# Deep Opinion-Unaware Blind Image Quality Assessment by Learning and Adapting from Multiple Annotators

**Zhihua Wang**[1,2] , **Xuelin Liu**[3] , **Jiebin Yan**[3] , **Jie Wen**[4] , **Wei Wang**[1] , **Chao Huang**[1]*

[1] Sun Yat-sen University, Shenzhen Campus
[2] City University of Hong Kong
[3] Jiangxi University of Finance and Economics
[4] Harbin Institute of Technology, Shenzhen

zhihua.wang@my.cityu.edu.hk , {xuelinliu-bill, jiebinyan}@foxmail.com
jiewen_pr@126.com , {wangwei29, huangch253}@mail.sysu.edu.cn

## Abstract

Existing deep neural network (DNN)-based blind image quality assessment (BIQA) methods primarily rely on human-rated datasets for training. However, collecting human labels is extremely time-consuming and labor-intensive, posing a significant bottleneck for practical applications. To address this challenge, we propose a **D**eep opinion-**U**naware **BI**QA model by learning and adapting from **M**ultiple **A**nnotators, termed DUBMA, thereby eliminating the need for human annotations. Specifically, we first generate a large-scale set of distorted image pairs and then assign relative quality rankings using existing full-reference IQA models. The resulting dataset is subsequently employed for training a DNN-based BIQA method. Due to the inherent discrepancies between synthetic and real-world distortions, a domain shift may occur. To address this, we propose an outlier-robust unsupervised domain adaptation approach leveraging optimal transport. This strategy effectively reduces the gap between synthetic and real-world distortion domains, thereby boosting the model's adaptability and overall performance. Extensive experiments show that DUBMA outperforms existing opinion-unaware BIQA methods in terms of prediction accuracy across multiple datasets.

**Code**-https://github.com/SMBU-MM/DUMBA

## 1 Introduction

Most reliable methods for automatic image quality assessment (IQA) are "full-reference", relying on the availability of the reference pristine image [Wang *et al.*, 2004]. Nonetheless, in many practical applications, the reference image is not accessible or even not in existence, necessitating the use of no-reference or blind IQA (BIQA) models [Mittal *et al.*, 2012]. The working mechanism of BIQA is close to our vision system since humans are able to assess the perceptual quality of distorted images without referring to their pristine counterparts. Presumably, our visual systems have, through evolutionary and developmental processes, learned to preferentially encode the prior knowledge of the "naturalness" of visual images - those may arise from the physical interactions of light, surfaces, and optics. Early attempts at BIQA models are mainly based on natural scene statistics (NSS), with the underlying assumption that natural images are innate with a number of quality-aware consistent statistical regularities. Spatially normalized coefficients and codebook-based representations are two representative pipelines that have achieved impressive performance on generic distortion types, *e.g.* blur, noise, and JPEG compression [Mittal *et al.*, 2012].

When generalizing to in-the-wild distortions, statistical regularities may not accurately model these distortions, resulting in poor performance. Recently, BIQA methods based on deep neural networks (DNNs) have shown remarkable performance, surpassing traditional NSS-based methods [Zhang *et al.*, 2018]. This success is attributed to the ability of DNNs to automatically learn informative feature representations for quality prediction. Training DNNs with millions of parameters typically requires extensive human quality annotations, such as mean opinion scores (MOSs). The acquisition of large-scale human-labeled datasets is both laborious and costly. For instance, the KonIQ-10k dataset [Hosu *et al.*, 2020] comprises over 10,000 images annotated by nearly 1,500 subjects, resulting in 1.2 million human ratings. To mitigate the reliance on human annotations, some studies explore the use of pseudo-labeled datasets to train DNN-based BIQA [Ye *et al.*, 2014; Ma *et al.*, 2017]. These approaches effectively mitigate the reliance on human annotations.

In this study, we further explore the use of synthetically distorted images with pseudo labels [Ma *et al.*, 2019] for BIQA by proposing a **D**eep opinion-**U**naware **BI**QA model that learns and adapts from **M**ultiple **A**nnotators, termed DUBMA. Specifically, we begin by generating a large-scale set of distorted image pairs, for which a collection of full-reference IQA (FR-IQA) models provides relative quality rankings. We then employ a learn-to-rank (L2R) approach [Wang and Ma, 2021] to learn from these FR-IQA models. During the training of DUBMA, each annotator is assigned a distinct reliability score, which is implicitly estimated via

[*] Corresponding author.

maximum likelihood estimation (MLE), thereby enabling robust label fusion. Previous studies, such as BLISS [Ye *et al.*, 2014] and Wu *et al.*[Wu *et al.*, 2020], have investigated the use of full-reference (FR)-IQA methods to generate quality scores for unlabeled images. BLISS employs unsupervised rank aggregation to combine scores from multiple FR-IQA models, while Wu *et al.*[Wu *et al.*, 2020] scale predictions from various FR-IQA models to a common range to produce pseudo-MOS. However, these approaches have a significant limitation: they do not consider the reliability of individual FR-IQA models, which can introduce noise into the aggregated scores due to varying prediction accuracies across different distortion types.

Due to the inherent discrepancies between synthetic and real-world distortions, a domain shift may arise, causing BIQA models trained solely on synthetic distortions to struggle in accurately assessing realistic camera distortions [Zhang *et al.*, 2021]. To overcome this challenge, we adopt unsupervised domain adaptation (UDA) methods [Borgwardt *et al.*, 2006], which have demonstrated great potential for enhancing model performance on unlabeled target domains by effectively transferring knowledge from source domains. At the heart of adapting from synthetic distortions (as the source) to realistic distortions (as the target) is the usage of optimal transport (OT) distances to quantify the distance between distortions, more precisely, the distance between feature-label pairs across domains [Flamary *et al.*, 2017]. This property is particularly applicable to our scenario, since it is difficult to ensure that the quality range of the synthetic images completely covers realistic ones.

In summary, this work makes two primary contributions:

- We propose DUBMA, an opinion-unaware BIQA method trained on synthetic data pseudo-labeled by multiple FR-IQA models. The framework incorporates a L2R strategy, assigning distinct reliability scores to each annotator to enable joint learning from multiple noisy FR-IQA models. The resulting model achieves robust performance on synthetic distortions.
- We integrate an outlier-robust UDA approach to adapt BIQA models trained on synthetic data for realistic distortions. Using unbalanced OT distances, we quantify discrepancies between features and labels across domains. The adapted models demonstrate strong generalization and effective handling of real-world distortions.
- We conduct extensive experiments across multiple human-rated IQA datasets and perform detailed ablation studies to validate the effectiveness and robustness of our opinion-unaware BIQA learner in challenging real-world scenarios.

## 2 Related Work

### 2.1 Opinion-Unaware BIQA

Early opinion-unaware BIQA methods primarily rely on quality-aware natural scene statistics (NSS) features. Mittal *et al.* [Mittal *et al.*, 2011] introduced TMIQ, which assesses image quality by comparing the similarity of NSS features between distorted and pristine images. Later, Mittal *et al.* [Mittal *et al.*, 2012] developed NIQE, which models the distribution of natural NSS features using a multivariate Gaussian (MVG) model. Similarly, Zhang *et al.* [Zhang *et al.*, 2015] extended the quality-aware NSS features to propose IL-NIQE. With the advent of DNNs, some studies have explored learning from pseudo-labels to train opinion-unaware BIQA models. Ye *et al.* [Ye *et al.*, 2014] introduced BLISS, which utilizes synthetic scores from full-reference IQA methods to train BIQA models. Liu *et al.* [Liu *et al.*, 2017] proposed RankIQA, where distortion intensities are used to determine the ranking order of image pairs. However, these methods, mainly designed for synthetic distortions, struggle with generalizing to realistic ones. In contrast, Zhu *et al.* [Zhu *et al.*, 2021] employed the trained discriminator of a Wasserstein GAN for quality assessment of realistic distortions. While they suggest that the discriminator may capture real-world image distributions, the performance remains suboptimal.

### 2.2 Deep Unsupervised Domain Adaptation

Popular UDA methods are primarily classified into statistical moment matching [Borgwardt *et al.*, 2006; Sun *et al.*, 2016; Damodaran *et al.*, 2018], domain style transfer, self-training, and feature-level adversarial learning [Ganin *et al.*, 2016; Shen *et al.*, 2018]. The testing benchmarks for deep UDA mainly focus on the image classification task. In the IQA/VQA setting, Chen *et al.* [Chen *et al.*, 2021a] developed the first Maximum Mean Discrepancy (MMD)-based domain adaptation for the quality assessment of screen content images (SCIs). Chen *et al.* [Chen *et al.*, 2021b] proposed a curriculum-style unsupervised domain adaptation to handle the cross-domain problem for no-reference video quality assessment. Recently, optimal transport (OT) distance, particularly Wasserstein distance [Flamary *et al.*, 2017], has gained attention for UDA. Flamary et al. [Flamary *et al.*, 2017] proposed a regularized OT method for aligning representations between source and target domains. However, OT treats all samples, including outliers, equally due to marginal constraints, which can lead to outliers disproportionately affecting the total mass. To address this, Benamou et al. [Benamou, 2003] introduced a penalization term to relax these constraints, resulting in unbalanced OT for handling unbalanced data. Furthermore, Fatras et al. [Fatras *et al.*, 2021] showed that unbalanced OT reduces bias caused by minibatches.

## 3 Proposed Method

Figure 1 illustrates the overall framework of DUBMA. Our learning paradigm assumes the existence of a training set consisting of $n_s$ synthetic image pairs $\mathcal{D}_s = \{(\boldsymbol{x}_s^{(i)}, \boldsymbol{y}_s^{(i)})\}_{i=1}^{n_s}$ and a test set $\mathcal{D}_t = \{\boldsymbol{x}_t^{(j)}\}_{j=1}^{n_t}$ of $n_t$ images. Given an instance $\boldsymbol{x}_s \in \mathcal{D}_s$, its true perceptual quality is denoted by $q(\boldsymbol{x}_s)$. We leverage $M$ objective IQA models $\{q_m\}_{m=1}^{M}$ to compute the quality prediction of $q(\boldsymbol{x}_s)$, which are collectively denoted by $\{q_m(\boldsymbol{x}_s)\}_{m=1}^{M}$. Since different objective IQA models may exhibit varying degrees of non-linearity and scale, we adopt the L2R strategy, where pseudo labels are organized in binary form to avoid being affected by non-linearity and scale, thereby mitigating these effects [Zhang *et*

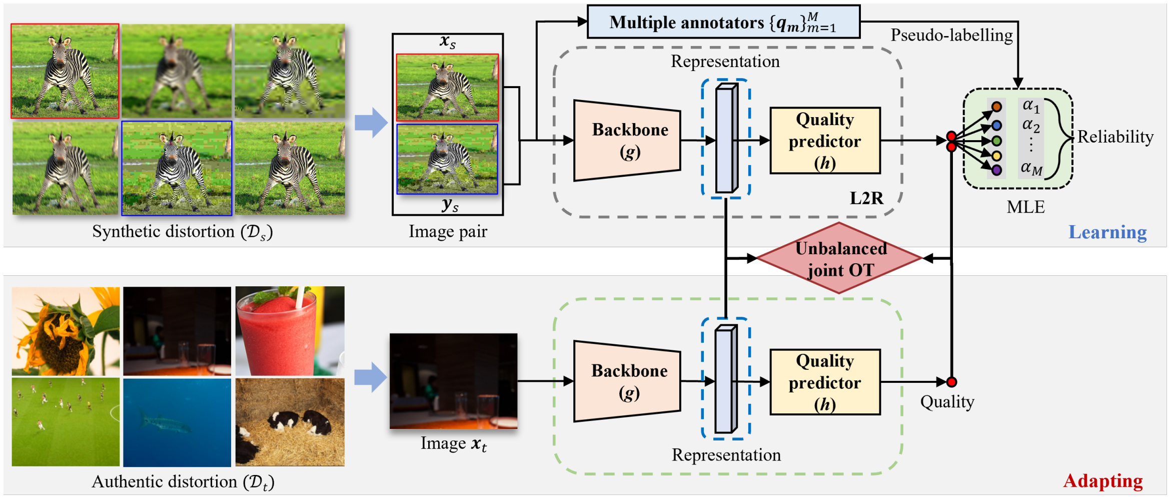

Figure 1: Framework of DUBMA. The first part illustrates the process of learning from synthetically-distorted images and evaluating their performance, while the entire framework encompasses learning from simulated distortions and adapting to the quality evaluation of authentically-distorted images.

*al.*, 2021]. Specifically, given an image pair $(\boldsymbol{x}_s, \boldsymbol{y}_s) \in \mathcal{D}_s$, we assign a binary label $r_m$ for this pair, where $r_m = 1$ if $q_m(\boldsymbol{x}_s) > q_m(\boldsymbol{y}_s)$ and $r_m = 0$ otherwise. Therefore, the pseudo-labeled training set is in the form of $\mathcal{D}_s = \{(\boldsymbol{x}_s^{(i)}, \boldsymbol{y}_s^{(i)}), r_1^{(i)}, \cdots, r_M^{(i)}\}_{i=1}^{n_s}$. We aim at training a DNN-based BIQA method on $\mathcal{D}_s$ to reliably estimate the perceptual quality of the sample in $\mathcal{D}_t$. Such a function includes two mappings: a differentiable embedding function $\boldsymbol{g}$, which maps the input into the latent embedding, and the quality predictor $h$, which maps the embedding to the quality score.

### 3.1 DUMBA for Synthetic Distortions

Since the quality estimate of each IQA annotator is noisy and not equally good (or bad) at labeling the input, we introduce a set of the reliability parameters $\boldsymbol{\alpha} = \{\alpha_m\}_{m=1}^M$ to explicitly model the hit rate and correct reject rate of each annotator, which are denoted as:

$$\alpha_m = \Pr(r_m = 1|r = 1) = \Pr(r_m = 0|r = 0), \qquad (1)$$

where $r = 1$ if $q(\boldsymbol{x}_s) \geq q(\boldsymbol{y}_s)$ and $r = 0$ otherwise. Under the hypothesis of Thurstone's Case V model [Zhang *et al.*, 2021], we assume $q(\boldsymbol{x}_s) \sim \mathcal{N}(h(\boldsymbol{g}(\boldsymbol{x}_s)), 1)$. The perceptual difference between a stimuli pair $(\boldsymbol{x_s}, \boldsymbol{y_s})$ takes the form $q(\boldsymbol{x}_s) - q(\boldsymbol{y}_s) \sim \mathcal{N}(h(\boldsymbol{g}(\boldsymbol{x}_s)) - h(\boldsymbol{g}(\boldsymbol{y}_s)), 2)$. From that, the strengths (probability) of preferring $\boldsymbol{x}_s$ over $\boldsymbol{y}_s$ perceptually in an image pair $(\boldsymbol{x}_s, \boldsymbol{y}_s)$ can be computed from the Gaussian cumulative distribution function (CDF) $\Phi(\cdot)$, which has a closed form as:

$$p(\boldsymbol{x}_s, \boldsymbol{y}_s) = \Phi\left(\frac{h(\boldsymbol{g}(\boldsymbol{x}_s)) - h(\boldsymbol{g}(\boldsymbol{y}_s))}{\sqrt{2}}\right). \qquad (2)$$

During training, we sample a mini-batch $\mathcal{B}_s$ from $\mathcal{D}_s$ in each iteration and minimize the negative logarithm maximum likelihood function [Ma *et al.*, 2019] to obtain the optimal model parameters along with the reliability variables $\boldsymbol{\alpha} = \{\alpha_m\}_{m=1}^M$, which is defined as:

$$\ell_q(\mathcal{B}_s; h, \boldsymbol{g}, \boldsymbol{\alpha}) = -\log \Pr(\mathcal{B}_s; h, \boldsymbol{g}, \boldsymbol{\alpha}), \qquad (3)$$

where

$$\Pr(\mathcal{B}_s; h, \boldsymbol{g}, \boldsymbol{\alpha}) = \prod_{i=1}^N \Bigg( p(\boldsymbol{x}_s^{(i)}, \boldsymbol{y}_s^{(i)}) \prod_{m=1}^M \Pr(r_m^{(i)}|r = 1) + (1 - p(\boldsymbol{x}_s^{(i)}, \boldsymbol{y}_s^{(i)})) \prod_{m=1}^M \Pr(r_m^{(i)}|r = 0) \Bigg). \qquad (4)$$

The resulting DUBMA model is capable of performing reliable quality assessment for synthetic distortions.

### 3.2 DUMBA for Realistic Distortions

The distribution shift between synthetic and realistic distortions may lead to underperformance of DUBMA, which is trained on synthetic data, when applied to realistic distortions. To address this issue, we incorporate UDA to align the distribution shift (see Figure 1).

**Optimal Transport.** The OT problem is concerned with finding the minimum mass of transport required to transform a source probability distribution $\boldsymbol{P}_s$ to a target probability distribution $\boldsymbol{P}_t$. Specifically, we assume that $\delta(x)$ denotes the Dirac function at location $\boldsymbol{x}$, $p_s^{(i)}$ and $p_t^{(j)}$ respectively represent the probability masses associated with the $i$-th source element $\boldsymbol{x}_s^{(i)}$ and $j$-th target element $\boldsymbol{x}_t^{(j)}$, belonging to the probability simplex, *i.e.*, $\sum_{i=1}^{n_s} p_s^{(i)} = \sum_{j=1}^{n_t} p_t^{(j)} = 1$. The empirical marginals $\boldsymbol{p}_s$ and $\boldsymbol{p}_t$ are written as:

$$\boldsymbol{p}_s = \sum_{i=1}^{n_s} p_s^{(i)} \delta(\boldsymbol{x}_s^{(i)}), \quad \boldsymbol{p}_t = \sum_{j=1}^{n_t} p_t^{(j)} \delta(\boldsymbol{x}_t^{(j)}), \qquad (5)$$

where $n_s$ and $n_t$ represent the number of source and target samples. The Kantorovitch's relaxation of the OT problem in discrete setting is formulated to seek the optimal transport plans or couplings $\boldsymbol{\gamma}^\star$ as:

$$\boldsymbol{\gamma}^\star = \underset{\boldsymbol{\gamma} \in \Pi(\boldsymbol{p}_s, \boldsymbol{p}_t)}{\operatorname{argmin}} \ \langle \boldsymbol{\gamma}, \boldsymbol{C} \rangle_F, \qquad (6)$$

where $\Pi(\boldsymbol{p}_s, \boldsymbol{p}_t) = \left\{\boldsymbol{\gamma} \geq 0 | \boldsymbol{\gamma}\mathbf{1}_{n_s} = \boldsymbol{p}_s, \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{1}_{n_t} = \boldsymbol{p}_t\right\}$ ($\mathbf{1}_d$ is a $d$-dimensional vector of ones) is a set of all joint distributions (couplings), $\langle\cdot,\cdot\rangle_F$ denotes the Frobenius dot product, $C \geq 0$ is the $n_s \times n_t$ cost function matrix, whose term $C_{ij} = c(\boldsymbol{x}_s^{(i)}, \boldsymbol{x}_t^{(j)})$ measures the pairwise cost to move a probability mass from $\boldsymbol{x}_s^{(i)}$ to $\boldsymbol{x}_t^{(j)}$. The minimal cost can be used to quantify the distance between two distributions and is referred to as the Kantorovich–Rubinstein–Wasserstein distance, *i.e.*,

$$\begin{aligned} W_{\mathrm{c}}(\boldsymbol{p}_s, \boldsymbol{p}_t) &= \min_{\boldsymbol{\gamma}\in\Pi(\boldsymbol{p}_s,\boldsymbol{p}_t)} \langle\boldsymbol{\gamma}, C\rangle_F \\ \text{s.t.} \quad & \boldsymbol{\gamma} \geq 0, \boldsymbol{\gamma}\mathbf{1}_{n_s} = \boldsymbol{p}_s, \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{1}_{n_t} = \boldsymbol{p}_t. \end{aligned} \tag{7}$$

**Unbalanced Joint OT.** The fundamental idea behind joint distribution alignment is to address the variations in both feature space and conditional label space [Courty *et al.*, 2017; Damodaran *et al.*, 2018]. However, since true quality labels are not available in either the source or target domains, we substitute them with quality predictions. To be more specific, assuming we have access to the source and target datasets $\mathcal{D}_s$ and $\mathcal{D}_t$ respectively, we redefine the joint empirical marginals $\boldsymbol{p}_s$ and $\boldsymbol{p}_t$ as:

$$\begin{aligned} \boldsymbol{p}_s &= \sum_{i=1}^{n_s} p_s^{(i)} \delta(\boldsymbol{g}(\boldsymbol{x}_s^{(i)}), h(\boldsymbol{g}(\boldsymbol{x}_s^{(i)}))), \\ \boldsymbol{p}_t &= \sum_{j=1}^{n_t} p_t^{(j)} \delta(\boldsymbol{g}(\boldsymbol{x}_t^{(j)}), h(\boldsymbol{g}(\boldsymbol{x}_t^{(j)}))), \end{aligned} \tag{8}$$

where $\boldsymbol{g}$ and $h$ represent the feature extractor and quality estimator, respectively. The cost function $c$ in Eq (7) needs to measure the distance between the features and the discrepancy between predictions in combination, which is defined as:

$$\begin{aligned} c(\boldsymbol{x}_s, \boldsymbol{x}_t) &= \|\boldsymbol{g}(\boldsymbol{x}_s) - \boldsymbol{g}(\boldsymbol{x}_t)\|_2 \\ &+ \beta\left(1 - \sqrt{pr} - \sqrt{(1-p)(1-r)}\right), \end{aligned} \tag{9}$$

where the second term measures the fidelity between $p(\boldsymbol{x}_s, \boldsymbol{y}_s)$ and $r = 0.5$, and $\beta$ is the trade-off to balance two terms.

In practical OT optimization, each sample is weighted equally due to the marginal constraints. This means that $p_s^{(i)} = 1/n_s$ for all source samples and $p_t^{(j)} = 1/n_t$ for all target samples. However, OT is sensitive to outliers due to the distribution's geometry. In our framework, it is inevitable to generate synthetic data with distributions that are far away from realistic data. To overcome this issue, we can introduce a "soft" penalty term that relaxes the "hard" marginal constraints. This assigns smaller weights to outliers compared to in-distribution samples, effectively ignoring them. This formulation is known as unbalanced OT, which is defined as:

$$\begin{aligned} D_{ub}(\boldsymbol{p}_s, \boldsymbol{p}_t) &= \min_{\boldsymbol{\gamma}\in\Pi(\hat{\boldsymbol{p}}_s,\hat{\boldsymbol{p}}_t)} \langle\boldsymbol{\gamma}, C\rangle_F \\ &+ \lambda\left(KL(\hat{\boldsymbol{p}}_s\|\boldsymbol{p}_s) + KL(\hat{\boldsymbol{p}}_t\|\boldsymbol{p}_t)\right), \end{aligned} \tag{10}$$

where $\hat{\boldsymbol{p}}_s$ and $\hat{\boldsymbol{p}}_t$ are the plan's marginals of unbalanced OT, $\lambda$ is the tradeoff that represents the strength of penalization; $KL$ is the Kullback-Leibler (KL) divergence. It should be noted that the marginals of $\boldsymbol{\gamma}$ are no longer equal to $(\boldsymbol{p}_s\, \boldsymbol{p}_t)$ in general. In DNN training, OT is computed over minibatches, with results averaged to approximate the full dataset's OT. However, minibatch sampling can introduce deviations between the true OT loss and its estimate [Fatras *et al.*, 2021]. Unbalanced OT addresses this by relaxing marginal constraints, improving robustness to minibatch effects and scaling well with large datasets.

**Optimization of Unbalanced Joint OT.** Directly optimizing the unbalanced OT is nontrivial and ill-posed. We will reformulate the optimization problem (10) as a non-negative penalized linear regression problem [Chapel *et al.*, 2021]. Let $\boldsymbol{t} = \mathrm{vec}(\boldsymbol{\gamma}), \boldsymbol{c} = \mathrm{vec}(\boldsymbol{C})$ and $\boldsymbol{p}^\top = \left[\boldsymbol{p}_s^\top, \boldsymbol{p}_t^\top\right]$. Problem (10) can be re-written as:

$$\min_{\boldsymbol{t}\geq 0} F_\lambda(\boldsymbol{t}) \overset{\text{def}}{=} \frac{1}{\lambda}\boldsymbol{c}^\top\boldsymbol{t} + KL(\boldsymbol{H}\boldsymbol{t}, \boldsymbol{p}), \tag{11}$$

where $\boldsymbol{H}^\top = [\boldsymbol{H}_r^\top, \boldsymbol{H}_c^\top]$ concatenates the sums of the rows and columns of $\boldsymbol{\gamma}$, respectively; $KL(\boldsymbol{H}\boldsymbol{t}, \boldsymbol{p})$ denotes the data fitting term. This formulation is well-known in inverse problems and non-negative matrix factorization [Lee and Seung, 2000]. Here, we leverage the majorization-minimization (MM) algorithm to solve the problem in the form (11), which leads to efficient multiplicative updates for the KL penalty [Chapel *et al.*, 2021].

**Optimization Objective over Minibatches.** The proposed BIQA model is trained in an end-to-end manner using a variant of the stochastic gradient descent method. Specifically, we sample mini-batches $\mathcal{B}_s$ and $\mathcal{B}_t$ from $\mathcal{D}_s$ and $\mathcal{D}_t$, respectively, in each iteration. Then the empirical loss function on $\mathcal{B}_s$ and $\mathcal{B}_t$ is defined as:

$$\ell(\mathcal{B}_s, \mathcal{B}_t; h, \boldsymbol{g}) = \ell_q(\mathcal{B}_s; h, \boldsymbol{g}, \boldsymbol{\alpha}) + \eta\ell_d(\mathcal{B}_s, \mathcal{B}_t; h, \boldsymbol{g}), \tag{12}$$

where

$$\begin{aligned} \ell_d(\mathcal{B}_s, \mathcal{B}_t; h, \boldsymbol{g}) &= D_{ub}\left(\boldsymbol{p}_s, \boldsymbol{p}_t\right), \\ \boldsymbol{p}_s &= \sum_{i=1}^{|\mathcal{B}_s|} p_s^{(i)} \delta(\boldsymbol{g}(\boldsymbol{x}_s^{(i)}), h(\boldsymbol{g}(\boldsymbol{x}_s^{(i)}))), \\ \boldsymbol{p}_t &= \sum_{j=1}^{|\mathcal{B}_t|} p_t^{(j)} \delta(\boldsymbol{g}(\boldsymbol{x}_t^{(j)}), h(\boldsymbol{g}(\boldsymbol{x}_t^{(j)}))), \end{aligned}$$

$\ell_q$ is defined in Eq. (3), $p_s^{(i)} = 1/n_s$ for all $i$, $p_t^{(j)} = 1/n_t$ for all $j$, and $\eta$ trades off the two terms. In Eq. (12), there are two sets of parameters to optimize: the OT couplings $\boldsymbol{\gamma}$ and the model parameters $\{h, \boldsymbol{g}, \boldsymbol{\alpha}\}$. During training, we use the alternative minimization approach [Damodaran *et al.*, 2018] to update these parameters. Firstly, we solve the optimal transport plan $\boldsymbol{\gamma}^\star$ with fixed model parameters $\{h, \boldsymbol{g}\}$ using non-negative penalized linear regression. Then, we optimize the model parameters $\{h, \boldsymbol{g}, \boldsymbol{\alpha}\}$ using stochastic gradient descent with the obtained $\boldsymbol{\gamma}^\star$.

## 4 Experiments and Results

### 4.1 DUMBA for Synthetic Distortions

**Datasets.** We build the pseudo-labeled dataset $\mathcal{D}_s$ based on the Waterloo Exploration Database [Ma *et al.*, 2016]. We take

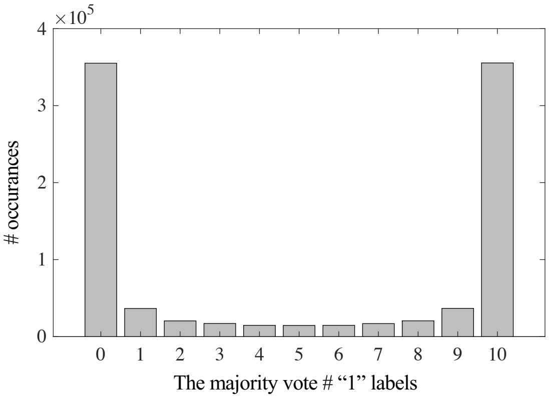

Figure 2: Histogram of the number of IQA models voting a "1".

these images as references and simulate twenty-five distortion types [1], each at five levels. We employ ten state-of-the-art (SOTA) FR-IQA models for pseudo labeling *i.e.*, SSIM [Wang *et al.*, 2004], MS-SSIM [Wang *et al.*, 2003], SR-SIM [Zhang and Li, 2012], FSIM [Zhang *et al.*, 2011], VSI [Zhang *et al.*, 2014], VIF [Sheikh and Bovik, 2006], GMSD [Xue *et al.*, 2013b], MDSI [Nafchi *et al.*, 2016], NLPD [Hadizadeh *et al.*, 2017], and A-DISTS [Ding *et al.*, 2021]. We select these models based on their performance evaluated by D-test, L-test, and P-test on the generated dataset [Ma *et al.*, 2016]. We generate over 90k image pairs and randomly select an appropriate number for training. Figure 2 illustrates the consistency of the IQA annotations across all pairs, counting the number of models that selected the first image in each pair as the higher quality one. Overall, the ten IQA annotators show high agreement, with complete consistency on roughly 80% of the image pairs. We evaluate the DUBMA performance on three commonly used subject-rated image sets - LIVE [Sheikh *et al.*, 2006], CSIQ [Larson and Chandler, 2010], and KADID-10k [Lin *et al.*, 2019].

**Details of Training and Testing.** We use ResNet-18 pretrained on ImageNet as the backbone $\boldsymbol{g}$ for feature extraction, followed by two fully connected layers with ReLU activation as the quality predictor $h$. The model is trained on $\mathcal{D}_s$ using the Adam optimizer, minimizing the loss in Eq. (3). The learning rates for $\boldsymbol{g}, h$ and $\alpha$ are set to $10^{-4}$ and $10^{-3}$, respectively, with a decay factor of 2 every two epochs. The batch size is 64, and training lasts for seven epochs. The best model is selected based on the average entropy of 5,000 randomly sampled image pairs from KADID-10k, with the model showing the least entropy chosen for performance comparison. We utilize the Spearman rank correlation coefficient (SRCC) and the Pearson linear correlation coefficient (PLCC) to quantitatively assess the model's performance. In addition to this, we employ the group MAximum Differentiation (gMAD) competition [Zhang *et al.*, 2021] to qualitatively evaluate method's generalizability.

| Metric | SRCC | | | PLCC | | |
|---|---|---|---|---|---|---|
| Dataset | LIVE | CSIQ | KADID-10k | LIVE | CSIQ | KADID-10k |
| PieAPP | 0.919 | 0.892 | 0.836 | 0.908 | 0.896 | 0.839 |
| LPIPS | 0.932 | 0.876 | 0.843 | 0.934 | 0.896 | 0.839 |
| DISTS | 0.954 | 0.929 | 0.887 | 0.954 | 0.928 | 0.886 |
| QAC | 0.868 | 0.490 | 0.239 | 0.863 | 0.708 | 0.390 |
| PIQE | 0.840 | 0.512 | 0.541 | 0.839 | 0.677 | 0.306 |
| LPSI | 0.818 | 0.522 | 0.148 | 0.826 | 0.718 | 0.443 |
| NIQE | 0.906 | 0.627 | 0.374 | 0.904 | 0.716 | 0.428 |
| ILNIQE | 0.898 | 0.815 | 0.531 | 0.903 | 0.854 | 0.573 |
| SISBLIM | 0.774 | 0.660 | 0.209 | 0.807 | 0.737 | 0.388 |
| SNP-NIQE | 0.907 | 0.609 | 0.371 | 0.766 | 0.731 | 0.422 |
| NPQI | 0.912 | 0.634 | 0.391 | 0.904 | 0.805 | 0.400 |
| RankIQA[1] | 0.897 | 0.808 | 0.569 | 0.891 | 0.832 | 0.569 |
| dipIQ | **0.938** | 0.527 | 0.304 | **0.935** | 0.779 | 0.402 |
| CaHDC | 0.928 | 0.758 | 0.540 | 0.918 | 0.827 | 0.574 |
| EONSS | 0.927 | 0.677 | 0.413 | 0.918 | 0.766 | 0.453 |
| Ma19 | 0.919 | **0.915** | 0.466 | 0.917 | **0.926** | 0.501 |
| Zhu21 | 0.135 | 0.391 | 0.185 | 0.517 | 0.481 | 0.342 |
| ContentSep | 0.748 | 0.587 | 0.506 | 0.700 | 0.589 | 0.486 |
| MDFS | **0.936** | 0.777 | **0.598** | **0.933** | 0.827 | **0.625** |
| DUBMA | 0.930 | **0.840** | **0.863** | 0.918 | **0.859** | **0.864** |

[1] https://github.com/YunanZhu/Pytorch-TestRankIQA.

Table 1: Correlations between model predictions and MOSs on LIVE, CSIQ, and KADID-10k, respectively. Bold indicates the top two results among BIQA models

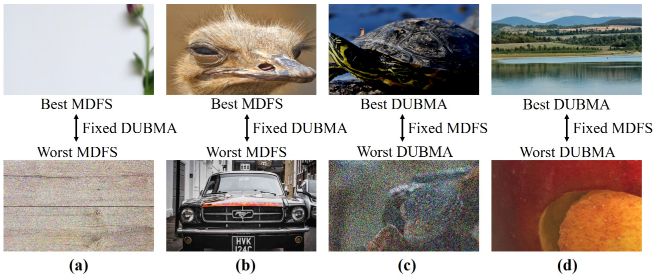

Figure 3: Representative gMAD pairs between DUBMA and MDFS [Ni *et al.*, 2024] on synthetic distortions. **(a)** Fixing DUBMA at the low-quality level. **(b)** Fixing DUBMA at the high-quality level. **(c)** Fixing the MDFS at the low-quality level. **(d)** Fixing the MDFS at the high-quality level.

**Quantitative Results.** We compare the proposed method with eight knowledge-driven BIQA methods - QAC [Xue *et al.*, 2013a], PIQE [Venkatanath *et al.*, 2015], LPSI [Wu *et al.*, 2015], NIQE [Mittal *et al.*, 2012], ILNIQE [Zhang *et al.*, 2015], SISBLIM [Gu *et al.*, 2014], SNP-NIQE [Liu *et al.*, 2019], NPQI [Liu *et al.*, 2020], and eight data-driven BIQA models - RankIQA [Liu *et al.*, 2017], dipIQ [Ma *et al.*, 2017], CaHDC [Wu *et al.*, 2020], EONSS [Wang *et al.*, 2019], Ma19 [Ma *et al.*, 2019] Zhu21 [Zhu *et al.*, 2021], ContentSep [Babu *et al.*, 2023], MDFS [Ni *et al.*, 2024]. All of these models are **opinion-unaware**. We also include three DNN-based full-reference IQA methods - PieAPP [Prashnani *et al.*, 2018], LPIPS [Zhang *et al.*, 2018], and DISTS [Ding *et al.*, 2022], trained on human-labeled datasets as reference. The implementations of all these methods are obtained from the respective authors. Table 1 compares our method with

[1] These include Gaussian blur, lens blur, motion blur, color diffusion, color shift, color quantization, color saturation 1/2, JPEG2000 compression, JPEG compression, white noise, whiter noise in color component, impulse noise, multiplicative noise, denoise, brighten, darken, mean shift, jitter, non-eccentricity patch, pixelate, quantization, color block, high sharpen, and contrast change.

| Metric | SRCC | | | PLCC | | |
|---|---|---|---|---|---|---|
| Dataset | LIVE | KonIQ-10k | SPAQ | LIVE | KonIQ-10k | SPAQ |
| QAC | 0.868 | 0.092 | 0.340 | 0.863 | 0.244 | 0.371 |
| PIQE | 0.840 | 0.245 | 0.232 | 0.839 | 0.210 | 0.251 |
| LPSI | 0.818 | 0.224 | 0.001 | 0.826 | 0.107 | 0.276 |
| NIQE | 0.906 | 0.530 | 0.703 | 0.904 | 0.538 | 0.712 |
| ILNIQE | 0.898 | 0.506 | 0.714 | 0.903 | 0.531 | 0.721 |
| SISBLIM | 0.774 | 0.616 | 0.701 | 0.807 | 0.619 | 0.718 |
| SNP-NIQE | 0.907 | 0.628 | 0.540 | 0.766 | 0.618 | 0.727 |
| NPQI | 0.912 | 0.613 | 0.600 | 0.904 | 0.556 | 0.627 |
| RankIQA | 0.897 | 0.483 | 0.584 | 0.891 | 0.482 | 0.587 |
| dipIQ | **0.938** | 0.236 | 0.385 | **0.935** | 0.435 | 0.497 |
| CaHDC | 0.928 | 0.423 | 0.562 | 0.918 | 0.441 | 0.594 |
| EONSS | 0.927 | 0.191 | 0.348 | 0.918 | 0.206 | 0.376 |
| Ma19 | 0.919 | 0.456 | 0.379 | 0.917 | 0.462 | 0.391 |
| Zhu21 | 0.135 | 0.636 | 0.683 | 0.517 | 0.641 | 0.688 |
| ContentSep | 0.748 | 0.640 | 0.708 | 0.700 | 0.645 | 0.706 |
| MDFS | **0.936** | **0.733** | **0.741** | **0.933** | **0.737** | **0.754** |
| DUBMA | 0.928 | **0.703** | **0.834** | 0.924 | **0.740** | **0.841** |

Table 2: Correlations between model predictions and MOSs on LIVE, KonIQ-10k and SPAQ, respectively. Bold indicates the top two results

other BIQA models. We observe DUBMA achieves significant improvements on KADID-10k [Lin *et al.*, 2019], demonstrating the effectiveness of our learning framework. Unlike CaHDC [Wu *et al.*, 2020] and EONSS [Wang *et al.*, 2019], which use equally weighted pseudo-MOSs from FR-IQA models, DUBMA outperforms these methods by assigning varying reliability levels to each annotator, reducing the impact of noisy labels. Additionally, DUBMA, trained on a larger dataset with more accurate FR-IQA annotators, shows better generalization on LIVE and KADID-10k compared to Ma19 [Ma *et al.*, 2019], which uses a smaller model. Remarkably, our method surpasses PieAPP [Prashnani *et al.*, 2018] and LPIPS [Zhang *et al.*, 2018] without relying on human data and matches the performance of DISTS [Ding *et al.*, 2022] on KADID-10k, proving the effectiveness of learning from pseudo-labels.

**gMAD Results.** We conduct a gMAD competition game between DUBMA and MDFS [Ni *et al.*, 2024], the second-best-performing BIQA model on the synthetic dataset. The competition takes place on the Wang and Ma [Wang and Ma, 2021] dataset, which includes 100,000 images generated from 1,000 different scenes. Fig 3 illustrates the gMAD pairs between DUBMA and MDFS. In (a) and (b), where DUBMA and MDFS act as the defender and attacker, respectively, the top images have similar quality compared to the corresponding bottom images, indicating that DUBMA successfully resists the attack from MDFS. When the roles are reversed (see (c) and (d)), DUBMA consistently identifies the failure cases of MDFS, with pairs of images showing substantial differences in quality according to human perception. These findings suggest that DUBMA generalizes better than existing opinion-unaware BIQA models on synthetic distortions.

## 4.2 DUMBA for Realistic Distortions

| Method | KonIQ-10k | | SPAQ | |
|---|---|---|---|---|
| | SRCC | PLCC | SRCC | PLCC |
| Source only | 0.353 | 0.335 | 0.611 | 0.634 |
| CORAL | 0.350 | 0.361 | 0.617 | 0.632 |
| DANN | 0.642 | 0.642 | 0.771 | 0.773 |
| MMD | 0.657 | 0.691 | 0.794 | 0.797 |
| WDGRL | **0.701** | 0.721 | 0.810 | 0.815 |
| DeepJDOT | 0.680 | **0.725** | **0.817** | **0.821** |
| SWD | 0.684 | 0.693 | 0.814 | 0.820 |
| Ours | **0.703** | **0.740** | **0.834** | **0.841** |

Table 3: Correlations between model predictions and MOSs trained with different UDA methods, and bold indicates the top two results

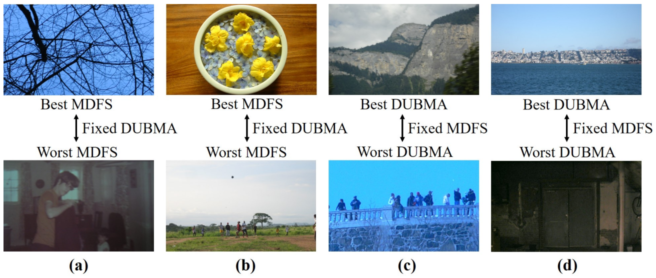

Figure 4: Representative gMAD pairs between DUBMA and MDFS [Ni *et al.*, 2024] on authentic distortions. **(a)** Fixing DUBMA at the low-quality level. **(b)** Fixing DUBMA at the high-quality level. **(c)** Fixing MDFS at the low-quality level. **(d)** Fixing MDFS at the high-quality level.

**Datasets.** We also leverage the Waterloo Exploration Database [Ma *et al.*, 2016], and simulate five distortion types [2] at five levels to construct $\mathcal{D}_s$. The pseudo labels are annotated by the aforementioned ten FR-IQA models. We generate about 25k image pairs for training. We use KonIQ-10k [Hosu *et al.*, 2020] and SPAQ [Fang *et al.*, 2020] as the target datasets. Both of them include more than $10,000$ authentically distorted images.

**Details of Training and Testing.** We use the same network architecture as the experiment on synthetic distortions. We train our model on the synthetic dataset $\mathcal{D}_s$ and adapt to the realistic dataset $\mathcal{D}_t$ by minimizing the loss function of Eq. (12). We rely on the POT package [Flamary *et al.*, 2017] to compute the exact OT solver and the Geomloss package for the unbalanced Sinkhorn divergence. The entropy of the probability output is employed as the criterion for validation. We also adopt the SRCC and PLCC for accuracy evaluation and gMAD competition to assess the generalization capacity.

**Quantitative Results.** We first compare the performance of the proposed method against the aforementioned sixteen BIQA models (see Section 4.1). Table 2 lists the SRCC and PLCC results between model predictions and MOSs on KonIQ-10k [Hosu *et al.*, 2020] and SPAQ [Fang *et al.*, 2020], respectively. We also include the performance on LIVE as a reference to measure the performance on synthetic distortions. It is easily observed that the BIQA methods initially tailored for synthetic distortions work well on LIVE but perform poorly on two realistic datasets with diverse content and complex distortions, especially the catastrophic performance

[2] These include Gaussian blur, Gaussian noise, motion blur, JPEG compression, and JPEG2000 compression.

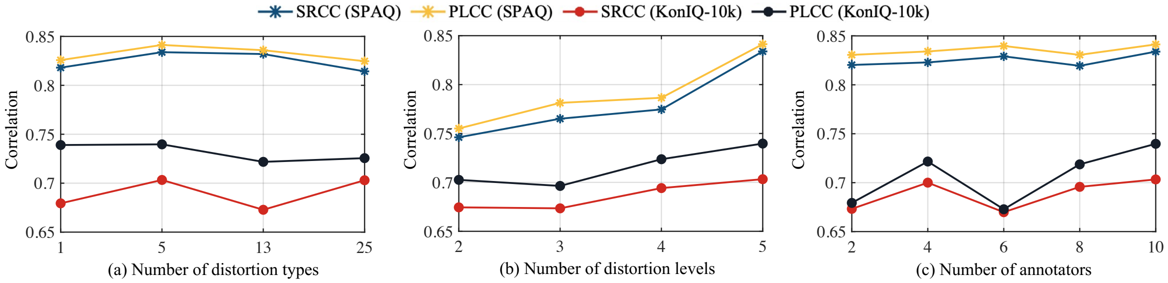

Figure 5: Ablation analysis of the number of (a) distortion types, (b) distortion levels, and (c) annotators.

of QAC. One possible reason is that the NSS is inadequate to model realistic distortions, characterized by complex image content and distortion types. Our proposed DUBMA achieves competitive performance on these three datasets, which confirms the effectiveness of our proposed method.

We then compare unbalanced OT against the *six* frequently used SOTA UDA methods - CORAL [Sun *et al.*, 2016], DANN [Ganin *et al.*, 2016], MMD [Borgwardt *et al.*, 2006], WDGRL [Shen *et al.*, 2018], DeepJDOT [Damodaran *et al.*, 2018] and SWD [Lee *et al.*, 2019]. CORAL, DANN, MMD, and WDGRL perform domain alignment on the feature space, while DeepJDOT and SWD minimize the discrepancy of joint deep feature and label domains. Table 3 reports the comparison results, where source only means that the BIQA model is only trained on $\mathcal{D}_s$ directly. We find UDA methods bring obvious performance gains compared to source only, except in the case of CORAL on SPAQ. Besides, joint alignment of both feature and label space is slightly superior to only matching feature space. Moreover, the proposed method outperforms all competing models by a significant margin.

**gMAD Results.** We qualitatively evaluate DUBMA and second-best MDFS [Ni *et al.*, 2024] using the gMAD competition. The competition performs the Wang *et al.* [Wang *et al.*, 2021] dataset, which includes 100,000 authentic images collected from the Internet. Figure 4 shows representative gMAD pairs between the two models. Initially, we allow MDFS to identify the failure cases of DUBMA, and we find that MDFS fails to detect these cases. In contrast, when we reverse the roles, DUBMA successfully identifies the failure cases of MDFS (see (c) and (d)). These results indicate that DUBMA outperforms MDFS for in-the-wild BIQA tasks.

**Ablation Studies.** We first probe the effect of the number of distortion types, and Figure 5 (a) shows the results. It could be observed that even synthesizing only one type of distortion to construct $\mathcal{D}_s$, the BIQA model can achieve impressive performance. With the increase of distortion types, the performance on KonIQ-10k and SPAQ is influenced slightly. We argue that UDA benefits BIQA models in learning distortion-unaware features. We then test the influence of distortion levels highly related to the image quality ranges. Figure 5 (b) shows the performance changes versus distortion levels. As we can see, the model performance is improved with more distortion levels, and five distortion levels achieve the relatively best performance. We further study the impact of the number of annotators. The results can be found in Figure 5 (c). It could be observed that the performance is improved with the increase of annotators, and it is best when the number of annotators is 10. Thus, we could leverage more reliable annotators to boost model performance. We finally verify the effectiveness of joint adaptation of both the marginal distribution and conditional distribution between domains [Courty *et al.*, 2017; Damodaran *et al.*, 2018]. Table 4 reports the comparison results obtained using the joint or non-joint cost function. We observe an obvious performance gain achieved by jointly aligning feature and conditional label space. This gain might indicate that only adapting the marginal distributions is not enough for the transfer learning of the proposed model.

| Method | KonIQ-10k | | SPAQ | |
|---|---|---|---|---|
| | SRCC | PLCC | SRCC | PLCC |
| Not Joint | 0.645 | 0.650 | 0.749 | 0.751 |
| **Joint** | 0.703 | 0.740 | 0.834 | 0.841 |

Table 4: Performance of the joint/non-joint alignment

## 5 Conclusion

We propose DUBMA in this paper that is trained on scalable pseudo-labeled synthetic pairs to handle both synthetic and real distortions. The approach uses supervised signals annotated by multiple existing FR-IQA models, eliminating the need for human ratings. To address the domain shift between the synthetic source data and the realistic target test data, we use robust joint UDA to align the two distributions. This allows us to train BIQA models on large-scale datasets that can be easily collected and scaled up without human input. Our extensive experiments demonstrate that our optimized model performs favorably against current opinion-unaware BIQA models. However, while the model's performance on authentic databases is promising, it has not yet achieved the performance level of models trained on human-rated datasets. We expect to improve the model performance by incorporating additional reliable IQA annotators and leveraging more efficient UDA solutions.

## Contributions

Zhihua Wang and Xuelin Liu are co-first authors.

## Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (62301323, 62301621, 62461028), in part by the Shenzhen Science and Technology Program (20231121172359002, JCYJ20241202125904007), in part by the Guangdong Basic and Applied Basic Research Foundation (2024A1515011164, 2025A1515011398), and in part by the China Postdoctoral Science Foundation (2024T170364).

## References

[Babu *et al.*, 2023] Nithin C Babu, Vignesh Kannan, and Rajiv Soundararajan. No reference opinion unaware quality assessment of authentically distorted images. In *WACV*, pages 2459–2468, 2023.

[Benamou, 2003] Jean-David Benamou. Numerical resolution of an "unbalanced" mass transport problem. *ESAIM: M2AN*, 37(5):851–868, 2003.

[Borgwardt *et al.*, 2006] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):149–157, 2006.

[Chapel *et al.*, 2021] Laetitia Chapel, Rémi Flamary, Haoran Wu, Cédric Févotte, and Gilles Gasso. Unbalanced optimal transport through non-negative penalized linear regression. In *NeurIPS*, pages 23270–23282, 2021.

[Chen *et al.*, 2021a] Baoliang Chen, Haoliang Li, Hongfei Fan, and Shiqi Wang. No-reference screen content image quality assessment with unsupervised domain adaptation. *TIP*, 30:5463–5476, 2021.

[Chen *et al.*, 2021b] Pengfei Chen, Leida Li, Jinjian Wu, Weisheng Dong, and Guangming Shi. Unsupervised curriculum domain adaptation for no-reference video quality assessment. In *ICCV*, pages 5178–5187, 2021.

[Courty *et al.*, 2017] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *NeurIPS*, 2017.

[Damodaran *et al.*, 2018] Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. DeepJDOT: Deep joint distribution optimal transport for unsupervised domain adaptation. In *ECCV*, pages 447–463, 2018.

[Ding *et al.*, 2021] Keyan Ding, Yi Liu, Xueyi Zou, Shiqi Wang, and Kede Ma. Locally adaptive structure and texture similarity for image quality assessment. In *ACM-MM*, pages 2483–2491, 2021.

[Ding *et al.*, 2022] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *TPAMI*, 44(5):2567–2581, 2022.

[Fang *et al.*, 2020] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *CVPR*, pages 3677–3686, 2020.

[Fatras *et al.*, 2021] Kilian Fatras, Thibault Séjourné, Rémi Flamary, and Nicolas Courty. Unbalanced minibatch optimal transport; applications to domain adaptation. In *ICML*, pages 3186–3197, 2021.

[Flamary *et al.*, 2017] R Flamary, N Courty, D Tuia, and A Rakotomamonjy. Optimal transport for domain adaptation. *TPAMI*, 39(9):1853–1865, 2017.

[Ganin *et al.*, 2016] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 17(1):2096–2030, 2016.

[Gu *et al.*, 2014] Ke Gu, Guangtao Zhai, Xiaokang Yang, and Wenjun Zhang. Hybrid no-reference quality metric for singly and multiply distorted images. *TBC*, 60(3):555–567, 2014.

[Hadizadeh *et al.*, 2017] Hadi Hadizadeh, Atiyeh Rajati, and Ivan V Bajić. Saliency-guided just noticeable distortion estimation using the normalized laplacian pyramid. *SPL*, 24(8):1218–1222, 2017.

[Hosu *et al.*, 2020] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *TIP*, 29:4041–4056, 2020.

[Larson and Chandler, 2010] Eric Cooper Larson and Damon Michael Chandler. Most apparent distortion: Full-reference image quality assessment and the role of strategy. *JEI*, 19(1):1–21, 2010.

[Lee and Seung, 2000] Daniel Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *NeurIPS*, pages 556–562, 2000.

[Lee *et al.*, 2019] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *CVPR*, pages 10285–10295, 2019.

[Lin *et al.*, 2019] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. KADID-10k: A large-scale artificially distorted IQA database. In *QoMEx*, pages 1–3, 2019.

[Liu *et al.*, 2017] Xialei Liu, Joost van de Weijer, and Andrew D Bagdanov. RankIQA: Learning from rankings for no-reference image quality assessment. In *ICCV*, pages 1040–1049, 2017.

[Liu *et al.*, 2019] Yutao Liu, Ke Gu, Yongbing Zhang, Xiu Li, Guangtao Zhai, Debin Zhao, and Wen Gao. Unsupervised blind image quality evaluation via statistical measurements of structure, naturalness, and perception. *TCSVT*, 30(4):929–943, 2019.

[Liu *et al.*, 2020] Yutao Liu, Ke Gu, Xiu Li, and Yongbing Zhang. Blind image quality assessment by natural scene statistics and perceptual characteristics. *TOMM*, 16(3):1–91, 2020.

[Ma *et al.*, 2016] 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. *TIP*, 26(2):1004–1016, 2016.

[Ma *et al.*, 2017] Kede Ma, Wentao Liu, Tongliang Liu, Zhou Wang, and Dacheng Tao. dipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs. *TIP*, 26(8):3951–3964, 2017.

[Ma *et al.*, 2019] Kede Ma, Xuelin Liu, Yuming Fang, and Eero P Simoncelli. Blind image quality assessment by learning from multiple annotators. In *ICIP*, pages 2344–2348, 2019.

[Mittal *et al.*, 2011] Anish Mittal, Gautam S Muralidhar, Joydeep Ghosh, and Alan C Bovik. Blind image quality assessment without human training using latent quality factors. *SPL*, 19(2):75–78, 2011.

[Mittal *et al.*, 2012] Anish Mittal, Ravi Soundararajan, and Alan C Bovik. Making a 'completely blind' image quality analyzer. *SPL*, 20(3):209–212, 2012.

[Nafchi *et al.*, 2016] Hossein Ziaei Nafchi, Atena Shahkolaei, Rachid Hedjam, and Mohamed Cheriet. Mean deviation similarity index: Efficient and reliable full-reference image quality evaluator. *IEEE Access*, 4:5579–5590, 2016.

[Ni *et al.*, 2024] Zhangkai Ni, Yue Liu, Keyan Ding, Wenhan Yang, Hanli Wang, and Shiqi Wang. Opinion-unaware blind image quality assessment using multi-scale deep feature statistics. *TMM*, 2024.

[Prashnani *et al.*, 2018] Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. PieAPP: Perceptual image-error assessment through pairwise preference. In *CVPR*, pages 1808–1817, 2018.

[Sheikh and Bovik, 2006] Hamid R Sheikh and Alan C Bovik. Image information and visual quality. *TIP*, 15(2):430–444, 2006.

[Sheikh *et al.*, 2006] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *TIP*, 15(11):3440–3451, 2006.

[Shen *et al.*, 2018] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *AAAI*, 2018.

[Sun *et al.*, 2016] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *AAAI*, 2016.

[Venkatanath *et al.*, 2015] Narasimhan Venkatanath, D Praneeth, Maruthi Chandrasekhar Bh, Sumohana S Channappayya, and Swarup S Medasani. Blind image quality evaluation using perception based features. In *NCC*, pages 1–6, 2015.

[Wang and Ma, 2021] Zhihua Wang and Kede Ma. Active fine-tuning from gMAD examples improves blind image quality assessment. *TPAMI*, 44(9):4577 – 4590, 2021.

[Wang *et al.*, 2003] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *ACSSC*, pages 1398–1402, 2003.

[Wang *et al.*, 2004] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *TIP*, 13(4):600–612, 2004.

[Wang *et al.*, 2019] Zhongling Wang, Shahrukh Athar, and Zhou Wang. Blind quality assessment of multiply distorted images using deep neural networks. In *ICIAR*, pages 89–101, 2019.

[Wang *et al.*, 2021] Zhihua Wang, Haotao Wang, Tianlong Chen, Zhangyang Wang, and Kede Ma. Troubleshooting blind image quality models in the wild. In *CVPR*, pages 16256–16265, 2021.

[Wu *et al.*, 2015] Qingbo Wu, Zhou Wang, and Hongliang Li. A highly efficient method for blind image quality assessment. In *ICIP*, pages 339–343, 2015.

[Wu *et al.*, 2020] 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. *TIP*, 29:7414–7426, 2020.

[Xue *et al.*, 2013a] Wufeng Xue, Lei Zhang, and Xuanqin Mou. Learning without human scores for blind image quality assessment. In *CVPR*, pages 995–1002, 2013.

[Xue *et al.*, 2013b] Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan C Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *TIP*, 23(2):684–695, 2013.

[Ye *et al.*, 2014] Peng Ye, Jayant Kumar, and David Doermann. Beyond human opinion scores: Blind image quality assessment based on synthetic scores. In *CVPR*, pages 4241–4248, 2014.

[Zhang and Li, 2012] Lin Zhang and Hongyu Li. SR-SIM: A fast and high performance IQA index based on spectral residual. In *ICIP*, pages 1473–1476, 2012.

[Zhang *et al.*, 2011] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. FSIM: A feature similarity index for image quality assessment. *TIP*, 20(8):2378–2386, 2011.

[Zhang *et al.*, 2014] Lin Zhang, Ying Shen, and Hongyu Li. VSI: A visual saliency-induced index for perceptual image quality assessment. *TIP*, 23(10):4270–4281, 2014.

[Zhang *et al.*, 2015] Lin Zhang, Lei Zhang, and Alan C Bovik. A feature-enriched completely blind image quality evaluator. *TIP*, 24(8):2579–2591, 2015.

[Zhang *et al.*, 2018] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018.

[Zhang *et al.*, 2021] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. Uncertainty-aware blind image quality assessment in the laboratory and wild. *TIP*, 30:3474–3486, 2021.

[Zhu *et al.*, 2021] Yunan Zhu, Haichuan Ma, Jialun Peng, Dong Liu, and Zhiwei Xiong. Recycling discriminator: Towards opinion-unaware image quality assessment using Wasserstein GAN. In *ACM-MM*, pages 116–125, 2021.