $\hat{q} : \mathbb{R}^N \mapsto \mathbb{R}$ is to predict the perceptual quality of $\boldsymbol{x}$, close to its MOS $q(\boldsymbol{x}) \in \mathbb{R}$. We also work with a Likert-scale of five quality levels [11, 54]: $c \in \mathcal{C} = \{1, 2, 3, 4, 5\} = \{$“bad”, “poor”, “fair”, “good”, “perfect”$\}$, and relate c to $\hat{q}$ by

$$\hat{q}(\boldsymbol{x}) = \sum_{c=1}^{C} \hat{p}(c|\boldsymbol{x}) \times c, \tag{1}$$

where $C = 5$ is the number of quality levels and $\hat{p}(c|\cdot)$ is the marginal probability of c to be estimated.

Apart from BIQA, we also include a conceptually conflicting task - scene classification and a closely related task - distortion type identification. We consider nine scene categories: $s \in \mathcal{S} = \{$“animal”, “cityscape”, “human”, “indoor scene”, “landscape”, “night scene”, “plant”, “still-life”, and “others”$\}$. An image may contain multiple scene labels. We do not discriminate between synthetic and realistic distortions, and instead identify the dominant distortion in the image: $d \in \mathcal{D} = \{$“blur”, “color-related”, “contrast”, “JPEG compression”, “JPEG2000 compression”, “noise”, “over-exposure”, “quantization”, “under-exposure”, “spatially-localized”, and “others”$\}$ with eleven in total. According to our characterization, the “others” category includes images with no distortions (*i.e.*, of pristine quality). It is then natural to create a textual template to put together labels from the three tasks: “*a photo of a(n)* $\{s\}$ *with* $\{d\}$ *artifacts,* *which is of* $\{c\}$ *quality*", and we have $5 \times 9 \times 11 = 495$ candidate textual descriptions.

3.2. Vision-Language Correspondence

Joint Probability over Multiple Tasks. The proposed LIQE relies on a pre-trained CLIP model [49] for feature embeddings. CLIP consists of an image encoder $\boldsymbol{f}_{\boldsymbol{\phi}} : \mathbb{R}^N \mapsto \mathbb{R}^K$ and a language encoder $\boldsymbol{g}_{\boldsymbol{\varphi}} : \mathcal{T} \mapsto \mathbb{R}^K$, parameterized by $\boldsymbol{\phi}$ and $\boldsymbol{\varphi}$, respectively, where $\mathcal{T}$ denotes the text corpus. Note that $\boldsymbol{f}_{\boldsymbol{\phi}}$ and $\boldsymbol{g}_{\boldsymbol{\varphi}}$ share the same feature dimension by design. As also part of the model design, $\boldsymbol{f}_{\boldsymbol{\phi}}$ accepts input images with a fixed spatial size (*i.e.*, $224 \times 224 \times 3$). A naïve resizing (*e.g.*, by bilinear interpolation) may change the perceptual quality, which corresponds to adjusting the effective viewing distance. As a result, we prefer cropping U sub-images from $\boldsymbol{x}$, and obtain the visual embedding matrix $\boldsymbol{F}(\boldsymbol{x}) \in \mathbb{R}^{U \times K}$. Similarly, given a total of V candidate textual descriptions by the lower-cased byte pair encoding (BPE) [53], we use the language encoder to obtain the textual embedding matrix $\boldsymbol{G}(\boldsymbol{x}) \in \mathbb{R}^{V \times K}$, where $V = 495$.

We then compute the cosine similarity between the visual embedding of the u-th sub-image $\boldsymbol{F}_{u\bullet}$ (as a row vector and for $1 \leq u \leq U$) and the v-th candidate textual embedding $\boldsymbol{G}_{v\bullet}$ (corresponding to a particular set of $\{c, s, d\}$), averaging across U sub-images to obtain the image-level correspondence score:

$$\text{logit}(c, s, d|\boldsymbol{x}) = \frac{1}{U} \sum_{u=1}^{U} \frac{\boldsymbol{F}_{u\bullet}(\boldsymbol{x}) \boldsymbol{G}_{v\bullet}^{\intercal}(\boldsymbol{x})}{\|\boldsymbol{F}_{u\bullet}(\boldsymbol{x})\|_2 \|\boldsymbol{G}_{v\bullet}(\boldsymbol{x})\|_2}. \quad (2)$$

After corresponding the image to all candidate textual descriptions, we apply a softmax function to compute a joint probability with a learnable temperature parameter τ_1:

$$\hat{p}(c, s, d|\boldsymbol{x}) = \frac{\exp\left(\text{logit}(c, s, d|\boldsymbol{x})/\tau_1\right)}{\sum_{c,s,d} \exp\left(\text{logit}(c, s, d|\boldsymbol{x})/\tau_1\right)}. \quad (3)$$

Ideally, we shall minimize the statistical distance between the predicted joint probability and the ground-truth joint probability $p(c, s, d|\boldsymbol{x})$ for optimizing the parameter vector $\boldsymbol{\theta} = \{\boldsymbol{\phi}, \boldsymbol{\varphi}, \tau\}$. However, $p(c, s, d|\boldsymbol{x})$ may not be inferred accurately due to the fact that existing IQA datasets only provide continuous quality scores instead of discrete quality levels. Moreover, different IQA datasets have different perceptual scales due to differences in subjective testing [78], which further complicates quality score-to-level conversion.

Loss for BIQA. Given $\hat{p}(c, s, d|\boldsymbol{x})$, we marginalize it to obtain $\hat{p}(c|\boldsymbol{x})$, and compute the quality estimate $\hat{q}(\boldsymbol{x}) \in \mathbb{R}$ by Eq. (1). We consider the pairwise learning-to-rank model estimation for BIQA. Specifically, for an image pair $(\boldsymbol{x}, \boldsymbol{y})$ from the same IQA dataset, we compute a binary label according to their ground-truth MOSs:

$$p(\boldsymbol{x}, \boldsymbol{y}) = \begin{cases} 1 & \text{if } q(\boldsymbol{x}) \geq q(\boldsymbol{y}) \\ 0 & \text{otherwise} \end{cases}. \quad (4)$$

Under the Thurstone's model [57], we estimate the probability of $\boldsymbol{x}$ perceived better than $\boldsymbol{y}$ as

$$\hat{p}(\boldsymbol{x}, \boldsymbol{y}) = \Phi\left(\frac{\hat{q}(\boldsymbol{x}) - \hat{q}(\boldsymbol{y})}{\sqrt{2}}\right), \quad (5)$$

where $\Phi(\cdot)$ is the standard Normal cumulative distribution function, and the variance is fixed to one. We adopt the fidelity loss [58] as the statistical distance measure:

$$\begin{aligned} \ell_q(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta}) = 1 &- \sqrt{p(\boldsymbol{x}, \boldsymbol{y}) \hat{p}(\boldsymbol{x}, \boldsymbol{y})} \\ &- \sqrt{(1 - p(\boldsymbol{x}, \boldsymbol{y}))(1 - \hat{p}(\boldsymbol{x}, \boldsymbol{y}))}. \end{aligned} \quad (6)$$

Loss for Scene Classification. In our setting, an image can be assigned to one or more scene categories, leading to a multi-label classification problem. Thus, we compute an average of S binary fidelity losses:

$$\begin{aligned} \ell_s(\boldsymbol{x}; \boldsymbol{\theta}) = &\frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \Big(1 - \sqrt{p(s|\boldsymbol{x}) \hat{p}(s|\boldsymbol{x})} \\ &- \sqrt{(1 - p(s|\boldsymbol{x}))(1 - \hat{p}(s|\boldsymbol{x}))}\Big), \end{aligned} \quad (7)$$

where $p(s|\boldsymbol{x}) = 1$ if $\boldsymbol{x}$ is in the s category and zero otherwise. $\sum_s p(s|\boldsymbol{x}) = S$, where $1 \leq S \leq |\mathcal{S}|$ is the number of target categories to which $\boldsymbol{x}$ belongs. $\hat{p}(s|\boldsymbol{x})$ is the marginal probability computed from Eq. (3). We also try a softmax loss formulation:

$$\ell_s(\boldsymbol{x}; \boldsymbol{\theta}) = \left(1 - \sum_{s \in \mathcal{S}} \sqrt{p(s|\boldsymbol{x}) \hat{p}(s|\boldsymbol{x})}\right), \quad (8)$$

where equal probabilities are assigned to the target scene categories with $\sum_s p(s|\boldsymbol{x}) = 1$. Similar results are obtained, which we attribute to the relatively simple setting of our scene classification problem with only nine categories.

Loss for Distortion Type Identification. Because we only consider the dominant distortion type in $\boldsymbol{x}$, distortion type identification can be formulated as a standard multi-class classification problem. We use the multi-class fidelity loss:

$$\ell_d(\boldsymbol{x}; \boldsymbol{\theta}) = \left(1 - \sum_{d \in \mathcal{D}} \sqrt{p(d|\boldsymbol{x}) \hat{p}(d|\boldsymbol{x})}\right), \quad (9)$$

where $p(d|\boldsymbol{x}) = 1$ if $\boldsymbol{x}$ contains the d artifacts, and zero otherwise. $\hat{p}(d|\boldsymbol{x})$ is computed similarly by marginalizing the joint probability in Eq. (3).

Final Loss for Multitask Learning. Similar to [78], we choose to jointly train a BIQA model on M ($M \geq 2$) IQA datasets. At the t-th training iteration and from the m-th dataset, we sample a mini-batch $\mathcal{B}_t^{(m)}$, and form all possible pairs of images with ground-truths (by Eq. (4)), which are

Table 1. Median SRCC and PLCC results across ten sessions along with the standard deviation in the bracket. CLIVE stands for the LIVE Challenge Database. The top two results are highlighted in bold.

Dataset	LIVE [54]	CSIQ [22]	KADID-10k [27]	BID [7]	CLIVE [11]	KonIQ-10k [17]
Criterion	SRCC					
NIQE	0.908 (± 0.017)	0.631 (± 0.038)	0.389 (± 0.019)	0.573 (± 0.044)	0.446 (± 0.065)	0.415 (± 0.019)
ILNIQE	0.887 (± 0.032)	0.808 (± 0.039)	0.565 (± 0.013)	0.548 (± 0.044)	0.469 (± 0.063)	0.509 (± 0.021)
Ma19	0.922 (± 0.024)	0.926 (± 0.017)	0.465 (± 0.019)	0.373 (± 0.059)	0.336 (± 0.038)	0.360 (± 0.013)
PaQ2PiQ	0.544 (± 0.033)	0.697 (± 0.040)	0.403 (± 0.021)	0.719 (± 0.043)	0.732 (± 0.036)	0.722 (± 0.012)
KonCept	0.673 (± 0.040)	0.631 (± 0.064)	0.503 (± 0.025)	0.816 (± 0.029)	0.778 (± 0.024)	0.911 (± 0.005)
MUSIQ	0.837 (± 0.011)	0.697 (± 0.040)	0.572 (± 0.027)	0.744 (± 0.038)	0.785 (± 0.029)	**0.915** (± 0.003)
DBCNN	0.963 (± 0.012)	**0.940** (± 0.015)	0.878 (± 0.023)	**0.864** (± 0.016)	0.835 (± 0.022)	0.864 (± 0.007)
HyperIQA	**0.966** (± 0.012)	0.934 (± 0.031)	0.872 (± 0.017)	0.848 (± 0.033)	**0.855** (± 0.021)	0.900 (± 0.006)
TreS	0.965 (± 0.019)	0.902 (± 0.041)	0.881(± 0.019)	0.853 (± 0.031)	0.846 (± 0.020)	0.907 (± 0.007)
UNIQUE	0.961 (± 0.005)	0.902 (± 0.052)	**0.884** (± 0.013)	0.852 (± 0.027)	0.854 (± 0.020)	0.895 (± 0.008)
LIQE	**0.970** (± 0.004)	**0.936** (± 0.025)	**0.930** (± 0.009)	**0.875** (± 0.020)	**0.904** (± 0.014)	**0.919** (± 0.004)
Criterion	PLCC					
NIQE	0.904 (± 0.089)	0.719 (± 0.022)	0.442 (± 0.019)	0.618 (± 0.045)	0.507 (± 0.054)	0.438 (± 0.015)
ILNIQE	0.894 (± 0.025)	0.851 (± 0.029)	0.611 (± 0.027)	0.494 (± 0.046)	0.518 (± 0.051)	0.534 (± 0.020)
Ma19	0.923 (± 0.019)	0.929 (± 0.011)	0.501 (± 0.013)	0.399 (± 0.059)	0.405 (± 0.033)	0.398 (± 0.009)
PaQ2PiQ	0.558 (± 0.023)	0.766 (± 0.028)	0.448 (± 0.012)	0.700 (± 0.032)	0.755 (± 0.022)	0.716 (± 0.009)
KonCept	0.619 (± 0.039)	0.645 (± 0.043)	0.515 (± 0.018)	0.825 (± 0.026)	0.799 (± 0.016)	**0.924** (± 0.003)
MUSIQ	0.818 (± 0.011)	0.766 (± 0.028)	0.584 (± 0.016)	0.774 (± 0.030)	0.828 (± 0.017)	**0.937** (± 0.003)
DBCNN	**0.966** (± 0.010)	**0.954** (± 0.013)	0.878 (± 0.022)	**0.883** (± 0.017)	0.854 (± 0.015)	0.868 (± 0.006)
HyperIQA	**0.968** (± 0.011)	**0.946** (± 0.022)	0.869 (± 0.018)	0.868 (± 0.027)	0.878 (± 0.015)	0.915 (± 0.004)
TreS	0.963 (± 0.016)	0.923 (± 0.031)	0.879 (± 0.019)	0.871 (± 0.028)	0.877 (± 0.016)	**0.924** (± 0.006)
UNIQUE	0.952 (± 0.007)	0.921 (± 0.048)	**0.885** (± 0.011)	0.875 (± 0.019)	**0.884** (± 0.014)	0.900 (± 0.005)
LIQE	0.951 (± 0.006)	0.939 (± 0.024)	**0.931** (± 0.009)	**0.900** (± 0.016)	**0.910** (± 0.013)	0.908 (± 0.002)

collectively denoted by $\mathcal{P}_t^{(m)}$. We combine $\{\mathcal{B}_t^{(m)}\}_{i=1}^M$ and $\{\mathcal{P}_t^{(m)}\}_{i=1}^M$ to form $\mathcal{B}_t$ and $\mathcal{P}_t$, respectively, based on which we define the final loss as a linearly weighted summation of the three individual losses:

$$\ell(\mathcal{B}, t; \boldsymbol{\theta}) = \frac{1}{|\mathcal{P}|} \sum_{(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}} \lambda_q(t) \ell_q(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta}) + \frac{1}{|\mathcal{B}|} \sum_{\boldsymbol{x} \in \mathcal{B}} \big(\lambda_s(t) \ell_s(\boldsymbol{x}) + \lambda_d(t) \ell_d(\boldsymbol{x})\big). \quad (10)$$

The weighting vector $\boldsymbol{\lambda}(t) = [\lambda_q(t), \lambda_s(t), \lambda_d(t)]^\intercal$ at the t-th iteration can be automatically computed according to the relative descending rate [31]:

$$\lambda_j(t) = \frac{\exp\left(w_j(t-1)/\tau_2\right)}{\sum_i \exp\left(w_i(t-1)/\tau_2\right)}, w_j(t-1) = \frac{\ell_j(t-1)}{\ell_j(t-2)}, \quad (11)$$

where $i, j \in \{q, s, d\}$, and τ_2 is another fixed temperature parameter. When τ_2 is sufficiently large, we approach a uniform distribution, where different losses are weighted equally. In our experiments, we follow the suggestions in [31], and compute $w_j(t-1)$ through the moving average loss over the recent iterations in each epoch.

4. Experiments

In this section, we first present the experimental setups, upon which we compare LIQE with several state-of-the-art BIQA methods. We then evaluate LIQE's ability to re-align MOSs from different datasets qualitatively and quantitatively. The rationality of each design choice is verified through a series of ablation studies. Finally, we provide an analysis of task relationships.

4.1. Experimental Setups

We conduct experiments on six IQA datasets, among which LIVE [54], CSIQ [22], and KADID-10k [27] contain synthetic distortions, while LIVE Challenge [11] (denoted as CLIVE in Table 1), BID [7], and KonIQ-10K [17] include realistic distortions. We randomly sample 70% and 10% images from each dataset to construct the training and validation set, respectively, leaving the remaining 20% for testing. Regarding the three datasets with synthetic distortions, we split the train/val/test sets according to the reference images in order to ensure content independence. We repeat this procedure ten times, and report median SRCC and PLCC results as prediction monotonicity and precision measures, respectively.

We adopt ViT-B/32 [49] as the visual encoder and GPT-2 [50] with a base size of 63M parameters as the text encoder. We train the model by minimizing the objective in Eq. (10) using AdamW [34] with a decoupled weight decay regularization of 10^{-3}. The initial learning rate is set to 5×10^{-6}, which is scheduled by a cosine annealing rule [33]. We optimize LIQE for 80 epochs with a mini-batch size of 4 on the LIVE, CSIQ, BID, and LIVE Challenge datasets, and size of 16 on KonIQ-10k and KADID-10k. During training and inference, we randomly crop 3 and 15 sub-images with a spatial size of $224 \times 224 \times 3$ from original images without changing their aspect ratios. All experiments are conducted on a single NVIDIA GeForce RTX 3090 GPU.

4.2. Main Results

We compare the performance of the proposed LIQE against three opinion-unaware BIQA models, including NIQE [45], ILNIQE [73], and Ma19 [38], and seven data-driven DNN-based methods, including PaQ2PiQ [72], KonCept [17], MUSIQ [19], DBCNN [76], HyperIQA [55], TreS [13], and UNIQUE [78]. For competing models, we either directly adopt the publicly available implementations [4, 17, 19, 72], or re-train them on our datasets with the training codes provided by the respective authors [13,55,76,78]. DBCNN, HyperIQA, and TreS are separately trained on each individual dataset, while UNIQUE is jointly trained on six datasets. We summarize the median SRCC and PLCC results across ten sessions in Table 1, from which we draw several insightful observations. First, although opinion-unaware methods aim ambitiously for handling arbitrary distortions, they only perform well on two legacy datasets LIVE [54] and CSIQ [22], which contain the most basic distortions in the field of IQA. Second, despite being trained on a relatively large-scale dataset, PaQ2PiQ [72] only presents reasonable quality prediction accuracy on datasets with realistic distortions [7, 11, 17]. Equipped with more sophisticated backbone networks and trained on KonIQ-10k [17], KonCept [17] and MUSIQ [19] deliver better performance. However, neither of them presents promising generalization capabilities towards complex synthetic distortions in KADID-10k [27], demonstrating the challenges in handling cross-distortion scenarios.

Being separately trained on each individual dataset with a separate group of parameters, DBCNN [76], HyperIQA [55], and TreS [13] achieve promising performance on all six datasets. The pairwise learning-to-rank training strategy enables UNIQUE [78] and LIQE to learn from multiple datasets simultaneously with a single set of weights. Their competitive performance against the separately trained BIQA methods is clearly demonstrated. Moreover, LIQE outperforms UNIQUE with clear margins on the three datasets with realistic distortions [7, 11, 17] and KADID-10k [27] (commonly regarded as the most challenging BIQA benchmark with synthetic distortions), which verifies the effectiveness of the proposed multitask learning scheme based on the vision-language correspondence.

Table 2. SRCC results on the three IQA datasets under the cross-dataset setup. The subscripts "s" and "r" stand for models trained on KADID-10K [27] and KonIQ-10K [17], respectively. Best results are highlighted in bold.

Dataset	TID2013 [47]	SPAQ [9]	PIPAL [14]
NIQE	0.314	0.578	0.153
DBCNN$_r$	0.471	0.801	0.413
DBCNN$_s$	0.686	0.412	0.321
PaQ2PiQ	0.423	0.823	0.400
MUSIQ$_r$	0.584	0.853	0.450
UNIQUE	0.768	0.838	0.444
LIQE	**0.811**	**0.881**	**0.478**

4.3. Cross-Dataset Evaluation

We compare the generalizability of LIQE against competitive BIQA models in a more challenging cross-dataset setting. Specifically, we employ the full TID2013 [47] and SPAQ [9] as the test sets, which contain synthetic and realistic camera distortions, respectively. In addition, we also test BIQA models using images from the training set[2] of PIPAL, whose MOSs are publicly available [14]. PIPAL gathers and annotates images enhanced by various image restoration algorithms. The introduced algorithm-dependent distortions, especially those arising from generative adversarial network (GAN)-based methods, appear quite different from those covered in the six datasets described in Sec. 4.1. It is clear from Table 2 that only UNIQUE [78] and LIQE are capable of handling both synthetic and realistic distortions well with single sets of parameters, which highlights the promise of the joint training on multiple datasets. Due to the significant distributional shifts in distortion patterns, none of the tested methods presents promising results on PIPAL [14], suggesting that task-specific model training may be a viable option to handle such algorithm-dependent distortions. Nevertheless, LIQE achieves a slightly higher SRCC result, which we believe is attributed to the better commonsense knowledge of visual quality learned from the vision-language correspondence.

4.4. Realignment of Quality Annotations

Despite the promising performance on each individual dataset, it is unclear whether the proposed LIQE suffices to realign well the quality annotations from different IQA

[2] We are intended to participate in the NTIRE challenge competition, where evaluations on the validation and test sets can only be performed online by registered participants.

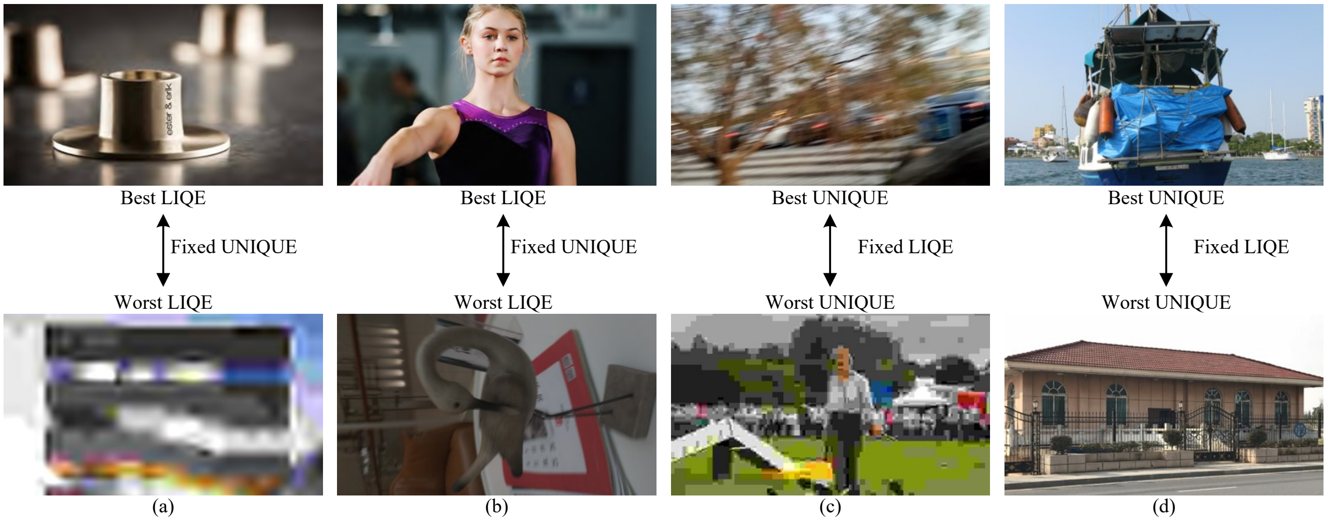

Figure 3. gMAD competition results between UNIQUE [78] and LIQE. **(a)** Fixed UNIQUE at the low-quality level. **(b)** Fixed UNIQUE at the high-quality level. **(c)** Fixed LIQE at the low-quality level. **(d)** Fixed LIQE at the high-quality level.

Table 3. Median SRCC results across ten sessions in the multi-dataset realignment experiment.

Method	UNIQUE	LIQE
SRCC	0.851	0.879

datasets in a common perceptual scale. In this subsection, we conduct two sessions of experiments to probe this question both qualitatively and quantitatively.

Qualitative Results. We rely on the gMAD [35] competition to obtain qualitative results, which seeks pairs of images that are estimated to be of similar quality by one model but of substantially different quality according to another model. As a result, at least one of the two models will be falsified with inconsistent judgments against human opinions. We combine the full Waterloo Exploration Database [36] and SPAQ [9] to build a gMAD playground with both realistic and synthetic distortions. We let LIQE compete against UNIQUE [78], both of which are trained jointly on all six datasets stated in Sec. 4.1. As shown in Fig. 3, UNIQUE makes similar quality predictions for pairs of images in (a) and (b), which is clearly inconsistent with human opinions as well as quality predictions by LIQE. When the roles of the two models switch, UNIQUE claims that the top image in the pair of (c) is of higher quality than the bottom one. However, both images in pair (c) are of low quality to humans, which are contaminated by realistic motion blur and synthetic JPEG compression, respectively. In contrast, LIQE makes a successful defense by rating both images in pair (c) as poor perceived quality. Similar conclusions can be drawn from pair (d).

Quantitative Results. MOSs of images obtained in different sessions of subjective testing [54] or via different testing methodologies [48] are not directly comparable. This hinders direct combination of MOSs from different datasets for training and evaluation. To test quantitatively the perceptual scale realignment performance of LIQE, we follow the method in [54], and sample 150 images from the six datasets, on which we conduct a separate perceptual scale realignment experiment to obtain the realigned MOSs. After that, we fit six monotonic nonlinear mapping functions to compute the realigned MOSs for all six datasets separately. We list in Table 3 the median SRCC results of LIQE and UNIQUE [78] on the same test sets used in Sec. 4.2. We find that LIQE achieves a higher SRCC.

In summary, we show convincingly that the proposed multitask learning scheme enabled by the vision-language correspondence improves the perceptual scale realignment performance on top of the pairwise learning-to-rank training strategy (used by both UNIQUE and LIQE).

4.5. Ablation Studies

We conduct a series of ablation studies to verify the design rationality of LIQE following the setups in Sec. 4. We first (1) implement a pre-trained CLIP baseline which adopts a Likert-scale of two quality levels as in [62], *i.e.*, $c \in \mathcal{C} = \{1, 2\} = \{$“bad”, “good”$\}$, and also (2) fine-tune it on IQA datasets. The subsequent ablations adopt the same multitask learning scheme as LIQE, while differing in three alternative design choices: (3) freezing the language encoder $\boldsymbol{g}_{\varphi}$ during training; (4) training with three separate textual templates with respect to quality prediction, scene classification, and distortion type identification, respectively; (5) using equal task weightings instead of the dynamic loss weighting. As a reference, we also (6) fine-

Table 4. Median SRCC results of LIQE variants across ten sessions. The top results are highlighted in bold.

Variants	LIVE	CSIQ	KADID-10k	BID	CLIVE	KonIQ-10k	Mean
(1) Pretrained CLIP ($C = 2$)	0.801	0.777	0.642	0.668	0.595	0.677	0.693
(2) Fine-Tuned CLIP ($C = 2$)	0.934	0.900	0.901	0.841	0.868	0.898	0.890
(3) Frozen Textual Encoder $\boldsymbol{g}_{\varphi}$	0.968	0.933	0.926	0.866	0.901	0.916	0.918
(4) Separate Task Templates	0.969	0.937	0.924	0.856	0.903	0.905	0.916
(5) Equal Task Weightings	0.968	**0.938**	0.927	0.870	0.899	0.917	0.920
(6) Fine-Tuned Visual Encoder $\boldsymbol{f}_{\phi}$ Only	0.969	0.923	0.921	0.863	0.901	0.910	0.915
LIQE	**0.970**	0.936	**0.930**	**0.875**	**0.904**	**0.919**	**0.922**

Table 5. Median mean (mSRCC) and mean accuracy (mACC) results of LIQE variants across ten sessions, trained on different task combinations of the six datasets. The subscripts "s" and "r" stand for the accuracy of scene classification and distortion type identification, respectively.

Task Combination	SRCC	ACC$_s$	ACC$_d$
Quality	0.915	–	–
Scene	–	0.908	–
Distortion	–	–	0.847
Quality + Scene	0.915	0.898	–
Quality + Distortion	0.920	–	0.829
All Tasks (LIQE)	0.922	0.894	0.831

tune the visual encoder $\boldsymbol{f}_{\phi}$ using the training strategy in [78] with a fully connected layer attached to the visual encoder. From Table 4, we observe that the pre-trained CLIP model is not good at BIQA, whose performance can be improved by finetuning on IQA datasets. Performance is negatively affected when the language encoder is frozen, which may be because quality-related concepts have not been sufficiently captured during the pre-training stage of CLIP. We also find that LIQE works better with the Likert-scale of five quality levels than two. The full method performs favorably against the variant trained with three separate textual templates, verifying the effectiveness of the proposed joint probability formulation. It is desirable to also include the dynamic loss weighting scheme, which not only achieves improvements over adopting equal task weightings but also liberates us from laborious hyperparameter tuning. Finally, the superiority of LIQE over the fine-tuned visual encoder corroborates our major contribution of multitask learning via the vision-language correspondence.

4.6. Task Relationship Analysis

We train LIQE variants on different combinations of tasks, and summarize the mean SRCC and accuracy results over six datasets in Table 5, from which we have three observations. First, including the task of distortion type identification is beneficial for BIQA, reflecting the cooperative relationship between them. Second, as a conceptually conflicting task, bringing only the scene classification task in neither improves nor impairs BIQA; however, the best mSRCC result is achieved when LIQE is trained on all three tasks, suggesting that the proposed multitask learning scheme successfully leverages an intermediate task like distortion type identification as a bridge in soliciting contributory features. Third, in turn, training with BIQA neither improves scene classification nor distortion type identification, indicating that the visual encoder representation optimized for BIQA moves away from the pre-trained CLIP, which are more transferable to the other two tasks.

5. Conclusion

We have formulated BIQA from a multitask learning perspective via the vision-language correspondence. During training, we simultaneously optimized a pair of image and language encoders on multiple IQA datasets for BIQA, scene classification, and distortion type identification jointly. We designed three fidelity losses to train the model, and employed a simple and efficient dynamic weighting scheme to automate the weighted summation of the multi-task losses. We presented the effectiveness of the proposed LIQE, and verified the rationality of various design choices. We also showed that the learned model realigns MOSs from different datasets in a more perceptually meaningful way. We believe the proposed multitask learning perspective will shed some light on the development of next-generation BIQA models as well as computational models for other machine vision applications.

Acknowledgments

The authors would like to thank Chengguang Zhu for coordinating the psychophysical experiment. This work was supported in part by Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), the Fundamental Research Funds for the Central Universities, and the National Natural Science Foundation of China under Grants 61901262 and 62071407.

References

[1] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Multi-task feature learning. *Advances in Neural Information Processing Systems*, 19, 2006. 3

[2] Shahrukh Athar and Zhou Wang. A comprehensive performance evaluation of image quality assessment algorithms. *IEEE Access*, 7:140030–140070, Sep. 2019. 1

[3] Mathieu Blondel, Olivier Teboul, Quentin Berthet, and Josip Djolonga. Fast differentiable sorting and ranking. In *International Conference on Machine Learning*, pages 950–959, 2020. 2

[4] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing*, 27(1):206–219, Jan. 2018. 1, 6

[5] Matthew R. Boutell, Jiebo Luo, Xipeng Shen, and Christopher M. Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, Sep. 2004. 2

[6] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International Conference on Machine Learning*, pages 794–803, 2018. 3

[7] Alexandre Ciancio, A. L. N. T. Targino da Costa, E. A. B. da Silva, Amir Said, Ramin Samadani, and Pere Obrador. No-reference blur assessment of digital pictures based on multi-feature classifiers. *IEEE Transactions on Image Processing*, 20(1):64–75, Jan. 2011. 2, 5, 6

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2

[9] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3677–3686, 2020. 2, 6, 7

[10] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135, 2017. 3

[11] Deepti Ghadiyaram and Alan C. Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, Jan. 2016. 2, 3, 5, 6

[12] Deepti Ghadiyaram and Alan C. Bovik. Perceptual quality prediction on authentically distorted images using a bag of features approach. *Journal of Vision*, 17(1):32–32, Jan. 2017. 2

[13] S. Alireza Golestaneh, Saba Dadsetan, and Kris M. Kitani. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1220–1230, 2022. 2, 6

[14] Jinjin Gu, Haoming Cai, Haoyu Chen, Xiaoxing Ye, Jimmy S. Ren, and Chao Dong. PIPAL: a large-scale image quality assessment dataset for perceptual image restoration. In *European Conference on Computer Vision*, pages 633–651, 2020. 6

[15] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2022. 3

[16] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *CoRR*, abs/1903.12261, 2019. 2

[17] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, Jan. 2020. 1, 2, 5, 6

[18] Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1733–1740, 2014. 2

[19] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *IEEE International Conference on Computer Vision*, pages 5148–5157, 2021. 2, 6

[20] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7482–7491, 2018. 3

[21] Iasonas Kokkinos. UberNet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6129–6138, 2017. 3

[22] Eric C. Larson and Damon M. Chandler. Most apparent distortion: Full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):1–21, Jan. 2010. 2, 5, 6

[23] Boyi Li, Kilian Q. Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022. 3

[24] Dingquan Li, Tingting Jiang, and Ming Jiang. Norm-in-norm loss with faster convergence and better performance for image quality assessment. In *ACM International Conference on Multimedia*, pages 789–797, 2020. 2

[25] Dingquan Li, Tingting Jiang, Weisi Lin, and Ming Jiang. Which has better visual quality: The clear blue sky or a blurry animal? *IEEE Transactions on Multimedia*, 21(5):1221–1234, May 2019. 2

[26] Liunian H. Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 3

[27] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. KADID-10k: A large-scale artificially distorted IQA database. In *International Conference on Quality of Multimedia Experience*, pages 1–3, 2019. 2, 5, 6

[28] Xi Lin, Hui-Ling Zhen, Zhenhua Li, Qing-Fu Zhang, and Sam Kwong. Pareto multi-task learning. In *Advances in*

Neural Information Processing Systems, volume 32, pages 12037–12047, 2019. 3

[29] Jianzhao Liu, Wei Zhou, Xin Li, Jiahua Xu, and Zhibo Chen. LIQA: Lifelong blind image quality assessment. *IEEE Transactions on Multimedia*, to appear, 2022. 2

[30] Shikun Liu, Stephen James, Andrew J. Davison, and Edward Johns. Auto-Lambda: Disentangling dynamic task relationships. *Transactions on Machine Learning Research*, 2022. 3

[31] Shikun Liu, Edward Johns, and Andrew J. Davison. End-to-end multi-task learning with attention. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1871–1880, 2019. 2, 3, 5

[32] Xialei Liu, Joost van de Weijer, and Andrew D. Bagdanov. RankIQA: Learning from rankings for no-reference image quality assessment. In *IEEE International Conference on Computer Vision*, pages 1040–1049, 2017. 1

[33] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. 6

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 6

[35] Kede Ma, Zhengfang Duanmu, Zhou Wang, Qingbo Wu, Wentao Liu, Hongwei Yong, Hongliang Li, and Lei Zhang. Group maximum differentiation competition: Model comparison with few samples. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):851–864, Apr. 2020. 2, 7

[36] Kede Ma, Zhengfang Duanmu, Qingbo Wu, Zhou Wang, Hongwei Yong, Hongliang Li, and Lei Zhang. Waterloo Exploration Database: New challenges for image quality assessment models. *IEEE Transactions on Image Processing*, 26(2):1004–1016, Feb. 2017. 7

[37] Kede Ma, Wentao Liu, Kai Zhang, Zhengfang Duanmu, Zhou Wang, and Wangmeng Zuo. End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing*, 27(3):1202–1213, Mar. 2018. 1, 2

[38] Kede Ma, Xuelin Liu, Yuming Fang, and Eero P. Simoncelli. Blind image quality assessment by learning from multiple annotators. In *IEEE International Conference on Image Processing*, pages 2344–2348, 2019. 1, 6

[39] Rui Ma, Hanxiao Luo, Qingbo Wu, King Ngi Ngan, Hongliang Li, Fanman Meng, and Linfeng Xu. Remember and reuse: Cross-task blind image quality assessment via relevance-aware incremental learning. In *ACM International Conference on Multimedia*, pages 5248–5256, 2021. 2

[40] Pavan C. Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C. Bovik. Image quality assessment using contrastive learning. *IEEE Transactions on Image Processing*, 31:4149–4161, Jun. 2022. 1

[41] Arun Mallya, Dillon Davis, and Svetlana Lazebnik. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *European Conference on Computer Vision*, pages 72–88, 2018. 3

[42] R. Timothy Marler and Jasbir S. Arora. Survey of multi-objective optimization methods for engineering. *Structural and Multidisciplinary Optimization*, 26(6):369–395, Mar. 2004. 3

[43] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3994–4003, 2016. 3

[44] Anish Mittal, Anush K. Moorthy, and Alan C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, Dec. 2012. 2

[45] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, Mar. 2013. 2, 6

[46] Anush K. Moorthy and Alan C. Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Transactions on Image Processing*, 20(12):3350–3364, Dec. 2011. 2

[47] Ponomarenko Nikolay, Jin Lina, Ieremeiev Oleg, Lukin Vladimir, Egiazarian Karen, Astola Jaakko, Vozel Benoit, Chehdi Kacem, Carli Marco, Battisti Federica, and C.-C. Jay Kuo. Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30:57–77, Jan. 2015. 6

[48] Maria Perez-Ortiz, Aliaksei Mikhailiuk, Emin Zerman, Vedad Hulusic, Giuseppe Valenzise, and Rafał K Mantiuk. From pairwise comparisons and rating to a unified quality scale. *IEEE Transactions on Image Processing*, 29:1139–1151, Aug. 2020. 2, 7

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021. 2, 4, 6

[50] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 6

[51] Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, and Anders Søgaard. Latent multi-task architecture learning. In *AAAI Conference on Artificial Intelligence*, pages 4822–4829, 2019. 3

[52] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, pages 525–536, 2018. 3

[53] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Association for Computational Linguistics*, pages 16–1162, 2016. 4

[54] Hamid R. Sheikh, Muhammad F. Sabir, and Alan C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15(11):3440–3451, Nov. 2006. 2, 3, 5, 6, 7

[55] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In

IEEE Conference on Computer Vision and Pattern Recognition, pages 3667–3676, 2020. 2, 6

[56] Ximeng Sun, Rameswar Panda, Rogerio Feris, and Kate Saenko. AdaShare: Learning what to share for efficient deep multi-task learning. *Advances in Neural Information Processing Systems*, 33:8728–8740, 2020. 3

[57] Louis L. Thurstone. A law of comparative judgment. *Psychological Review*, 34:273–286, Jul. 1927. 4

[58] Ming-Feng Tsai, Tie-Yan Liu, Tao Qin, Hsin-Hsi Chen, and Wei-Ying Ma. FRank: A ranking method with fidelity loss. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 383–390, 2007. 2, 4

[59] Simon Vandenhende, Stamatios Georgoulis, Luc Van Gool, and Bert De Brabandere. Branched multi-task networks: Deciding what layers to share. In *British Machine Vision Conference*, 2020. 3

[60] Yael Vinker, Ehsan Pajouheshgar, Jessica Y. Bo, Roman Christian Bachmann, Amit H. Bermano, Daniel Cohen-Or, Amir Zamir, and Ariel Shamir. CLIPasso: semantically-aware object sketching. *ACM Transactions on Graphics*, 41(4):86:1–86:11, Aug. 2022. 3

[61] Matthew Wallingford, Hao Li, Alessandro Achille, Avinash Ravichandran, Charless Fowlkes, Rahul Bhotika, and Stefano Soatto. Task adaptive parameter sharing for multi-task learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7561–7570, 2022. 3

[62] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. *CoRR*, abs/2207.12396, 2022. 3, 7

[63] Zhou Wang and Alan C. Bovik. *Modern Image Quality Assessment*. Morgan & Claypool, 2006. 1

[64] Zhou Wang, 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(4):600–612, Apr. 2004. 1

[65] Zhihua Wang and Kede Ma. Active fine-tuning from gMAD examples improves blind image quality assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4577–4590, 2022. 2

[66] Zhihua Wang, Haotao Wang, Tianlong Chen, Zhangyang Wang, and Kede Ma. Troubleshooting blind image quality models in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 16256–16265, 2021. 1, 2

[67] Jinjian Wu, Jupo Ma, Fuhu Liang, Weisheng Dong, Guangming Shi, and Weisi Lin. End-to-end blind image quality prediction with cascaded deep neural network. *IEEE Transactions on Image Processing*, 29:7414–7426, Jun. 2020. 1

[68] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. GroupViT: Semantic segmentation emerges from text supervision. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022. 3

[69] Jingtao Xu, Peng Ye, Qiaohong Li, Haiqing Du, Yong Liu, and David Doermann. Blind image quality assessment based on high order statistics aggregation. *IEEE Transactions on Image Processing*, 25(9):4444–4457, Sep. 2016. 2

[70] Sidi Yang, 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1191–1200, 2022. 2

[71] Peng Ye, Jayant Kumar, Le Kang, and David Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1098–1105, 2012. 2

[72] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan C. Bovik. From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3572–3582, 2020. 2, 6

[73] Lin Zhang, Lei Zhang, and Alan C. Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, Aug. 2015. 6

[74] Weixia Zhang, Dingquan Li, Chao Ma, Guangtao Zhai, Xiaokang Yang, and Kede Ma. Continual learning for blind image quality assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):2864–2878, 2023. 2

[75] Weixia Zhang, Dingquan Li, Xiongkuo Min, Guangtao Zhai, Guodong Guo, Xiaokang Yang, and Kede Ma. Perceptual attacks of no-reference image quality models with human-in-the-loop. In *Advances in Neural Information Processing Systems*, pages 2916–2929, 2022. 2

[76] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1):36–47, Jan. 2020. 1, 2, 6

[77] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. Task-specific normalization for continual learning of blind image quality models. *CoRR*, abs/2107.134290, 2021. 2

[78] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Transactions on Image Processing*, 30:3474–3486, Mar. 2021. 1, 2, 4, 6, 7, 8

[79] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, Aug. 2022. 3

[80] Hancheng Zhu, Leida Li, Jinjian Wu, Weisheng Dong, and Guangming Shi. MetaIQA: Deep meta-learning for no-reference image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 14131–14140, 2020. 1, 2