Source-free Unsupervised Domain Adaptation for Blind Image Quality Assessment

Jianzhao Liu, Xin Li, Shukun An and Zhibo Chen, Senior Member, IEEE

Abstract—Existing learning-based methods for blind image quality assessment (BIQA) are heavily dependent on large amounts of annotated training data, and usually suffer from a severe performance degradation when encountering the domain/distribution shift problem. Thanks to the development of unsupervised domain adaptation (UDA), some works attempt to transfer the knowledge from a label-sufficient source domain to a label-free target domain under domain shift with UDA. However, it requires the coexistence of source and target data, which might be impractical for source data due to the privacy or storage issues. In this paper, we take the first step towards the source-free unsupervised domain adaptation (SFUDA) in a simple yet efficient manner for BIQA to tackle the domain shift without access to the source data. Specifically, we cast the quality assessment task as a rating distribution prediction problem. Based on the intrinsic properties of BIQA, we present a group of well-designed self-supervised objectives to guide the adaptation of the BN affine parameters towards the target domain. Among them, minimizing the prediction entropy and maximizing the batch prediction diversity aim to encourage more confident results while avoiding the trivial solution. Besides, based on the observation that the IQA rating distribution of single image follows the Gaussian distribution, we apply Gaussian regularization to the predicted rating distribution to make it more consistent with the nature of human scoring. Extensive experimental results under cross-domain scenarios demonstrated the effectiveness of our proposed method to mitigate the domain shift.

I. INTRODUCTION

Blind image quality assessment (BIQA) aims to automatically predict the perceptual image quality without any information of reference images. Recent years have witnessed the impressive success of deep learning in BIQA area. However, learning-based methods rely heavily on abundant annotated data. For BIQA whose annotation process is laborious and time-consuming, it’s hard to obtain large amount of data, leading to the poor generalization under the cross-domain/cross-dataset scenario. To alleviate the dependence of annotation, some works [1], [2] utilize the unsupervised domain adaptation (UDA) technique to transfer the learned knowledge from label-sufficient source domain to label-free target domain in IQA area. UCDA [1] adopts two-stage adaptation method, where in the first stage a coarse adaptation is applied between the source and target domain, and in the second stage a fine-level adaptation is applied between the confident and uncertain subdomains of target domain. Inspired by the transferability of pairwise quality relationship, RankDA [2] adopts maximum mean discrepancy (MMD) loss to align the ranking features extracted from source and target domains. However, existing UDA methods for BIQA are required to access the source data when learning to adapt the model, which is impractical when the source data is not available during training (e.g., those from surveillance cameras or hospitals). In this paper, we address an interesting but challenging UDA scenario in IQA where only a trained source model and the unlabeled target data are provided.

A natural idea is to directly adopt the source-free unsupervised domain adaptation (SFUDA) methods invented in high-level classification tasks. For example, SHOT [3] fine-tunes the feature encoder by information maximization loss and self-supervised pseudo-labeling. Unfortunately, without taking into account the essential difference between classification and IQA, these methods will lead to inferior performance when applied directly to the IQA task. Conventional IQA methods only regress the scalar image quality score, while in this paper we cast the quality assessment task as a rating distribution prediction problem [1], [4], [5] with a novel probabilistic modeling method, in order to bridge the gap between the classification and IQA. In spite of this, there still exist large differences that hamper the performance of applying high-level SFUDA methods to IQA.

As shown in Fig. 1, the label of the classification task and the IQA task are quite different. For the classification task, the label is one-hot encoding with the probability of certain
class being 100%. However, for the IQA task, one distorted image will receive divergent opinions from different subjects and people rarely reach an absolute consensus, thus resulting in the uncertainty of the label. In order to further explore the characteristics of the rating distributions of IQA, we collect images with each subject’s annotation from KonIQ-10K [6]. KonIQ-10K adopts the single stimulus absolute category rating method to collect the discrete rating ranging from 1 to 5. The mean opinion score (MOS) is obtained by averaging all subjects’ ratings. As shown in Fig. 2, we first split the MOS value into 7 segments and cluster the rating distributions of images whose mean quality score lies in a specific range. Following [7], we use k-means to generate 5 clusters for each quality range, and visualize the rating distributions of these clusters. From Fig. 2 we can see that the clusters of rating distributions tend to be approximated by Gaussian. Besides, we employ three distributions: Gaussian, Gamma and Weibull to conduct the Goodness-of-Fit (GoF) experiments respectively, and the RMSE measuring results are shown in Table I which reveals that Gaussian functions can perform adequately for simulating the rating distributions of IQA.

![Fig. 2: Clusters of distributions for images with different MOS ranges. The legend of each plot shows the percentage of these images associated with each cluster. The score distributions trend to be Gaussian.](image)

**Table I: Goodness-of-Fit (GoF) measured by RMSE between distributions we used to model the score distributions of KonIQ-10K.** The percentage in the bracket denotes the percentage of these images whose MOSs are within the corresponding range.

| RMSE Model | MOS range | 1.0-1.5 | 1.5-2.0 | 2.0-2.5 | 2.5-3.0 | 3.0-3.5 | 3.5-4.0 | 4.0-4.5 | Weighted Avg. |
|------------|-----------|---------|---------|---------|---------|---------|---------|---------|--------------|
| Gaussian   |           | 0.0911  | 0.0387  | 0.0336  | **0.0158** | **0.0160** | 0.0360  | 0.0446  | **0.0254**  |
| Gamma      |           | **0.0256** | 0.0456  | **0.0217** | 0.0362  | 0.0310  | 0.0852  | 0.0907  | 0.0494      |
| Weibull    |           | 0.0664  | 0.0597  | 0.0633  | 0.0325  | 0.0433  | **0.0249** | 0.0473  | 0.0381      |

Based on the above observations, we introduce three self-supervised training objectives respectively to make the predicted distributions more consistent with the human scoring nature:

- **Entropy loss.** We encourage more confident predictions by minimizing the entropy of the predicted rating distribution for each single image.
- **Diversity loss.** We encourage the diversity of the predictions between different images by maximizing the entropy of the averaged batch predictions. The diversity loss can effectively avoid the trivial solution of entropy loss where all the unlabeled data have the same predictions that tend to be one-hot encoding.
- **Gaussian regularization loss.** We constrain that the predicted distributions should follow the Gaussian distribution.

Apart from the self-supervised training objectives specially tailored for IQA, we also propose a more efficient and simple
way to fine-tune the trained source network. Instead of fine-
tuning the whole network or the whole feature extractor
like most existing SFUDA methods [3], [9], we propose to
re-calculate the BN statistics and only optimize the affine
parameters of batch normalization (BN) layers towards target
domain. Inspired by some prior works aligning the BN statis-
tics between the source domain and the target domain [9], [10],
we think that the affine parameters of BN can well calibrate the
“source-style” features to “target-style” features. Specially, in
order to avoid the conflicts between multiple target domains,
we employ domain specific batch normalization (DSBN) [11]
to separate domain-specific information, thus making it pos-
tible to process multiple target domains simultaneously. Be-
sides, thanks to the DSBN, the proposed method can also be
extended to an unsupervised continual learning scenario where
each target domain comes sequentially.

In summary, our contributions can be summarized as fol-
lows:

- We take the first step towards the source-free unsuper-
  vised domain adaptation in the IQA area, which only
  utilizes a trained source model and the unlabeled target
domain data to solve UDA problems.
- We design three self-supervised training objectives ac-
  cording to three observed IQA intrinsic properties, and
demonstrate that simply fine-tuning the affine parameters
  of the BN layers can achieve efficient and fast adaptation
to the target domain.
- We use domain-specific batch normalization layer to
  avoid the conflicts between different domains, which
  makes it possible to process multiple target domains
  simultaneously. Besides, the proposed method can also be
  extended to unsupervised continual learning scenario.

II. RELATED WORKS

A. Blind image quality assessment

BIQA aims to automatically predict the subjective quality of
a distorted image without accessing any reference information.
It can be roughly divided into two categories: distortion-
specific and general-purpose approaches [12]. Distortion-
specific methods are designed for a particular distortion type
(e.g. JPEG [13], blur [14] and super-resolution [15]). General-
purpose methods design models based on the assumption that
distortions will break the regular statistics extracted from natu-
ral images [16], [17]. Recently, deep learning has significantly
advanced the development of BIQA. Typically, learning-based
BIQA methods [18], [19] usually employ the two-step net-
work, i.e., feature extraction and quality prediction. Pan et al.
[20] employed a fully convolutional neural network to
predict a pixel-by-pixel similar quality map and used a deep
pooling network to regress the quality map into a quality score.
In [21], Zhang et al. proposed a deep bilinear model that works
for both synthetically and authenticely distorted images.

It is known that learning-based BIQA models are vulnerable
to cross-distortion scenario generalization due to the limited
number of images with subjective ratings, and therefore some
works are proposed to alleviate the poor generalizability of
BIQA models. RankIQA [22] ranked images in terms of
image quality by using synthetically generated distortions,
to address the problem of limited IQA dataset size. Zhang
et al. [23] proposed an uncertainty-aware BIQA model and
a method of training it on multiple datasets simultaneously.
Liu et al. [24] and Zhang et al. [25] attempted to solve the
continual learning of BIQA. UCDA [1] and RankDA [2] employed the unsupervised domain adaptation technique to
transfer knowledge from the label-sufficient source domain to
the label-free target domain. Although the above UDA-based
methods have achieved impressive performance, they are all
designed to utilize the source data when adapting the model,
which is impractical when the source data is unavailable
due to storage or private issues. In contrast, our work tries
to investigate how to achieve a more practical and flexible
adaptation without access to the source data.

B. Unsupervised domain adaptation

Unsupervised domain adaptation (UDA) is proposed to
tackle the domain shift when the labeled source data and the
unlabeled target data fall from different distributions. The goal
of UDA is to align the distributions between the source and the
target domains. The mainstream UDA methods can be catego-
ized into two directions: discrepancy-based approaches and
adversary-based approaches. Discrepancy-based approaches
[26], [27] mostly use distance measures such as maximum
mean discrepancy (MMD) [28], Wasserstein metric, correla-
tion alignment (CORAL) [29], Kullback-Leibler (KL) diver-
gence [30], and contrastive domain discrepancy (CDD) [31]
to measure the dissimilarity across domains. Adversary-based
approaches [32], [33] tend to obtain transferable and domain-
invariant features through an adversarial objective. However,
the source data is not always available due to storage or
privacy issues in many real-world applications, which makes
the above UDA approaches unable to use. To address this
issue, source-free UDA is proposed, which adapts to the target
domain only with target data and the model pre-trained on
source domain provided. SHOT [3] proposed an information
maximization loss and a clustering-based pseudo labeling
method to implicitly align representations from the target
domains to the source hypothesis. Ishii et al. [8] proposed to
minimize both the BN-statistics matching loss and the in-
formation maximization loss to fine-tune the feature extractor.
TENT [34] proposed to optimize the BN affine parameters
by simply minimizing the entropy of the predictions during
test-time adaptation. Although several works have attempted
to utilize UDA technique to solve the domain shift in IQA, the
source-free UDA in IQA area is still under-explored, which
is an interesting but challenging topic.

III. PROBLEM DEFINITION

For a vanilla UDA task, we are usually given \( n_s \) labeled
samples \( \{x_i^s, y_i^s\}_{i=1}^{n_s} \) from the source domain \( \mathcal{D}_s \)
where \( x_i^s \in \mathcal{X}_s, y_i^s \in \mathcal{Y}_s \), and \( n_t \) unlabeled samples \( \{x_i^t\}_{i=1}^{n_t} \) from the target
domain \( \mathcal{D}_t \), where \( x_i^t \in \mathcal{X}_t \). The goal of vanilla UDA is to
predict the target labels \( \{y_i^t\}_{i=1}^{n_t} \), where \( y_i^t \in \mathcal{Y}_t \). We denote the
source task and target task as \( \mathcal{X}_s \rightarrow \mathcal{Y}_s \) and \( \mathcal{X}_t \rightarrow \mathcal{Y}_t \),
respectively. While for the source-free UDA, we are expected
to learn a target function \( f_t : \mathcal{X}_t \rightarrow \mathcal{Y}_t \) and infer \( \{q^l_i\}_{i=1}^n \) with only a pre-trained source function \( f_s : \mathcal{X}_s \rightarrow \mathcal{Y}_s \) and the unlabeled target data \( \{x_t^i\}_{i=1}^n \) provided.

IV. APPROACH

In this section, we first illustrate how to cast the quality assessment task to a rating distribution prediction task. Then we introduce the source-free adaptation process of our proposed method.

A. Discrete rating distribution construction and learning

To bridge the gap between the classification and the IQA, we aim to cast the quality assessment task to a discrete rating distribution prediction task, which is helpful to transfer the technologies from classification to IQA. Given the mean quality score \( \mu \) and the variance \( \sigma^2 \) of a single image, we first model the rating distribution as a truncated Gaussian distribution \( \mathcal{N}(\mu, \sigma^2) \) as Eq. (1)

\[
p(l|\mu, \sigma^2) = \frac{\varphi(\mu, \sigma^2; l)}{\int_{r_1}^{r_2} \varphi(\mu, \sigma^2; t)dt}, r_1 \leq l \leq r_2
\]

where \( l \) is the variable of the quality score, \( r_1 \) and \( r_2 \) are the lower and upper bounds of the quality range. \( \varphi \) is the standard normal distribution where \( \varphi(\mu, \sigma^2; l) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(l-\mu)^2}{2\sigma^2}\right) \).

To make this distribution discrete, we sample \( C \) quality levels at equal intervals, where \( C \)-level is corresponding to \( C \) classes in classification. The \( k \)-th rating level \( l_k \) is:

\[
l_k = \frac{k-1}{C-1} * (r_2 - r_1) + r_1, \quad k \in [1, C]
\]

Then we construct the discrete rating distribution \( q = (q_1, \ldots, q_C) \) using:

\[
q_k = \frac{p(l_k|\mu, \sigma^2)}{\sum_j p(l_j|\mu, \sigma^2)}, \quad \sum_k q_k = 1.
\]

where \( \sum_{k=1}^{C} q_k = 1 \). \( L = (l_1, \ldots, l_C) \) is the rating level set and \( \sum_{k=1}^{C} q_k l_k \approx \mu \). A deep IQA network is usually composed of two parts: feature extractor and quality prediction head (several fully-connected layers). Given an input image \( x \) with ground truth mean quality score \( \mu \) and rating distribution \( q \), the activation of the last fully connected layer is \( z = \phi(x; \theta) = (z_1, \ldots, z_C) \), where \( \theta \) is the parameters of the IQA network \( \phi \). We use a softmax function to transform \( z \) into a probability distribution:

\[
\hat{q}_k = \frac{\exp(z_k)}{\sum_j \exp(z_j)}.
\]

The goal is to find \( \theta \) to generate a distribution \( \hat{q} \) that is similar to \( q \) with the predicted mean quality score \( \hat{\mu} \) keeps close to \( \mu \). During training on the source domain, we construct the following learning objective:

\[
\mathcal{L} = \mathcal{L}_{KL} + \mathcal{L}_{MSE} = \sum_{k=1}^{C} \mathbb{E}_{\{x^i\}_{i=1}^B | x, z_k = \hat{q}_k} \left[ -\sum_k q_k \log \hat{q}_k + ||\mu - \hat{\mu}||^2 \right],
\]

where \( \hat{\mu} = \sum_{k=1}^{C} \hat{q}_k l_k \) and \( B \) is the mini-batch size. \( \mathcal{L}_{KL} \) minimizes Kullback-Leibler (KL) divergence between the predicted distribution and the ground-truth distribution. In this way, we cast the quality assessment task as a rating distribution prediction task.

B. Adaptation to target domain

As shown in Fig. 3 during the adaptation process, we freeze the parameters of all the network except for the affine parameters within the domain-specific batch normalization (DSBN) layers. DSBN is just like the ordinary BN, but can work for multiple domains simultaneously. A BN layer
### Table II: Description of IQA databases. DisNum refers to the number of synthetic distortion types. MOS refers to Mean Opinion Score and a higher value denotes better perceptual quality. DMOS refers to Differential Mean Opinion Score and is inversely proportional to MOS. DisImageNum refers to the number of distorted images.

| Database  | Scenario | DisNum | Subjective Testing Methodology | Annotation | Range | DisImageNum |
|-----------|----------|--------|--------------------------------|------------|-------|-------------|
| LIVE [36] | Synthetic | 5      | Single stimulus continuous quality rating | DMOS, Variance | [0,100] | 779         |
| CSIQ [37] | Synthetic | 6      | Multi stimulus absolute category rating | DMOS, Variance | [0,1]  | 866         |
| KADID-10K [38] | Synthetic | 25     | Double stimulus absolute category rating with crowdsourcing | MOS, Variance | [1.5]  | 10,125      |
| BID [39]  | Authentic | -      | Single stimulus continuous quality rating | DMOS, Variance | [0.5]  | 386         |
| KonIQ-10K [6] | Authentic | -      | Single stimulus absolute category rating with crowdsourcing | MOS, Variance | [1.5]  | 10,073      |

![MOS=2.265](image1)

![MOS=3.7706](image2)

Fig. 4: Comparison between different combinations of objectives. Adaptation results using Entro., Entro.+Div. and Entro.+Div.+Gau. are shown from left to right, respectively. Our method can produce more accurate results consistent with the nature of human scoring (Gaussian distribution). To avoid overlap, we shift the x-coordinate a little for adapted results.

where during testing stage, $\tilde{\mu}_d$ and $\tilde{\sigma}_d^2$ are directly used for the testing examples in each domain $d$.

For the $\tau$-th target domain $t_\tau$, we update the BN statistics and optimize the affine-parameters {$\gamma_{t\tau,m,h}, \beta_{t\tau,m,h}$} for each BN layer $m$ and channel $h$ in the DSBN, where $\tau \in \{1, \ldots, T\}$. According to the three observed characteristics of IQA distributions, we design three self-supervised objectives: Entropy loss, Diversity loss and Gaussian regularization loss.

**Entropy loss.** Based on the observation that all the ratings are fluctuated around the mean quality scores, we minimize the entropy of the predicted rating distribution for each single image to encourage more confident predictions:

$$
L_{ent} = -E_{(x^i)\in X_{t\tau}} \sum_{k=1}^C \hat{q}_k^i \log \hat{q}_k^i
$$

where $\hat{q}_k^i = \delta(\phi(x^i; \theta))$ is the softmax output of the IQA network $\phi$ with parameters $\theta$, $\delta$ is the softmax function and $B$ is the mini-batch size.

**Diversity loss.** Based on the observation that the rating distributions of different images are diverse, we maximize the entropy of the averaged prediction within a mini-batch to make

$$
(\tilde{\sigma}_d^{ite+1})^2 = (1 - \alpha)(\tilde{\sigma}_d^{ite})^2 + \alpha(\tilde{\sigma}_d^{ite+1})^2,
$$

where $\alpha$ is a small constant to avoid divide-by-zero, and $\tilde{\mu}_d$ and $\tilde{\sigma}_d$ are statistics (the mean and variance) of activations within a mini-batch:

$$
\tilde{\mu}_d = \frac{\sum_b \sum_{i,j} v_d[b, i, j]}{B \cdot H \cdot W},
$$

$$
\tilde{\sigma}_d^2 = \frac{\sum_b \sum_{i,j} (v_d[b, i, j] - \tilde{\mu}_d)^2}{B \cdot H \cdot W}.
$$

During training, DSBN estimates the mean and variance of activations for each domain separately by exponential moving average with update factor $\alpha$:

$$
\tilde{\mu}_d^{ite+1} = (1 - \alpha)\tilde{\mu}_d^{ite} + \alpha \tilde{\mu}_d^{ite+1}.
$$

$$
\tilde{\sigma}_d^{ite+1} = (1 - \alpha)\tilde{\sigma}_d^{ite} + \alpha \tilde{\sigma}_d^{ite+1}.
$$
the results globally diverse:

$$L_{\text{div}} = - \sum_{k=1}^{C} \hat{q}_k \log q_k, \quad (13)$$

where $\hat{q} = E_{(x^i)_{i=1}^n \in \mathcal{X}_s} [\delta(\phi(x^i; \theta))]$ is the averaged prediction within a mini-batch.

**Gaussian regularization loss.** Based on the observation that the rating distributions can be closely approximated by Gaussian functions, we design the Gaussian regularization loss to encourage the predicted predictions to be more consistent with the nature of the human scoring:

$$L_{\text{Gau}} = -E_{(x^i)_{i=1}^n \in \mathcal{X}_s} \sum_{k=1}^{C} \hat{q}_k \log \hat{q}_k, \quad (14)$$

where $\hat{q}$ is the pseudo rating distribution:

$$\hat{q}_k = \frac{p(l_k | \hat{\mu}, \hat{\sigma}^2)}{\sum_j p(l_j | \hat{\mu}, \hat{\sigma}^2)}. \quad (15)$$

$p$ is the truncated normal distribution defined in Eq. [1] $\hat{\mu}$ and $\hat{\sigma}^2$ are the estimated mean and variance of the predicted rating distribution respectively. $\hat{\mu} = \sum q_k l_k$ and $\hat{\sigma}^2 = \sum q_k (l_k - \hat{\mu})^2$. The total loss on $T$ target domains can be defined by:

$$L_{\text{total}} = \frac{1}{T} \sum_{\tau} (L_{\text{ent}} - \lambda_{\text{div}} L_{\text{div}} + \lambda_{\text{Gau}} L_{\text{Gau}}), \quad (16)$$

where $\lambda_{\text{div}}$ and $\lambda_{\text{Gau}}$ are hyperparameters of $\tau$-th target domain which balance the weights of corresponding loss items.

### V. Experiments

In this section, we first introduce the IQA datasets and the implementation details of our experiments. Then, we compare our method with several state-of-the-art methods and conduct ablation studies to verify the effectiveness of each item of our designed self-supervised loss. Finally, we show the performance of our method under the unsupervised continual learning setting.

#### A. Datasets

We conduct cross-domain experiments on five IQA datasets, among which three are synthetically distorted (LIVE [36], CSIQ [37], KADID-10K [38]) and the others are authentically distorted (BID [39] and KonIQ-10K [38]). The summarization of the five datasets is shown in Table [1].

**LIVE** database includes 779 synthetically distorted images, which are generated from 29 reference images by corrupting them with five distortion types. DMOS of each distorted image ranges from 0 to 100 and is collected using the single stimulus continuous quality rating method. **CSIQ** database consists of 866 synthetically distorted images which are derived from 30 original images distorted with six distortion types at four to five different intensity levels. The ratings are reported in the form of DMOS ranging from 0 to 1 using multi-stimulus absolute category rating method. **KADID-10K** consists of 10,125 distorted images derived from 81 pristine images considering 25 different distortion types at 5 intensity levels. The MOS of each image ranges from 1 to 5 and is collected using double stimulus absolute category rating with crowdsourcing. **BID** database contains 586 authentically distorted pictures taken by human users in a variety of scenes, camera apertures, and exposition times. The distorted images are mostly blurred, which include not only typical, easy-to-model blurring cases but also more complex, realistic ones. The MOS of each image ranges from 0 to 5 and is collected using the single stimulus continuous rating method. **KonIQ-10K** database consists of 10,073 authentically distorted images selected from a massive public multimedia database, YFCC100m [40]. The MOS of each image ranges from 1 to 5 and is collected by the single stimulus absolute category rating method with crowdsourcing.

#### B. Implementation details

To evaluate the performance of our proposed method under cross-domain scenario, we utilize the above 5 datasets to construct $5 \times 4$ cross-domain settings. For each dataset, we randomly sampled 80% images for training, 10% for validation and the left 10% for testing. Specially, for the synthetic datasets, we split the datasets according to the reference images in order to ensure content dependence. During training we randomly cropped the images into 384 × 384, and during validation and testing, we cropped the images to 384 × 384 in the center. We set the quality range lower bound $r_1$ and the upper bound $r_2$ to 1 and 5, respectively. Then we linearly rescaled the MOS/DMOS of each image to the range of [1-5] for all datasets, and the variance was also correspondingly scaled. We empirically set the number of sampled rating levels $C$ to 5.

We employ a pre-trained ResNet-18 (without the final FC layers) as the feature extractor and use two $FC – \text{ReLU}$ layers followed by a softmax function as the quality prediction head. We empirically set $\lambda_{\text{div}}$ to 1.0 and $\lambda_{\text{Gau}}$ to 0.2 for all target domains in all experiments. When training on the source domain, the Adam optimizer [46] with a leaning rate of $1 \times 10^{-4}$ was adopted to minimize the loss defined in Eq. [5]. We adopted an early-stopping strategy and chose the model that performed the best on the validation set as the pre-trained source model. During the adaptation stage, we employed the Adam optimizer with a leaning rate of $5 \times 10^{-5}$ to minimize the loss defined in Eq. [16]. We choose the adapted model which achieves the lowest loss for testing. We adopted two performance criteria: Spearman rank-order correlation coefficient (SROCC) and Pearson linear correlation coefficient (PLCC), to measure prediction monotonicity and precision, respectively. Before computing PLCC, the predicted quality scores are passed through a non-linear logistic mapping function:

$$\hat{\mu} = \beta_1 \left( \frac{1}{2} - \frac{1}{\exp(\beta_2 (\hat{\mu} - \beta_3))} \right) + \beta_4 \hat{\mu} + \beta_5 \quad (17)$$

We repeated 5 times for all experiments with different seeds and got the average values of PLCC and SROCC for report.
TABLE III: SROCC and PLCC performance comparisons of different algorithms under various cross-domain scenarios.

| Source | SROCC/PLCC Target | KADID-10K | KonIQ-10K | BID | CSIQ | LIVE |
|--------|-------------------|-----------|-----------|-----|------|------|
| Alg.   |                   | KADID-10K | KonIQ-10K | BID | CSIQ | LIVE |
|        |                   | 0.3464/0.3792 | 0.4469/0.4600 | 0.3640/0.4046 | 0.4895/0.5064 | 0.8830/0.8876 |
| Trad.  |                  | 0.2424/0.3325 | 0.0843/0.1995 | 0.2376/0.2437 | 0.5512/0.6721 | 0.7787/0.7881 |
|        |                   | 0.2890/0.3704 | 0.3910/0.3815 | 0.2494/0.3341 | 0.5853/0.5912 | 0.9456/0.9399 |
| Lean.  |                  | 0.8603/0.8666 | 0.5032/0.4584 | 0.5231/0.5252 | 0.7408/0.6673 | 0.8685/0.8296 |
|        |                  | 0.8416/0.8429 | 0.4290/0.3855 | 0.3792/0.2937 | 0.7334/0.6069 | 0.8936/0.8304 |
|        |                  | 0.8400/0.8521 | 0.5876/0.6017 | 0.4817/0.4933 | 0.7455/0.7748 | 0.8262/0.8489 |
|        |                  | 0.7406/0.7335 | 0.3517/0.2884 | 0.2290/0.2521 | 0.6290/0.6109 | 0.8063/0.8082 |
|        |                  | 0.8156/0.8246 | 0.5270/0.6362 | 0.2121/0.2017 | 0.5515/0.7077 | 0.8667/0.9071 |
| NoAdapt|                  | 0.6371/0.6233 | 0.5378/0.5084 | 0.7452/0.7666 | 0.8783/0.8847 |
| UDA    |                  | 0.4958/0.5010 | 0.5499/0.5277 | 0.6794/0.7189 | 0.8960/0.8919 |
|        |                  | 0.6364/0.5662 | 0.5273/0.578 | 0.7561/0.6402 | 0.7152/0.7024 |
|        |                  | 0.6735/0.6004 | 0.6186/0.6200 | 0.7482/0.7752 | 0.8515/0.8556 |
|        |                  | 0.6839/0.6937 | 0.5622/0.5798 | 0.7588/0.7312 | 0.8553/0.8634 |
|        |                  | 0.7221/0.7124 | 0.5946/0.5982 | 0.7658/0.7855 | 0.8942/0.8867 |
| KonIQ-10K|                 | 0.5518/0.5821 | 0.8654/0.8784 | 0.7511/0.7645 | 0.7105/0.7608 | 0.7266/0.7226 |
|        |                  | 0.5829/0.6024 |                         | 0.7574/0.7766 | 0.7699/0.8121 | 0.8577/0.8444 |
| BID    |                  | 0.3761/0.4371 | 0.5751/0.6192 | 0.8082/0.7805 | 0.4481/0.4355 | 0.6159/0.5762 |
|        |                  | 0.3960/0.4606 | 0.6552/0.6579 | 0.7437/0.4432 | 0.6259/0.5933 |
| CSIQ   |                  | 0.3214/0.3836 | 0.6182/0.6118 | 0.5958/0.5447 | 0.9608/0.9472 | 0.9107/0.9024 |
|        |                  | 0.5243/0.5973 | 0.6669/0.6744 | 0.5955/0.5996 | 0.9053/0.9035 |
| LIVE   |                  | 0.4768/0.5230 | 0.7210/0.7387 | 0.6718/0.6926 | 0.7223/0.8003 | 0.9731/0.9573 |
|        |                  | 0.5319/0.5387 | 0.7165/0.7146 | 0.6983/0.6799 | 0.7502/0.8025 |

C. Performance evaluation

In order to validate the effectiveness of our proposed method under cross-domain scenario, we utilize the 5 datasets described in Section V-A to construct 5 × 4 cross-domain settings, covering synthetic → synthetic, synthetic → authentic, authentic → synthetic and authentic → authentic. We compare the performance of ours against three traditional methods (NIQE [16], PIQE [41], and BRISQUE [42]), five learning-based methods (DBCNN [43], HyperIQA [44], RankIQA [22], MetaIQA [45] and NIMA [4]), two UDA-based methods for quality assessment (UCDA [1] and RankDA [2]) and two SUDUA methods adapted from the classification task (TENT [34] and SHOT [3]). One should be noted that among the learning-based methods, NIMA also predicts the rating distributions instead of the quality score, but it approximates the distributions from the available MOS values through maximum entropy optimization which is not consistent with human scoring nature (when the MOS value is equal to 3, the approximated distribution is a uniform distribution where the probability of each rating category is equal to 0.2). We evaluate the competitors’ performance under the same experimental settings as ours, and the results are shown in Table III: From the results, we have several observations:

1. All algorithms, no matter traditional methods or learning-based methods, will suffer from performance degradation when crossing domain shift.

2. Comparing the “NoAdapt” results of ours against other methods, we can find that due to the rating distribution construction and learning, our method can achieve comparatively better generalization ability under cross-dataset setup.

3. Compared with UDA-based IQA methods, ours can steadily achieve better performance. In the second stage of UCDA, it tries to align the confident subdomain and uncertain domain, which can be seen as implicit entropy minimization. In RankDA, it aligns the rank feature which reveals the pairwise quality relationship. However, it selects the rank pairs according to the inaccurate pseudo scores given by DBCNN [43] and the final quality scores are generated from inaccurate rank mos, thus it does not bring apparent improvement either.

4. Compared with SFUDA-based methods adapted from classification, ours can also obtain superior performance due to the objectives specially designed according to IQA intrinsic properties.

5. Our method can effectively mitigate the domain-shift under most cross-domain scenarios (e.g. the SROCC and PLCC are improved by 0.1311 and 0.1178 respectively under KonIQ-10K → LIVE scenario).

TABLE IV: SROCC and PLCC performance with respect to different combinations of the entropy loss, diversity loss and the Gaussian regularization loss. ↑ and ↓ mean performance improvement and degradation compared with NoAdapt results.

| SROCC/PLCC | KADID-10K | KonIQ-10K | BID | CSIQ | LIVE |
|------------|-----------|-----------|-----|------|------|
| NoAdapt    | 0.6378/0.6252 | 0.5400/0.5101 | 0.7452/0.7666 | 0.8783/0.8847 |
| Entro.     | 0.6812/0.6731 | 0.6186/0.6200 | 0.7482/0.7752 | 0.8515/0.8556 |
| Div.       | 0.6561/0.6593 | 0.5060/0.4915 | 0.7480/0.7770 | 0.8936/0.8816 |
| Gau.       | 0.6212/0.6161 | 0.5351/0.5499 | 0.7449/0.7625 | 0.9112/0.9027 |
| Entro.+Div. | 0.7058/0.6989 | 0.5989/0.6596 | 0.7550/0.7704 | 0.8431/0.8509 |
| Entro.+Gau. | 0.6748/0.6624 | 0.6212/0.6215 | 0.7614/0.7828 | 0.8590/0.8580 |
| Div.+ Gau.  | 0.6005/0.5897 | 0.5321/0.5346 | 0.7542/0.7820 | 0.9090/0.9082 |
| Ours       | 0.7221/0.7124 | 0.5946/0.5982 | 0.7658/0.7855 | 0.8942/0.8867 |
D. Ablation study

To verify the effectiveness of our designed self-supervised objectives, we conduct ablation studies under the “KADID-10K → Others” cross-domain scenario. We design 6 variants combinations of the entropy loss, diversity loss and the Gaussian regularization loss. The SROCC/PLCC performance are shown in Table [V]. From the table, we can see that each single objective can improve the performance of certain datasets. Besides, the combinations of the three objectives can obtain the best performance where the performances of all datasets get improved. To intuitively understand what each objective does, we visualize the predicted distributions as well as the corresponding quality scores of some samples in Fig. [4]. Adaptation results using Entro., Entro.+Div. and Entro.+Div.+Gau. are shown from left to right respectively. We can see that entropy loss is aimed to encourage more confident predictions by improving the probability of certain ranking category, but it will limit the predicted score around the initial quality score. At the same time, diversity loss will explore more diverse results and thus will probably change the peak of distribution. For example, in the second row of Fig. [4] the initial predicted quality score is 2.5265. When applying entropy loss, the distribution is sharper, but the predicted score gets worse. In contrast, when applying diversity loss, the peak of the predicted distribution shifts to the region near ranking 4. However, by compairing the circled segments, we can find the the results produced by Entro. and Entro.+Div. actually don’t obey the Gaussian distribution (the mean score does not corresponds to the peak). From the fourth column of Fig. [4] Kon we can find that when applying the gaussian regularization loss, the predicted distribution is well calibrated and tends to approximately be a Gaussian distribution. Meanwhile, the predicted score gets more accurate compared with simply using entropy loss and diversity loss.

E. Extension to unsupervised continual learning setting

Thanks to the DSBN layers, our method not only can process multiple target domains simultaneously, but also can be extended to process multiple domains sequentially which is called a continual learning setting. We sequentially learn multiple datasets following the order of LIVE→CSIQ→BID→CLIVE→KonIQ-10K→KADID-10K in the [24]. We employ separate parameters to each new task by DSBN, thus avoiding the catastrophic forgetting of previously learned knowledge when learning new task (similar ideas can be found in prior works [48], [49]). Instead of learning new tasks in a supervised manner, we adopt unsupervised learning using our proposed self-supervised objectives. We compare our method with other five methods which learn new task in a supervised manner, including fine-tuning (FT), EWC [50], LWF [51], GFR [52] and LIQA [24]. We show the SROCC performance of five previously learned datasets at the last task session in Table [V]. From the table, we can surprisingly find that although ours adopts unsupervised learning, it still shows superior performance because it does not forget the learned knowledge during the whole continual learning process.

VI. Conclusions

In this paper, we first tackle the challenge of source-free unsupervised domain adaptation in IQA area, which is an interesting but significant topic when the source data is inavilable due to special issues. Following the observation that the rating distribution can be approximated by Gaussian, we cast the quality assessment task to a rating distribution prediction task by rating distribution construction and learning. Then to achieve a simple yet effective adaptation towards the target domain, we simply optimize the affine parameters of BN layers, which can modulate the “source-style” feature to the “target-style” feature. Considering the IQA intrinsic properties, we design three self-supervised objectives to make the predictions more consistent with the nature of human scoring nature. Extensive experiments demonstrate the effectiveness of our proposed method under cross-domain scenarios. Specially, we employ DSBN to avoid conflicts between various domains. Therefore, our method can process multiple target domains simultaneously and sequentially (continual learning setting), which is convenient and scalable.

REFERENCES

[1] P. Chen, L. Li, J. Wu, W. Dong, and G. Shi, “Unsupervised curriculum domain adaptation for no-reference video quality assessment,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 5178–5187.

[2] B. Chen, H. Li, H. Fan, and S. Wang, “No-reference screen content image quality assessment with unsupervised domain adaptation,” IEEE Transactions on Image Processing, vol. 30, pp. 5463–5476, 2021.

[3] J. Liang, D. Hu, and J. Feng, “Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation,” in International Conference on Machine Learning. PMLR, 2020, pp. 6028–6039.

[4] H. Talebi and P. Milanfar, “Nima: Neural image assessment,” IEEE Transactions on Image processing, vol. 27, no. 8, pp. 3998–4011, 2018.

[5] H. Zeng, L. Zhang, and A. C. Bovik, “A probabilistic quality representation approach to deep blind image quality prediction,” arXiv preprint arXiv:1708.08190, 2017.

[6] V. Hsu, H. Lin, T. Sziranyi, and D. Sauppe, “Konig-10k: An ecologically valid database for deep learning of blind image quality assessment,” IEEE Transactions on Image Processing, vol. 29, pp. 4041–4056, 2020.

[7] N. Murray, L. Marchesotti, and F. Perronnin, “Ava: A large-scale database for aesthetic visual analysis,” in 2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012, pp. 2408–2415.

[8] M. Ishii and M. Sugiyama, “Source-free domain adaptation via distributional alignment by matching batch normalization statistics,” arXiv preprint arXiv:2101.10842, 2021.

[9] Y. Li, N. Wang, J. Shi, X. Hou, and J. Liu, “Adaptive batch normalization for practical domain adaptation,” Pattern Recognition, vol. 80, pp. 109–117, 2018.

[10] F. Maria Carlucci, L. Porzi, B. Caputo, E. Ricci, and S. Rota Bulo, “Autoidal: Automatic domain alignment layers,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 5067–5075.
