Incorporating Semi-Supervised and Positive-Unlabeled Learning for Boosting Full Reference Image Quality Assessment

Yue Cao[1], Zhaolin Wan[2], Dongwei Ren[1(✉)], Zifei Yan[1], Wangmeng Zuo[1,3]

[1]School of Computer Science and Technology, Harbin Institute of Technology, China
[2]College of Artificial Intelligence, Dalian Maritime University, China [3]Peng Cheng Laboratory, China
cscaoyue@gmail.com, zlwan@dlmu.edu.cn, rendongweihit@gmail.com, {yanzifei, wmzuo}@hit.edu.cn

Abstract

Full-reference (FR) image quality assessment (IQA) evaluates the visual quality of a distorted image by measuring its perceptual difference with pristine-quality reference, and has been widely used in low-level vision tasks. Pairwise labeled data with mean opinion score (MOS) are required in training FR-IQA model, but is time-consuming and cumbersome to collect. In contrast, unlabeled data can be easily collected from an image degradation or restoration process, making it encouraging to exploit unlabeled training data to boost FR-IQA performance. Moreover, due to the distribution inconsistency between labeled and unlabeled data, outliers may occur in unlabeled data, further increasing the training difficulty. In this paper, we suggest to incorporate semi-supervised and positive-unlabeled (PU) learning for exploiting unlabeled data while mitigating the adverse effect of outliers. Particularly, by treating all labeled data as positive samples, PU learning is leveraged to identify negative samples (i.e., outliers) from unlabeled data. Semi-supervised learning (SSL) is further deployed to exploit positive unlabeled data by dynamically generating pseudo-MOS. We adopt a dual-branch network including reference and distortion branches. Furthermore, spatial attention is introduced in the reference branch to concentrate more on the informative regions, and sliced Wasserstein distance is used for robust difference map computation to address the misalignment issues caused by images recovered by GAN models. Extensive experiments show that our method performs favorably against state-of-the-arts on the benchmark datasets PIPAL, KADID-10k, TID2013, LIVE and CSIQ. The source code and model are available at https://github.com/happycaoyue/JSPL.

1. Introduction

The goal of image quality assessment is to provide computational models that can automatically predict the perceptual image quality consistent with human subjective perception. Over the past few decades, significant progress has been made in developing full reference (FR) image quality assessment (IQA) metrics, including peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [58], which have been widely used in various image processing fields. Recently, CNN-based FR-IQA models have attracted more attention, which usually learn a mapping from distorted and pristine images to mean opinion score.

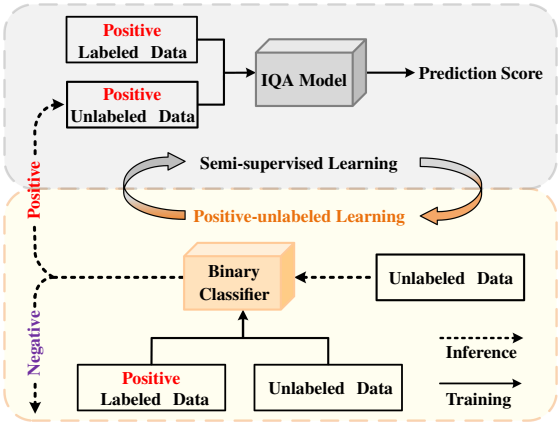


Figure 1. Illustration of joint semi-supervised and PU learning (JSPL) method, which mitigates the adverse effect of outliers in unlabeled data for boosting the performance of IQA model.

Most existing CNN-based FR-IQA models are trained using pairwise labeled data with mean opinion score (MOS), thus requiring extensive human judgements. To reduce the cost of collecting a large amount of labeled data, a potential alternative is semi-supervised learning for exploiting unlabeled samples which are almost free. Recently, considerable attention has been given to semi-supervised IQA algorithms [38,39,55,59,63] which show promising performance using both labeled and unlabeled data. However, unlabeled data can be collected in various unconstrained ways and may have a much different distribution from labeled

data. Consequently, outliers usually are inevitable and are harmful to semi-supervised learning [22].

In this paper, we incorporate semi-supervised and positive-unlabeled (PU) learning for exploiting unlabeled data while mitigating the adverse effect of outliers. PU learning aims at learning a binary classifier from a labeled set of positive samples as well as an unlabeled set of both positive and negative samples, and has been widely applied in image classification [8] and anomaly detection [68]. As for our task, the labeled images with MOS annotations can be naturally treated as positive samples. As shown in Fig. 1, PU learning is then exploited to find and exclude outliers, *i.e.*, negative samples, from the unlabeled set of images without MOS annotations. Then, semi-supervised learning (SSL) is deployed to leverage both labeled set and positive unlabeled images for training deep FR-IQA models. Moreover, the prediction by PU learning can also serve as the role of confidence estimation to gradually select valuable positive unlabeled images for SSL. Thus, our joint semi-supervised and PU learning (JSPL) method provides an effective and convenient way to incorporate both labeled and unlabeled sets for boosting FR-IQA performance.

Besides, we also present a new FR-IQA network for emphasizing informative regions and suppressing the effect of misalignment between distorted and pristine images. Like most existing methods, our FR-IQA network involves a Siamese (*i.e.*, dual-branch) feature extraction structure respectively for distorted and pristine images. The pristine and distortion features are then fed into the distance calculation module to generate the difference map, which is propagated to the score prediction network to obtain the prediction score. However, for GAN-based image restoration, the distorted image is usually spatially misaligned with the pristine image, making pixel-wise Euclidean distance unsuitable for characterizing the perceptual quality of distorted image [18, 19]. To mitigate this, Gu [18] introduced a pixel-wise warping operation, *i.e.*, space warping difference (SWD). In this work, we extend sliced Wasserstein distance to its local version (LocalSW) for making the difference map robust to small misalignment while maintaining its locality. Moreover, human visual system (HVS) usually pays more visual attention to the image regions containing more informative content [33, 44, 51, 60], and significant performance improvements have been achieved by considering the correlation with human visual fixation or visual region-of-interest detection [14, 32, 34]. Taking the properties of HVS into account, we leverage spatial attention modules on pristine feature for emphasizing more on informative regions, which are then used for reweighting distance map to generate the calibrated difference maps.

Extensive experiments are conducted to evaluate our JSPL method for FR-IQA. Based on the labeled training set, we collect unlabeled data by using several representative image degradation or restoration models. On the Perceptual Image Processing ALgorithms (PIPAL) dataset [19], the results show that both JSPL, LocalSW, and spatial attention contribute to performance gain of our method, which performs favorably against state-of-the-arts for assessing perceptual quality of GAN-based image restoration results. We further conduct experiments on four traditional IQA datasets, *i.e.*, LIVE [47], CSIQ [33], TID2013 [45] and KADID-10k [35], further showing the superiority of our JSPL method against state-of-the-arts.

To sum up, the main contribution of this work includes:

- A joint semi-supervised and PU learning (JSPL) method is presented to exploit images with and without MOS annotations for improving FR-IQA performance. In comparison to SSL, PU learning plays a crucial role in our JSPL by excluding outliers and gradually selecting positive unlabeled data for SSL.
- In FR-IQA network, spatial attention and local sliced Wasserstein distance are further deployed in computing difference map for emphasizing informative regions and suppressing the effect of misalignment between distorted and pristine image.
- Extensive experiments on five benchmark IQA datasets show that our JSPL model performs favorably against the state-of-the-art FR-IQA models.

2. Related Work

In this section, we present a brief review on learning-based FR-IQA, semi-supervised IQA, as well as IQA for GAN-based image restoration.

2.1. Learning-based FR-IQA Models

Depending on the accessibility to the pristine-quality reference, IQA methods can be classified into full reference (FR), reduced reference (RR) and no reference (NR) models. FR-IQA methods compare the distorted image against its pristine-quality reference, which can be further divided into two categories: traditional evaluation metrics and CNN-based models. The traditional metrics are based on a set of prior knowledge related to the properties of HVS. However, it is difficult to simulate the HVS with limited hand-crafted features because visual perception is a complicated process. In contrast, learning-based FR-IQA models use a variety of deep networks to extract features from training data without expert knowledge.

For deep FR-IQA, Gao *et al.* [15] first computed the local similarities of the feature maps from VGGNet layers between the reference and distorted images. Then, the local similarities are pooled together to get the final quality score. DeepQA [2] applied CNN to regress the sensitivity map to subjective score, which was generated from distorted images and error maps. Bosse *et al.* [6] presented a CNN-based FR-IQA method, where the perceptual

image quality is obtained by weighted pooling on patch-wise scores. Learned Perceptual Image Patch Similarity (LPIPS) [73] computed the Euclidean distance between reference and distorted deep feature representations, and can be flexibly embedded in various pre-trained CNNs, such as VGG [52] and AlexNet [30]. Benefiting from SSIM-like structure and texture similarity measures, Ding *et al*. [13] presented a Deep Image Structure and Texture Similarity metric (DISTS) based on an injective mapping function. Hammou *et al*. [23] proposed an ensemble of gradient boosting (EGB) metric based on selected feature similarity and ensemble learning. Ayyoubzadeh *et al*. [3] used Siamese-Difference neural network equipped with the spatial and channel-wise attention to predict the quality score. All the above metrics require a large number of labeled images to train the model. However, manual labeling is expensive and time-consuming, making it appealing to better leverage unlabeled images for boosting IQA performance.

2.2. Semi-Supervised IQA

In recent years, semi-supervised IQA algorithms have attracted considerable attention, as they use less expensive and easily accessible unlabeled data, and are beneficial to performance improvement [10]. Albeit semi-supervised learning (SSL) has been extensively studied and applied in vision and learning tasks, the research on semi-supervised IQA is still in its infancy. Tang *et al*. [55] employed deep belief network for IQA task, and the method was pre-trained with unlabeled data and then finetuned with labeled data. Wang *et al*. [59] utilized the semi-supervised ensemble learning for NR-IQA by combining labeled and unlabeled data, where unlabeled data is incorporated for maximizing ensemble diversity. Lu *et al*. [40] introduced semi-supervised local linear embedding (SS-LLE) to map the image features to the quality scores. Zhao *et al*. [75] proposed a SSL-based face IQA method, which exploits the unlabeled data in the target domain to finetune the network by predicting and updating labels. In the field of medical imaging, the amount of labeled data is limited, and the annotated labels are highly private. And SSL [38,39,63] provided an encouraging solution to address this problem by incorporating the unlabeled data with the labeled data to achieve better medical IQA performance. Nonetheless, the above studies assume that the labeled and unlabeled data are from the same distribution. However, the inevitable distribution inconsistency and outliers are harmful to SSL [22], but remain less investigated in semi-supervised IQA.

2.3. IQA for GAN-based Image Restoration

Generative adversarial networks (GAN) have been widely adopted in image restoration for improving visual performance of restoration results. However, these images usually suffer from texture-like artifacts aka GAN-based distortions that are seemingly fine-scale yet fake details. Moreover, GAN is prone to producing restoration results with spatial distortion and misalignment, which also poses new challenges to existing IQA methods. Recently, some intriguing studies have been proposed to improve the performance on IQA for GAN-based image restoration. SWDN [18] proposed a pixel-wise warping operation named space warping difference (SWD) to alleviate the spatial misalignment, by comparing the features within a small range around the corresponding position. Shi *et al*. [50] deployed the reference-oriented deformable convolution and a patch-level attention module in both reference and distortion branches for improving the IQA performance on GAN-based distortion. For modeling the GAN-generated texture-like noises, IQMA [21] adopted a multi-scale architecture to measure distortions, and evaluated images at a fine-grained texture level. IQT [9] combined CNN and transformer for IQA task, and achieved state-of-the-art performance. Although progress has been made in evaluating GAN-based distortion, existing methods are based on labeled data via supervised learning. In comparison, this work suggests a joint semi-supervised and PU learning method as well a new IQA network for leveraging unlabeled data and alleviating the spatial misalignment issue.

3. Proposed Method

3.1. Problem Setting

Denote by $\boldsymbol{x} = (\boldsymbol{I}_{Ref}, \boldsymbol{I}_{Dis})$ a two-tuple of pristine-quality reference image $\boldsymbol{I}_{Ref}$ and distorted image $\boldsymbol{I}_{Dis}$, and y the ground-truth MOS. Learning-based FR-IQA aims to find a mapping $f(\boldsymbol{x})$ parameterized by Θ^f to predict the quality score $\hat{y}$ for approximating y. Most existing FR-IQA methods are based on supervised learning where the collection of massive MOS annotations is very time-consuming and cumbersome. In this work, we consider a more encouraging and practically feasible SSL setting, *i.e*., training FR-IQA model using labeled data as well as unlabeled data with outliers. While SSL has been suggested to exploit unlabeled data for boosting IQA performance, we note that outliers usually are inevitable when unlabeled data are collected with diverse and unconstrained ways. For example, reference image quality of some unlabeled two-tuples may not meet the requirement. And the unlabeled data may also contain distortion types unseen in labeled data and non-necessary for IQA training.

Let $\mathbb{P} = \{\boldsymbol{x}_i, y_i\}_{i=1}^{N_p}$ denote the positive labeled data and $\mathbb{U} = \{\boldsymbol{x}_j\}_{j=1}^{N_u}$ denote unlabeled data. We present a joint semi-supervised and PU learning (JSPL) method for leveraging the unlabeled data with potential outliers. Besides the IQA model $f(\boldsymbol{x})$, our JSPL also learns a binary classifier $h(\boldsymbol{x}_j)$ parameterized by Θ^h for determining an unlabeled two-tuple is a negative (*i.e*., outlier) or a positive sample.

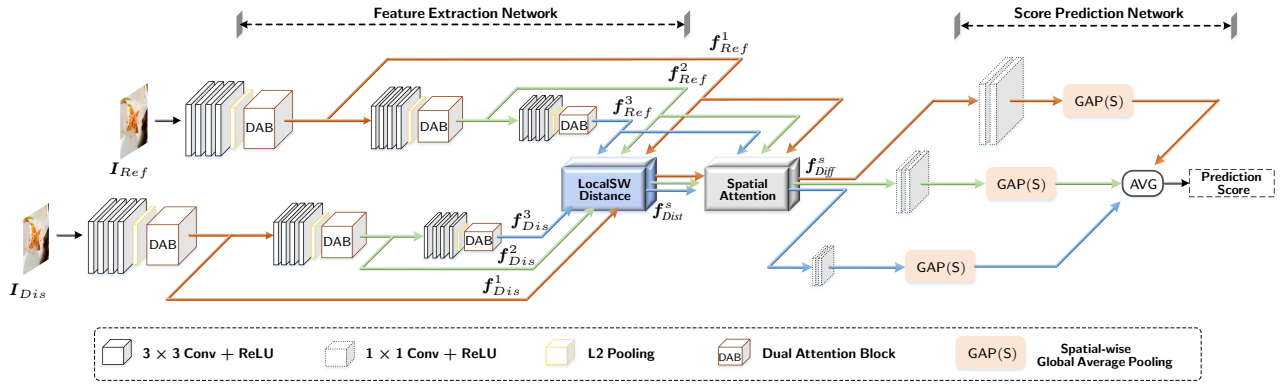


Figure 2. Illustration of our FR-IQA network. It adopts a dual-branch structure for feature extraction, *i.e.*, one for reference and another for distortion. The feature extraction network performs feature extraction on reference and distortion images at three scales. The distance calculation module generates the difference map between the above two features. The spatial attention module gives greater weight on more informative regions to obtain the calibrated difference map, which is then fed into score prediction network to predict the final score.

3.2. JSPL Model

A joint semi-supervised and PU learning (JSPL) model is presented to learn IQA model $f(\boldsymbol{x})$ and binary classifier $h(\boldsymbol{x})$ from the labeled data $\mathbb{P}$ and the unlabeled data $\mathbb{U}$. Particularly, PU learning is utilized to learn $h(\boldsymbol{x})$ for identifying positive unlabeled samples. And SSL is used to learn $f(\boldsymbol{x})$ from both labeled and positive unlabeled samples. In the following, we first describe the loss terms for PU learning and SSL, and then introduce our overall JSPL model.

PU Learning. In order to learn $h(\boldsymbol{x})$, we treat all samples in $\mathbb{P}$ as positive samples, and all samples in $\mathbb{U}$ as unlabeled samples. For a positive sample $\boldsymbol{x}_i$, we simply adopt the cross-entropy (CE) loss,

$$CE(h(\boldsymbol{x}_i)) = -\log h(\boldsymbol{x}_i). \tag{1}$$

Each unlabeled sample $\boldsymbol{x}_j$ should be either positive or negative sample, and we thus require the output $h(\boldsymbol{x}_j)$ to approach either 1 or 0. To this end, we introduce the entropy loss defined as,

$$\mathcal{H}(h(\boldsymbol{x}_j)) = -h(\boldsymbol{x}_j)\log h(\boldsymbol{x}_j) - (1-h(\boldsymbol{x}_j))\log(1-h(\boldsymbol{x}_j)). \tag{2}$$

We note that the entropy loss has been widely used in SSL [17]. When only using CE loss and entropy loss, $h(\boldsymbol{x})$ may simply produce 1 for any sample $\boldsymbol{x}$. To tackle this issue, for a given mini-batch $\mathcal{B}_u$ of unlabeled samples, we introduce a negative-enforcing (NE) loss for constraining that there is at least one negative sample in each mini-batch,

$$NE(\mathcal{B}_u) = -\log\left(1 - \min_{\boldsymbol{x}_j \in \mathcal{B}_u} h(\boldsymbol{x}_j)\right). \tag{3}$$

Combining the above loss terms, we define the PU learning loss as,

$$\mathcal{L}_{PU} = \sum\nolimits_i CE(h(\boldsymbol{x}_i)) + \sum\nolimits_j \mathcal{H}(h(\boldsymbol{x}_j)) + \sum\nolimits_{\mathcal{B}_u} NE(\mathcal{B}_u). \tag{4}$$

SSL. FR-IQA is a regression problem. For labeled sample $\boldsymbol{x}_i$ with ground-truth MOS y_i, we adopt the mean squared error (MSE) loss defined as,

$$\ell(f(\boldsymbol{x}_i), y_i) = \|f(\boldsymbol{x}_i) - y_i\|^2. \tag{5}$$

As for unlabeled data, only the positive unlabeled samples (*i.e.*, $h(\boldsymbol{x}_j) \geq \tau$) are considered in SSL. Here, τ (*e.g.*, $= 0.5$) is a threshold for selecting positive unlabeled samples. For positive unlabeled samples, we also adopt the MSE loss,

$$\ell(f(\boldsymbol{x}_j), y^*_j) = \|f(\boldsymbol{x}_j) - y^*_j\|^2, \tag{6}$$

where y^*_j denotes the pseudo MOS for $\boldsymbol{x}_j$. In SSL, sharpening is usually used for classification tasks to generate the pseudo label for unlabeled samples [4, 53], but is not suitable for regression tasks. Motivated by [31, 37], we use the moving average strategy to obtain y^*_j during training,

$$y^*_j(t) = \alpha \cdot y^*_j(t-1) + (1-\alpha) \cdot f^t(\boldsymbol{x}_j), \tag{7}$$

where α ($= 0.95$) is the momentum. $y^*_j(t)$ denotes the pseudo MOS after t iterations of training, and $f^t(\boldsymbol{x}_j)$ denotes the network output after t iterations of training. Therefore, we define the SSL loss as,

$$\mathcal{L}_{SSL} = \sum\nolimits_i \ell(f(\boldsymbol{x}_i), y_i) + \sum\nolimits_j \mathbb{I}_{h(\boldsymbol{x}_j)\geq\tau} \ell\left(f(\boldsymbol{x}_j), y^*_j\right). \tag{8}$$

$\mathbb{I}_{h(\boldsymbol{x}_j)\geq\tau}$ is an indicator function, where it is 1 if $h(\boldsymbol{x}_j) \geq \tau$ and 0 otherwise.

JSPL Model. Taking the losses for both SSL and PU learning into account, the learning objective for JSPL can be written as,

$$\min_{\Theta^f, \Theta^h} \mathcal{L} = \mathcal{L}_{SSL} + \mathcal{L}_{PU}. \tag{9}$$

We note that our JSPL is a joint learning model, where both the FR-IQA network $f(\boldsymbol{x})$ and binary classifier $h(\boldsymbol{x})$ can be learned by minimizing the above objective function. Particularly, for a given mini-batch of unlabeled samples, we first update the binary classifier by minimizing $\mathcal{L}_{PU}$. Then, pseudo MOS is updated for each unlabeled sample, and positive unlabeled samples are selected. Furthermore, the positive unlabeled samples are incorporated with the mini-batch of labeled samples to update the FR-IQA network by minimizing $\mathcal{L}_{SSL}$.

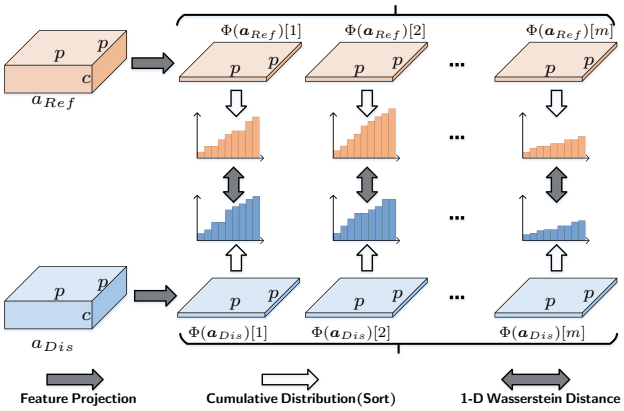


Figure 3. The proposed local sliced Wasserstein distance (LocalSW) calculation module which measures the 1-D Wasserstein distance between cumulative distribution of the projected reference and distortion feature maps.

3.3. FR-IQA Network Structure

As shown in Fig. 2, our proposed FR-IQA consists of a feature extraction network and a score prediction network. The feature extraction network adopts a Siamese (*i.e.*, dual-branch) structure, which respectively takes the reference image and the distorted image as the input. It is based on VGG16 [52] consisting of three different scales, *i.e.*, $s = 1$, 2 and 3. And we further modify the VGG16 network from two aspects. First, all max pooling layers in VGG are replaced with L_2 pooling [25] to avoid aliasing when downsampling by a factor of two. Second, to increase the fitting ability, dual attention blocks (DAB) used in [67] are integrated into different scales of backbone network. The reference image $\boldsymbol{I}_{Ref}$ and distorted image $\boldsymbol{I}_{Dis}$ are fed into the feature extraction network to obtain the reference feature $\boldsymbol{f}^s_{Ref}$ and distortion feature $\boldsymbol{f}^s_{Dis}$ ($s = 1, 2, 3$), respectively. Then, local sliced Wasserstein (LocalSW) distance is presented to produce distance map $\boldsymbol{f}^s_{Dist}$, and a spatial attention module is deployed for reweighting distance map to generate calibrated difference map $\boldsymbol{f}^s_{Diff}$ for each scale s. As shown in Fig. 2, the score prediction network has three branches, where each branch involves two 1×1 convolutional layers and a spatial-wise global averaging pooling layer. $\boldsymbol{f}^s_{Diff}$ is fed to the s-th branch to generate the score at scale s, and the scores at all scales are averaged to produce the final score.

In the following, we elaborate more on the LocalSW distance and difference map calibration.

LocalSW Distance. Given the reference feature $\boldsymbol{f}^s_{Ref}$ and distortion feature $\boldsymbol{f}^s_{Dis}$, one direct solution is the element-wise difference, *i.e.*, $|\boldsymbol{f}^s_{Ref} - \boldsymbol{f}^s_{Dis}|$. Here $|\cdot|$ denotes element-wise absolute value. However, GAN-based restoration is prone to producing results being spatially distorted and misaligned with the reference image, while the element-wise difference is not robust to spatial misalignment. Instead, we suggest local sliced Wasserstein (LocalSW) distance which measures the difference by comparing the distributions of feature maps. Previously sliced Wasserstein loss [12, 24] has been proposed to calculate the global sliced Wasserstein distance. Considering that the misalignment between $\boldsymbol{f}^s_{Ref}$ and $\boldsymbol{f}^s_{Dis}$ is usually local and within a small range, we adopt LocalSW distance by dividing $\boldsymbol{f}^s_{Ref}$ and $\boldsymbol{f}^s_{Dis}$ ($\in \mathbb{R}^{H\times W\times C}$) into J non-overlapped patches with resolution $p \times p$, *i.e.*, $J = (H/p) \times (W/p)$. Fig. 3 illustrates the computation of LocalSW distance by using a patch pair $\boldsymbol{a}_{Ref}$ and $\boldsymbol{a}_{Dis}$ ($\in \mathbb{R}^{p\times p\times C}$) as an example. In particular, we first use the projection operator Φ on $\boldsymbol{a}_{Ref}$ and $\boldsymbol{a}_{Dis}$ to obtain the projected features $\Phi(\boldsymbol{a}_{Ref})$ and $\Phi(\boldsymbol{a}_{Dis})$ ($\in \mathbb{R}^{p\times p\times m}$), where $m = C/2$. Then, we implement the cumulative distributions through sorting operation $\mathbf{Sort}(\cdot)$ on each channel (*i.e.*, slice) v of $\Phi(\boldsymbol{a}_{Ref})$ and $\Phi(\boldsymbol{a}_{Dis})$. And the LocalSW distance for slice v of this patch pair can be obtained by,

$$\mathcal{SW}[v] = \|\mathbf{Sort}(\Phi(\boldsymbol{a}_{Ref})[v]) - \mathbf{Sort}(\Phi(\boldsymbol{a}_{Dis})[v])\|\,. \quad (10)$$

Furthermore, we compute the LocalSW distance for all slices and all patches to form the LocalSW distance map $\boldsymbol{f}^s_{Dist} \in \mathbb{R}^{\frac{H}{p}\times\frac{W}{p}\times m}$.

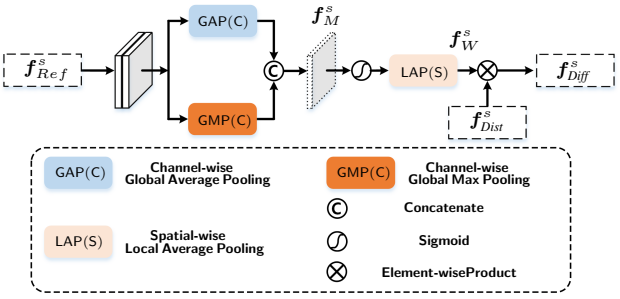


Figure 4. Spatial attention for difference map calibration, where spatial attention based on reference feature is used to reweight distance map for generating calibrated difference map.

Spatial Attention for Difference Map Calibration. Obviously, the contribution of image region to visual quality is spatially varying. Informative regions have more influences and should be emphasized more when predicting the final score. In learning-based FR-IQA, ASNA [3] computes spatial and channel attention based on decoder feature to improve MOS estimation. Actually, the importance of local region should be determined by the reference image instead of decoder feature and distance map. Thus, we adopt a much simple design by computing spatial attention based on reference feature while applying it on distance map to generate calibrated difference map. As show in Fig. 4, the spatial attention module takes reference feature $\boldsymbol{f}^s_{Ref}$ at scale s as input. Then, we use two 3×3 convolutional layers followed by global average pooling and max pooling along the channel dimension to form a feature map $\boldsymbol{f}^s_M$. Finally, a 1×1 convolutional layer followed by sigmoid activation

Table 1. Summary of five IQA databases, *i.e.*, LIVE [47], CSIQ [33], TID2013 [45], KADID-10k [35] and PIPAL [19]. DMOS is inversely proportional to MOS.

Dataset	#Ref.	#Dis.	#Dis. Type	#Rating	Rating Type	Score Range
LIVE [47]	29	779	5	25k	DMOS	[0, 100]
CSIQ [33]	30	866	6	5k	DMOS	[0, 1]
TID2013 [45]	25	3,000	24	524k	MOS	[0, 9]
KADID-10k [35]	81	10,125	25	30.4k	MOS	[1, 5]
PIPAL [19]	250	25,850	40	1.13m	MOS	[917, 1836]

and local average pooling is deployed to generate spatial weighting map $\boldsymbol{f}_W^s \in \mathbb{R}^{\frac{H}{p} \times \frac{W}{p}}$, where the size of the local average pooling region is set to $p \times p$. Calibrated difference map $\boldsymbol{f}_{Diff}^s$ can then be obtained by using $\boldsymbol{f}_W^s$ for reweighting each channel of distance map $\boldsymbol{f}_{Dist}^s$ in an element-wise manner, while final score can be predicted by feeding $\boldsymbol{f}_{Diff}^s$ into score prediction network.

3.4. Network Structure of Binary Classifier

The network structure of binary classifier is relatively simple, and contains two parts. The first part involves the first 12 convolutional layers in VGG16 (*i.e.*, 3 scales). The second part has the same structure as the score prediction network in our FR-IQA model.

4. Experiments

In this section, we first introduce experiment settings and implementation details of the proposed method. Then, we conduct ablation studies to analyze the proposed method, and compare it with state-of-the-art IQA methods on five benchmark datasets. Finally, we evaluate the generalization ability of our method.

4.1. Experiment Settings

Labeled Data. Five IQA datasets are employed in the experiments, including LIVE [47], CSIQ [33], TID2013 [45], KADID-10k [35] and PIPAL [19], whose configurations are presented in Table 1. LIVE [47], CSIQ [33] and TID2013 [45] are three relatively small-scale IQA datasets, where distorted images only contain traditional distortion types (*e.g.*, noise, downsampling, JPEG compression, *etc.*). KADID-10k [35] further incorporates the recovered results of a denoising algorithm into the distorted images, resulting in a medium-sized IQA dataset. Since the explicit splits of training, validation and testing are not given on these four datasets, we randomly partition the dataset into training, validation and testing sets by splitting reference images with ratios 60%, 20%, 20%, respectively. To reduce the bias caused by a random split, we run the random splits ten times. On these four datasets, the comparison results are reported as the average of ten times evaluation experiments.

PIPAL [19] is a large-scale IQA dataset. The training set consists of 200 reference images and $23,200$ distorted images with resolution of 288×288. The validation set consists of 25 reference images and $1,000$ distorted images. Since the testing set of PIPAL is not publicly available, we in this paper report the evaluation results on validation set via the online server[1]. The distorted images in PIPAL dataset include traditional distorted images and images restored by multiple types of image restoration algorithms (*e.g.*, denoising, super-resolution, deblocking, etc.) as well as GAN-based restoration models. It is worth noting that the distortion types in PIPAL validation set are unseen in the training set.

Unlabeled Data. We take $1,000$ image patches (288×288) randomly from DIV2K [1] validation set and Flickr2K [56] as reference images in unlabeled data. For the acquisition of distorted images, we adopt the following three manners: (i) ESRGAN Synthesis: All the reference images are downsampled, and then super-resolved using 50 groups of intermediate ESRGAN models. The restored images are regarded as distorted images in unlabeled data. (ii) DnCNN Synthesis: We add Gaussian noises to reference images to obtain degraded images, which are restored using 50 groups of intermediate DnCNN models. (iii) KADID-10k Synthesis: Following [35], we add 25 degradation types to reference images by randomly select 2 of 5 distortion levels for obtaining distortion images in unlabeled data. More details of intermediate models of ESRGAN and DnCNN can be found in the supplementary material. We note that ESRGAN and DnCNN are not adopted in validation set of PIPAL, guaranteeing non-intersection of distortion types in PIPAL validation set and our collected unlabeled data.

Evaluation Criteria. Two evaluation criteria are reported for each experimental setup, *i.e.*, Spearman Rank Correlation Coefficient (SRCC) for measuring prediction accuracy, and Pearson Linear Correlation Coefficient (PLCC) for measuring prediction monotonicity.

4.2. Implementation Details

We use the Adam optimizer [29] for all models presented in this paper with a batchsize of 32. We randomly crop the image patches with size 224×224, and perform flipping (horizontal/vertical) and rotating ($90°$, $180°$, or $270°$) on training samples for data augmentation.

Supervised Learning. We train the proposed FR-IQA model with labeled data for total 20,000 iterations. The learning rate is initialized to $1e$-4, and decreased to $1e$-5 after 10,000 iteration. Moreover, we have found empirically that even if the training iterations are further increased, the IQA model will not get any performance improvement.

Joint Semi-supervised and PU Learning. We initialize the network parameters using the pre-trained IQA model with the learning rate of $1e$-5 for 20,000 iterations. The pseudo MOS y_j^* is initialized with the pre-trained IQA

[1] https://competitions.codalab.org/competitions/28050

Table 2. PLCC / SRCC performance with ablation studies about network structure performed on the PIPAL [19] and KADID-10k [35].

NO.	DAB	SA	LocalSW	PIPAL PLCC / SRCC	KADID-10k PLCC / SRCC
1	✗	✗	✗	0.835 / 0.824	0.899 / 0.889
2	✓	✗	✗	0.843 / 0.837	0.908 / 0.905
3	✗	✓	✗	0.849 / 0.838	0.927 / 0.919
4	✓	✓	✗	0.852 / 0.849	0.941 / 0.940
5	✓	✗	✓	0.861 / 0.857	0.929 /0.925
6	✓	✓	✓	**0.868 / 0.868**	**0.943 / 0.944**

model. Hyper-parameter p, *i.e.*, the region size in local Sliced Wasserstein distance (LocalSW), is set to 8 and 2 for PIPAL and traditional IQA datasets, respectively. The momentum parameter α is set to 0.95. Hyperparameter τ changes with iterations, *i.e.*, $\tau = \max\{\tau_0^{t/T_0}, \tau_{\min}\}$ for t-th iteration, where parameters τ_0, T_0 and $\tau_{\min}$ are set as 0.9, $1,000$ and 0.5, respectively.

4.3. Ablation Study

All the ablation experiments are performed on PIPAL [19] and KADID-10k [35], considering that the distortion types of these two datasets are very different.

Network Structure. We first study the effects of our three architectural components, *i.e.*, Dual Attention Block (DAB), Spatial Attention (SA), and Local Sliced Wasserstein Distance (LocalSW). In Table 2, one can see that on PIPAL dataset, removing the LocalSW results in the greatest performance degradation, which is mainly due to the additional computational error introduced by the spatial misalignment in the GAN-based distorted images. When the SA module is eliminated, the IQA model assigns the same weight to different information content areas, resulting in low accuracy. Similarly, DAB also contributes to the final performance.

Training Strategy. We conduct ablation experiments on three different types of unlabeled data, *i.e.*, ESRGAN Synthesis, DnCNN Synthesis, KADID-10k Synthesis, and compare the proposed JSPL with semi-supervised learning (SSL), *i.e.*, combining labeled and unlabeled data without PU learning. From Table 3, we have the following observations: (i) First, compared to the other two syntheses types, the distribution of unlabeled data using ESRGAN Synthesis is more consistent with the labeled PIPAL dataset, leading to the greater performance gains. Similarly, the KADID-10k dataset has same distortion types with KADID-10k Synthesis. It indicates that the inconsistent distribution between labeled and unlabeled data is a key issue for semi-supervised learning. Therefore, in the subsequent experiments, we choose unlabeled data that are closer to the distribution of the labeled data. (ii) Second, from the six sets of comparative experiments on SSL and JSPL, we can see that JSPL performs better than SSL. This is because our JSPL can exclude negative outliers, making the distribution of labeled data and positive unlabeled data be more consistent, while SSL is adversely affected by these outliers.

Table 3. PLCC / SRCC results obtained using different data settings with SL, SSL or JSPL manners on PIPAL [19] and KADID-10k [35].

Methods	PIPAL Unlabeled Data	PIPAL PLCC / SRCC	KADID10k Unlabeled Data	KADID10k PLCC / SRCC
SL	-	0.868 / 0.868	-	0.943 / 0.944
SSL	ESRGAN Synthesis	0.872 / 0.870	ESRGAN Synthesis	0.930 / 0.932
	DnCNN Synthesis	0.870 / 0.868	DnCNN Synthesis	0.945 / 0.944
	KADID-10k Synthesis	0.867 / 0.866	KADID-10k Synthesis	0.959 / 0.958
JSPL	ESRGAN Synthesis	**0.877 / 0.874**	ESRGAN Synthesis	0.945 / 0.948
	DnCNN Synthesis	0.875 / 0.872	DnCNN Synthesis	0.959 / 0.957
	KADID-10k Synthesis	0.873 / 0.870	KADID-10k Synthesis	**0.963 / 0.961**

Table 4. Performance comparison of IQA methods on PIPAL [19] dataset. Some results are provided from the NTIRE 2021 IQA challenge report [20].

Methods	Category	PLCC	SRCC	Methods	Category	PLCC	SRCC
MA [41]	NR	0.203	0.201	PSNR	FR	0.292	0.255
PI [5]		0.166	0.169	SSIM [58]		0.398	0.340
NIQE [43]		0.102	0.064	LPIPS-Alex [73]		0.646	0.628
VIF [48]	FR	0.524	0.433	LPIPS-VGG [73]		0.647	0.591
VSNR [7]		0.375	0.321	PieAPP [46]		0.697	0.706
VSI [70]		0.516	0.450	WaDIQaM-FR [6]		0.654	0.678
MAD [33]		0.626	0.608	DISTS [13]		0.686	0.674
NQM [11]		0.416	0.346	SWD [19]		0.668	0.661
UQI [57]		0.548	0.486	EGB [23]		0.775	0.776
IFC [49]		0.677	0.594	DeepQA [2]		0.795	0.785
GSM [36]		0.469	0.418	ASNA [3]		0.831	0.824
RFSIM [71]		0.304	0.266	RADN [50]		0.867	0.866
SRSIM [69]		0.654	0.566	IQMA [21]		0.876	0.872
FSIM [72]		0.561	0.467	IQT [9]		0.876	0.865
FSIMc [72]		0.559	0.468	Ours(SL)		0.868	0.868
MS-SSIM [61]		0.563	0.486	Ours(JSPL)		**0.877**	**0.874**

4.4. Comparison with State-of-the-arts

4.4.1 Evaluation on PIPAL Dataset

As shown in Table 4, we compare 18 traditional evaluation metrics and 12 CNN-based FR-IQA models with the proposed model under two different learning strategies, *i.e.*, supervised learning (SL) and JSPL. Compared with traditional evaluation metrics, CNN-based FR-IQA models are proven to be more consistent with human subjective quality scoring. Albeit retraining on the PIPAL dataset, the performance of pioneering CNN-based FR-IQA models, *e.g.*, LPIPS [73], WaDIQaM-FR [6] and DISTS [13] are still limited. Although SWDN [18] designed a pixel-by-pixel alignment module to address the misalignment problem in GAN-based distortion, the corresponding feature extraction network is not sufficiently effective to achieve satisfactory result. In contrast, considering both the properties of GAN-based distortion and the design of the feature extraction network, IQT [9], IQMA [21] and RADN [50] achieve top3 performance on PIPAL in published literatures. Because of the spatial attention and the LocalSW module, the proposed method using supervised learning obtains superior performance than RADN [50] on PIPAL. Although our FR-IQA model by adopting supervised learning strategy is slightly inferior to IQT [9] and IQMA [21], the proposed JSPL strategy significantly boosts its performance by exploiting adequate positive unlabled data while mitigating the adverse

Table 5. Performance evaluation on the LIVE [47], CSIQ [33], TID2013 [45] and KADID-10k [35] databases.

Methods	Category	LIVE		CSIQ		TID2013		KADID-10k	
		SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
BRISQUE [42]	NR	0.939	0.935	0.746	0.829	0.604	0.694	-	-
FRIQUEE [16]	NR	0.940	0.944	0.835	0.874	0.680	0.753	-	-
CORNIA [66]	NR	0.947	0.950	0.678	0.776	0.678	0.768	0.541	0.580
M3 [65]	NR	0.951	0.950	0.795	0.839	0.689	0.771	-	-
HOSA [64]	NR	0.946	0.947	0.741	0.823	0.735	0.815	0.609	0.653
Le-CNN [26]	NR	0.956	0.953	-	-	-	-	-	-
BIECON [27]	NR	0.961	0.962	0.815	0.823	0.717	0.762	-	-
DIQaM-NR [6]	NR	0.960	0.972	-	-	0.835	0.855	-	-
WaDIQaM-NR [6]	NR	0.954	0.963	-	-	0.761	0.787	-	-
ResNet-ft [28]	NR	0.950	0.954	0.876	0.905	0.712	0.756	-	-
IW-CNN [28]	NR	0.963	0.964	0.812	0.791	0.800	0.802	-	-
DB-CNN [74]	NR	0.968	0.971	0.946	0.959	0.816	0.865	0.501	0.569
CaHDC [62]	NR	0.965	0.964	0.903	0.914	0.862	0.878	-	-
HyperIQA [54]	NR	0.962	0.966	0.923	0.942	0.729	0.775	-	-
PSNR	FR	0.873	0.865	0.810	0.819	0.687	0.677	0.676	0.675
SSIM [58]	FR	0.948	0.937	0.865	0.852	0.727	0.777	0.724	0.717
MS-SSIM [61]	FR	0.951	0.940	0.906	0.889	0.786	0.830	0.826	0.820
VSI [70]	FR	0.952	0.948	0.942	0.928	0.897	0.900	0.879	0.877
FSIMc [72]	FR	0.965	0.961	0.931	0.919	0.851	0.877	0.854	0.850
MAD [33]	FR	0.967	0.968	0.947	0.950	0.781	0.827	0.799	0.799
VIF [48]	FR	0.964	0.960	0.911	0.913	0.677	0.771	0.679	0.687
DeepSim [15]	FR	0.974	0.968	-	-	0.846	0.872	-	-
DIQaM-FR [6]	FR	0.966	0.977	-	-	0.859	0.880	-	-
WaDIQaM-FR [6]	FR	0.970	0.980	-	-	**0.940**	0.946	-	-
DISTS [13]	FR	0.955	0.955	0.946	0.946	0.830	0.855	0.887	0.886
PieAPP [46]	FR	0.918	0.909	0.890	0.873	0.670	0.749	0.836	0.836
LPIPS [73]	FR	0.932	0.934	0.903	0.927	0.670	0.749	0.843	0.839
Ours(SL)	FR	0.970	0.978	0.965	0.968	0.924	0.912	0.944	0.943
Ours(JSPL)	FR	**0.980**	**0.983**	**0.977**	**0.970**	**0.940**	**0.949**	**0.961**	**0.963**

effects of outliers.

4.4.2 Evaluation on Traditional Datasets

Our methods with two learning manners, *i.e.*, SL and JSPL, are compared with the competitors on the other four traditional IQA datasets, including LIVE [47], CSIQ [33], TID2013 [45] and KADID-10k [35]. From Table 5 we can observe that the FR-IQA models achieve a higher performance compared to the NR-IQA models, since the pristine-quality reference image provides more accurate reference information for quality assessment. Although WaDIQaM-FR [6] achieves almost the same performance with our method in terms of the SRCC metric on TID2013 dataset, but is inferior to ours on LIVE and PIPAL datasets, indicating its limited generalization ability. On all testing sets, the proposed FR-IQA model with SL strategy still delivers superior performance, which reveals the effectiveness of the proposed spatial attention and LocalSW module. By adopting JSPL strategy, our FR-IQA model achieves the best performance on all the four datasets. More comparisons on individual distortion types and cross-datasets are provided in supplementary material.

4.5. Evaluating Generalization Ability

Considering that distortion types in KADID-10k and PIPAL are not similar, we adopt these two datasets for evaluating generalization ability of our method as well as IQT [9], a state-of-the-art method in Table 4. As shown in Table 6, both IQT and our method can obtain satisfying performance when keeping consistent validation and training sets from PIPAL. However, significant performance degradations can be observed when applying the models learned based on KADID-10k to validation set of PIPAL. This is because the distribution discrepancy between KADID-10k and PIPAL is severe, which cannot be addressed by SL strategy. By adopting SSL and JSPL, unlabeled data using ESRGAN Synthesis is introduced. Although SSL utilizes unlabeled data, the performance drops can be observed for IQT and our method due to the effect of outliers, which demonstrates that the elimination of outliers is essential. In contrast, our JSPL can exclude negative outliers while exploiting positive unlabeled data, significantly boosting generalization ability of IQT and our method. In comparison to IQT with JSPL, our method with JSPL has better generalization ability, which can be attributed to the novel modules SA and LocalSW in our FR-IQA model.

Table 6. PLCC / SRCC assessment about IQA models trained on different settings, and tested on the PIPAL [19] Val.

Methods	Training Data Labeld Data (& Unlabeled Data)	PIPAL Val. PLCC / SRCC
IQT(SL)	PIPAL	0.876 / 0.865
IQT(SL)	KADID-10k	0.741 / 0.718
IQT(SSL)	KADID-10k & ESRGAN Synthesis	0.700 / 0.662
IQT(JSPL)	KADID-10k & ESRGAN Synthesis	0.794 / 0.783
Our(SL)	PIPAL	0.868 / 0.868
Ours(SL)	KADID-10k	0.756 / 0.770
Ours(SSL)	KADID-10k & ESRGAN Synthesis	0.733 / 0.766
Ours(JSPL)	KADID-10k & ESRGAN Synthesis	0.804 / 0.801

5. Conclusion

In this paper, we proposed a joint semi-supervised and PU learning (JSPL) to exploit unlabelled data for boosting performance of FR-IQA, while mitigating the adverse effects of outliers. We also introduced a novel FR-IQA network, embedding spatial attention and local sliced Wasserstein distance (LocalSW) for emphasizing informative regions and suppressing the effect of misalignment between distorted and pristine images, respectively. Extensive experimental results show that the proposed JSPL algorithm can improve the performance of the FR-IQA model as well as the generalization capability. In the future, the proposed JSPL algorithm can be extended to more challenging image quality assessment tasks, *e.g.*, NR-IQA.

Acknowledgement

This work was supported in part by National Key RD Program of China under Grant 2021ZD0112100, and National Natural Science Foundation of China under Grants No. 62172127, No. U19A2073 and No. 62102059.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: dataset and study. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 126–135, 2017. 6, 12, 14

[2] Sewoong Ahn, Yeji Choi, and Kwangjin Yoon. Deep learning-based distortion sensitivity prediction for full-reference image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 344–353, 2021. 2, 7

[3] Seyed Ayyoubzadeh and Ali Royat. (ASNA) An attention-based siamese-difference neural network with surrogate ranking loss function for perceptual image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 388–397, 2021. 3, 5, 7, 14

[4] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. MixMatch: A holistic approach to semi-supervised learning. In *IEEE International Conference on Advances in Neural Information Processing Systems*, volume 32, 2019. 4

[5] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 PIRM challenge on perceptual image super-resolution. In *European Conference on Computer Vision Workshops*, pages 0–0, 2018. 7

[6] 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, 2017. 2, 7, 8, 12, 13, 14

[7] Damon Chandler and Sheila Hemami. VSNR: A wavelet-based visual signal-to-noise ratio for natural images. *IEEE Transactions on Image Processing*, 16(9):2284–2298, 2007. 7

[8] Xuxi Chen, Wuyang Chen, Tianlong Chen, Ye Yuan, Chen Gong, Kewei Chen, and Zhangyang Wang. Self-PU: Self boosted and calibrated positive-unlabeled training. In *IEEE International Conference on Machine Learning*, pages 1510–1519. PMLR, 2020. 2

[9] Manri Cheon, Sung-Jun Yoon, Byungyeon Kang, and Junwoo Lee. Perceptual image quality assessment with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 433–442, 2021. 3, 7, 8, 12, 13, 14, 15

[10] Fabio Gagliardi Cozman, Ira Cohen, and Marcelo Cesar Cirelo. Semi-supervised learning of mixture models. In *IEEE International Conference on Machine Learning*, volume 4, page 24, 2003. 3

[11] Niranjan Damera-Venkata, Thomas Kite, Wilson Geisler, Brian Evans, and Alan Bovik. Image quality assessment based on a degradation model. *IEEE Transactions on Image Processing*, 9(4):636–650, 2000. 7

[12] Mauricio Delbracio, Hossein Talebi, and Peyman Milanfar. Projected distribution loss for image enhancement. *arXiv preprint arXiv:2012.09289*, 2020. 5

[13] Keyan Ding, Kede Ma, Shiqi Wang, and Eero Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 3, 7, 8, 12, 13, 14

[14] Ulrich Engelke, Vuong Nguyen, and Hans-Jurgen Zepernick. Regional attention to structural degradations for perceptual image quality metric design. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 869–872. IEEE, 2008. 2

[15] Fei Gao, Yi Wang, Panpeng Li, Min Tan, Jun Yu, and Yani Zhu. Deepsim: Deep similarity for image quality assessment. *Neurocomputing*, 257:104–114, 2017. 2, 8

[16] Deepti Ghadiyaram and Alan Bovik. Perceptual quality prediction on authentically distorted images using a bag of features approach. *Journal of Vision*, 17(1):32–32, 2017. 8

[17] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *International Conference on Neural Information Processing Systems*, pages 529–536, 2004. 4

[18] Jinjin Gu, Haoming Cai, Haoyu Chen, Xiaoxing Ye, Jimmy Ren, and Chao Dong. Image quality assessment for perceptual image restoration: A new dataset, benchmark and metric. *arXiv preprint arXiv:2011.15002*, 2020. 2, 3, 7

[19] Jinjin Gu, Haoming Cai, Haoyu Chen, Xiaoxing Ye, Jimmy 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. Springer, 2020. 2, 6, 7, 8, 13, 14, 16

[20] Jinjin Gu, Haoming Cai, Chao Dong, Jimmy Ren, Yu Qiao, Shuhang Gu, and Radu Timofte. Ntire 2021 challenge on perceptual image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 677–690, 2021. 7

[21] Haiyang Guo, Yi Bin, Yuqing Hou, Qing Zhang, and Hengliang Luo. IQMA network: Image quality multi-scale assessment network. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 443–452, 2021. 3, 7

[22] Lan-Zhe Guo, Zhen-Yu Zhang, Yuan Jiang, Yu-Feng Li, and Zhi-Hua Zhou. Safe deep semi-supervised learning for unseen-class unlabeled data. In *IEEE International Conference on Machine Learning*, pages 3897–3906. PMLR, 2020. 2, 3

[23] Dounia Hammou, Sid Fezza, and Wassim Hamidouche. EGB: Image quality assessment based on ensemble of gradient boosting. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 541–549, 2021. 3, 7

[24] Eric Heitz, Kenneth Vanhoey, Thomas Chambon, and Laurent Belcour. A sliced Wasserstein loss for neural texture synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9412–9420, 2021. 5

[25] Olivier Hénaff and Eero Simoncelli. Geodesics of learned representations. *arXiv preprint arXiv:1511.06394*, 2015. 5

[26] 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. 8

[27] Jongyoo Kim and Sanghoon Lee. Fully deep blind image quality predictor. *IEEE Journal of Selected Topics in Signal Processing*, 11(1):206–220, 2016. 8

[28] Jongyoo Kim, Hui Zeng, Deepti Ghadiyaram, Sanghoon Lee, Lei Zhang, and Alan Bovik. Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment. *IEEE Signal Processing Magazine*, 34(6):130–141, 2017. 8

[29] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[30] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. Imagenet classification with deep convolutional neural networks. *IEEE International Conference on Advances in Neural Information Processing Systems*, 25:1097–1105, 2012. 3

[31] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*, 2018. 4

[32] Eric Larson and Damon Chandler. Unveiling relationships between regions of interest and image fidelity metrics. In *Visual Communications and Image Processing*, volume 6822, page 68222A. International Society for Optics and Photonics, 2008. 2

[33] Eric Larson and Damon Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):011006, 2010. 2, 6, 7, 8, 16

[34] Eric Larson, Cuong Vu, and Damon Chandler. Can visual fixation patterns improve image fidelity assessment? In *IEEE International Conference on Image Processing*, pages 2572–2575. IEEE, 2008. 2

[35] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. KADID-10k: A large-scale artificially distorted IQA database. In *IEEE International Conference on Quality of Multimedia Experience*, pages 1–3. IEEE, 2019. 2, 6, 7, 8, 13, 14, 16

[36] Anmin Liu, Weisi Lin, and Manish Narwaria. Image quality assessment based on gradient similarity. *IEEE Transactions on Image Processing*, 21(4):1500–1512, 2011. 7

[37] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. In *IEEE International Conference on Advances in Neural Information Processing Systems*, volume 33, 2020. 4

[38] Siyuan Liu, Kim-Han Thung, Weili Lin, Dinggang Shen, and Pew-Thian Yap. Hierarchical nonlocal residual networks for image quality assessment of pediatric diffusion MRI with limited and noisy annotations. *IEEE Transactions on Medical Imaging*, 39(11):3691–3702, 2020. 1, 3

[39] Siyuan Liu, Kim-Han Thung, Weili Lin, Pew-Thian Yap, and Dinggang Shen. Real-time quality assessment of pediatric MRI via semi-supervised deep nonlocal residual neural networks. *IEEE Transactions on Image Processing*, 29:7697–7706, 2020. 1, 3

[40] Wen Lu, Ning Mei, Fei Gao, Lihuo He, and Xinbo Gao. Blind image quality assessment via semi-supervised learning and fuzzy inference. In *Applied Informatics*, volume 2, pages 1–20. SpringerOpen, 2015. 3

[41] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158:1–16, 2017. 7

[42] Anish Mittal, Anush Moorthy, and Alan Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. 8

[43] Anish Mittal, Rajiv Soundararajan, and Alan Bovik. Making a 'completely blind' image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2012. 7

[44] Jiri Najemnik and Wilson Geisler. Optimal eye movement strategies in visual search. *Nature*, 434(7031):387–391, 2005. 2

[45] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. Image database TID2013: Peculiarities, results and perspectives. *Signal processing: Image communication*, 30:57–77, 2015. 2, 6, 8, 16

[46] Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1808–1817, 2018. 7, 8, 12, 13, 14

[47] Hamid Sheikh. Image and video quality assessment research at live. *http://live. ece. utexas. edu/research/quality*, 2003. 2, 6, 8, 16

[48] Hamid Sheikh and Alan Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2):430–444, 2006. 7, 8

[49] Hamid Sheikh, Alan Bovik, and Gustavo De Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on Image Processing*, 14(12):2117–2128, 2005. 7

[50] Shuwei Shi, Qingyan Bai, Mingdeng Cao, Weihao Xia, Jiahao Wang, Yifan Chen, and Yujiu Yang. Region-adaptive deformable network for image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 324–333, 2021. 3, 7

[51] Eero Simoncelli and Bruno Olshausen. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24(1):1193–1216, 2001. 2

[52] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3, 5

[53] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin Raffel, Ekin Cubuk, Alexey Kurakin, and Chun-Liang Li. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *IEEE International Conference on Advances in Neural Information Processing Systems*, volume 33, 2020. 4

[54] 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*, June 2020. 8

[55] Huixuan Tang, Neel Joshi, and Ashish Kapoor. Blind image quality assessment using semi-supervised rectifier networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2877–2884, 2014. 1, 3

[56] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single

image super-resolution: Methods and results. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 114–125, 2017. 6, 12, 14

[57] Zhou Wang and Alan Bovik. A universal image quality index. *IEEE Signal Processing Letters*, 9(3):81–84, 2002. 7

[58] Zhou Wang, Alan Bovik, Hamid Sheikh, and Eero Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 1, 7, 8, 15

[59] Zhihua Wang, Dingquan Li, and Kede Ma. Semi-supervised deep ensembles for blind image quality assessment. *International Joint Conference on Artificial Intelligence Workshops*, 2021. 1, 3

[60] Zhou Wang and Qiang Li. Information content weighting for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 20(5):1185–1198, 2010. 2, 14

[61] Zhou Wang, Eero Simoncelli, and Alan Bovik. Multiscale structural similarity for image quality assessment. In *Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. IEEE, 2003. 7, 8, 15

[62] 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, 2020. 8

[63] Junshen Xu, Sayeri Lala, Borjan Gagoski, Esra Abaci Turk, P Ellen Grant, Polina Golland, and Elfar Adalsteinsson. Semi-supervised learning for fetal brain MRI quality assessment with roi consistency. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 386–395. Springer, 2020. 1, 3

[64] 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, 2016. 8

[65] Wufeng Xue, Xuanqin Mou, Lei Zhang, Alan Bovik, and Xiangchu Feng. Blind image quality assessment using joint statistics of gradient magnitude and laplacian features. *IEEE Transactions on Image Processing*, 23(11):4850–4862, 2014. 8

[66] 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. IEEE, 2012. 8

[67] Syed Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Khan, Ming-Hsuan Yang, and Ling Shao. CycleISP: Real image restoration via improved data synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2696–2705, 2020. 5

[68] Jiaqi Zhang, Zhenzhen Wang, Junsong Yuan, and Yap-Peng Tan. Positive and unlabeled learning for anomaly detection with multi-features. In *ACM International Conference on Multimedia*, pages 854–862, 2017. 2

[69] Lin Zhang and Hongyu Li. SR-SIM: A fast and high performance IQA index based on spectral residual. In *IEEE International Conference on Image Processing*, pages 1473–1476. IEEE, 2012. 7

[70] Lin Zhang, Ying Shen, and Hongyu Li. VSI: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 23(10):4270–4281, 2014. 7, 8

[71] Lin Zhang, Lei Zhang, and Xuanqin Mou. RFSIM: A feature based image quality assessment metric using riesz transforms. In *IEEE International Conference on Image Processing*, pages 321–324. IEEE, 2010. 7

[72] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. FSIM: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 2011. 7, 8

[73] Richard Zhang, Phillip Isola, Alexei Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 3, 7, 8, 12, 13, 14, 15

[74] 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, 2018. 8

[75] Xuan Zhao, Yali Li, and Shengjin Wang. Face quality assessment via semi-supervised learning. In *IEEE International Conference on Computing and Pattern Recognition*, pages 288–293, 2019. 3

Incorporating Semi-Supervised and Positive-Unlabeled Learning for Boosting Full Reference Image Quality Assessment Supplemental Materials

The content of this supplementary material includes:

A. Limitation and Negative Impact in Sec. A.
B. ESRGAN and DnCNN Synthesis Process in Sec. B.
C. More Comparisons on Individual Distortion Types and Cross-dataset in Sec. C.
D. More Ablation Studies in Sec. D.
E. Discussion in Sec. E
F. More Details on IQA Datasets in Sec. F.

A. Limitation and Negative Impact

The proposed FR-IQA model predicts image quality by measuring the fidelity deviation from its pristine-quality reference. Unfortunately, in the vast majority of practical applications, reference images are not always available or difficult to obtain, which indicates our method is limited especially for authentically-distorted images.

B. ESRGAN and DnCNN Synthesis Process

For ESRGAN Synthesis, we adopt the DIV2K [1] training set as clean high-resolution (HR) images and employ the bicubic downsampler with the scale factor 2 to obtain the low-resolution (LR) images. Then, we retrain the original ESRGAN model using HR-LR pairs with the size of 128×128 and 64×64 cropped from the training HR and LR images, respectively. The ESRGAN model is trained with the GAN loss for 50 epochs and 50 groups of intermediate ESRGAN models are obtained. The learning rate is initialized to $2e$-4 and then decayed to $2e$-5 after 20 epochs. We take $1,000$ image patches (288×288) randomly from DIV2K [1] validation set and Flickr2K [56] as reference images in unlabeled data, which are propagated into the bicubic downsampler to obtain the degraded images. The corresponding distorted images can be obtained by feeding the degraded images into 50 groups of intermediate ESRGAN models.

For synthetic noises in DnCNN Synthesis, we use the additive white Gaussian noise with noise level 25. DnCNN is trained to learn a mapping from noisy image to denoising result. The DnCNN model is trained with the MSE loss for 50 epochs and 50 groups of intermediate DnCNN models are obtained. The learning rate is fixed to $1e$-4 and then decayed to $1e$-5 after 25 epochs. Similarly, we also take same $1,000$ image patches as reference images in unlabeled data. The restored images can be achieved by feeding the noisy images into 50 groups of intermediate DnCNN models, which are regared as the corresponding distorted images in unlabeled data.

Table A. SRCC comparisons on individual distortion types on the LIVE database. Red and blue are utilized to indicate top 1st and 2nd rank, respectively.

Database	LIVE				
Type	WN	JPEG	JP2K	FF	GB
WaDIQaM-FR [6]	0.975	0.959	0.934	0.941	0.915
DISTS [13]	0.969	0.982	0.971	0.961	0.969
PieAPP [46]	0.963	0.941	0.885	0.920	0.867
LPIPS [73]	0.968	0.982	0.968	0.955	0.918
our(SL)	0.983	0.984	0.952	0.967	0.912
our(JSPL)	0.984	0.986	0.959	0.968	0.943

C. More Comparisons on Individual Distortion Types and Cross-dataset

Comparisons on Individual Distortion Types. To further investigate the behaviors of our proposed method, we exhibit the performance on individual distortion type and compare it with several competing FR-IQA models on LIVE. The LIVE dataset contains five distortion types, *i.e.*, additive white Gaussian noise (WN), JPEG compression (JPEG), JPEG2000 compression (JP2K), Gaussian blur (GB) and Rayleigh fast-fading channel distortion (FF). As shown in Table A, the average SRCC values of above ten groups are reported. It is worth noting that our methods achieve significant performance improvements on three distortion types, *i.e.*, WN, JPEG and FF. Overall, better consistency with subjective scores and the consistently stable performance across different distortion types of the proposed scheme makes it the best IQA metric among all the compared metrics.

Comparisons on Cross-dataset. To verify the generalization capability, we further evaluate the proposed method on three groups of cross-dataset settings. We compare five FR-IQA methods, including: WaDIQaM-FR [6], DISTS [13], PieAPP [46], LPIPS [73] and IQT [9] with the proposed model under two different learning strategies, *i.e.*, SL and JSPL. We retrain the DISTS [13], PieAPP [46] and

Table B. SRCC comparisons on different cross-dataset with the PIPAL as training set. Red and blue are utilized to indicate top 1st and 2nd rank, respectively.

Methods	Traingning Set Labeld Data (& Unlabeled Data)	Test Sets LIVE	CSIQ	TID2013	KADID-10k
WaDIQaM-FR [6]	PIPAL	0.910	0.877	0.802	0.713
DISTS [13]	PIPAL	0.913	0.876	0.803	0.706
PieAPP [46]	PIPAL	0.904	0.875	0.762	0.699
LPIPS [73]	PIPAL	0.908	0.863	0.795	0.717
IQT [9]	PIPAL	0.917	0.880	0.796	0.718
our(SL)	PIPAL	0.919	0.873	0.804	0.717
our(JSPL)	PIPAL & KADID-10k Synthesis	0.930	0.894	0.812	0.776

Table C. SRCC comparisons on different cross-dataset with the KADID10k as training set. Red and blue are utilized to indicate top 1st and 2nd rank, respectively.

Methods	Traingning Set Labeld Data (& Unlabeled Data)	Test Sets LIVE	CSIQ	TID2013	PIPAL Val.
WaDIQaM-FR [6]	KADID-10k	0.948	0.931	0.861	0.712
DISTS [13]	KADID-10k	0.954	0.939	0.881	0.703
PieAPP [46]	KADID-10k	0.917	0.936	0.856	0.633
LPIPS [73]	KADID-10k	0.932	0.917	0.821	0.671
IQT [9]	KADID-10k	0.970	0.943	0.899	0.718
our(SL)	KADID-10k	0.973	0.951	0.908	0.770
our(JSPL)	KADID-10k & KADID-10k Synthesis	0.974	0.953	0.910	-
our(JSPL)	KADID-10k & ESRGAN Synthesis	-	-	-	0.801

Table D. SRCC comparisons on different cross-dataset with the TID2013 as training set. Red and blue are utilized to indicate top 1st and 2nd rank, respectively.

Methods	Traingning Set Labeld Data (& Unlabeled Data)	Test Sets LIVE	CSIQ	KADID-10k	PIPAL Val.
WaDIQaM-FR [6]	TID2013	0.911	0.913	0.760	0.552
DISTS [13]	TID2013	0.923	0.914	0.737	0.458
PieAPP [46]	TID2013	0.888	0.886	0.573	0.401
LPIPS [73]	TID2013	0.895	0.913	0.761	0.595
IQT [9]	TID2013	0.940	0.929	0.775	0.639
our(SL)	TID2013	0.944	0.932	0.762	0.651
our(JSPL)	TID2013 & KADID-10k Synthesis	0.948	0.934	0.795	-
our(JSPL)	TID2013 & ESRGAN Synthesis	-	-	-	0.699

LPIPS [73] by the source codes provided by the authors. Although the source training code for WaDIQaM-FR and IQT is not publicly available, we reproduce WaDIQaM-FR [6] and IQT [9], and achieve the similar performance of the original paper. From Table B, all FR-IQA models with supervised learning (SL) are trained using the largest human-rated IQA dataset, *i.e.*, PIPAL, so the results on the other four test datasets are relatively close. Because our approach with JSPL makes full use of unlabeled KADID-10k Synthesis which contains the same distortion types with KADID-10k, the higher performance on KADID-10k can be obtained.

From Table. C, all FR-IQA models with supervised learning (SL) are trained on KADID-10k, which contains the most diverse traditional distortion types. Therefore, compared to training on PIPAL or TID2013, all the FR-IQA methods achieve the best performance on traditional IQA datasets, *e.g.*, LIVE and CSIQ. Compared to other FR-IQA models, the proposed FR-IQA designs the spatial attention to deploy in computing difference map for emphasizing informative regions, and achieves the best performance in all FR-IQA models with supervised learning. However, when testing on PIPAL which contains distortion images restored by multiple types of image restoration algorithms as well as GAN-based restoration, significant performance degradation can be observed due to the distribution variation among different datasets. To alleviate this problem, the proposed JSPL strategy can improve performance to some extent for the use of unlabeled data.

From Table. D, all FR-IQA models with supervised learning (SL) are trained on TID2013. Due to fewer human-annotations and distorted samples are provided in TID2013, compared to KADID-10k, performance drop can be observed on traditional datasets, *e.g.*, LIVE and CSIQ, which indicates the collection of massive MOS annotations is beneficial to the performance improvement. However, the collection of massive MOS annotations is very time-consuming and cumbersome. In this work, we consider a more encouraging and practically feasible SSL setting, *i.e.*, training FR-IQA model using labeled data as well as unlabeled data. Based on three groups of cross-dataset experiments, the proposed JSPL can exploit positive unlabeled data, and significantly boost the performance and the generalization ability of FR-IQA.

Table E. PLCC / SRCC results for computing spatial attention based on different features.

Based on	PIPAL Val.
Reference feature $\boldsymbol{f}^s_{Ref}$	**0.868 / 0.868**
Distortion feature $\boldsymbol{f}^s_{Dis}$	0.861 / 0.860
Distance map $\boldsymbol{f}^s_{Dist}$	0.864 / 0.864

Table F. Performance on different attention mechanism on PIPAL.

Attention Mechanism Spatial	Channel	SRCC
✗	✗	0.857
✓	✗	**0.868**
✗	✓	0.840
✓	✓	0.859

Table G. PLCC / SRCC results for varying threshold parameter (*i.e.*, τ_{min}) on PIPAL [19] and KADID-10k [35].

τ_{min}	PIPAL PLCC / SRCC	KADID-10k PLCC / SRCC
0.4	0.872 / 0.870	0.951 / 0.949
0.5	**0.877 / 0.874**	**0.963 / 0.961**
0.6	0.874 / 0.872	0.955 / 0.955

Table H. SRCC performance on different sliced Wasserstein. p denotes local region size.

Methods		PIAPL	KADID-10k
Global		0.755	0.509
Local	$p = 32$	0.820	0.881
	$p = 16$	0.862	0.928
	$p = 8$	**0.868**	0.933
	$p = 4$	0.866	0.939
	$p = 2$	0.864	**0.944**
	$p = 1$	0.857	0.940

Table I. PLCC / SRCC comparisons on different FR-IQA with SL or JSPL training on PIPAL. Red and blue are utilized to indicate top 1st and 2nd rank, respectively.

Method	SL	JSPL
WaDIQaM-FR [6]	0.778 / 0.761	0.793 / 0.775
DISTS [13]	0.813 / 0.806	0.822 / 0.812
PieAPP [46]	0.785 / 0.778	0.806 / 0.796
LPIPS [73]	0.790 / 0.790	0.809 / 0.802
IQT [9]	0.876 / 0.865	0.876 / 0.873
our	0.868 / 0.868	0.877 / 0.874

D. More Ablation Studies

Spatial Attention. As far as the design of spatial attention, we adopt a much simple design by computing spatial attention based on the reference feature while applying it to the distance map to generate calibrated difference map. We conduct the ablation study by computing spatial attention based on different features, *i.e.*, the reference feature $\boldsymbol{f}^s_{Ref}$, the distortion feature $\boldsymbol{f}^s_{Dis}$ and the distance map $\boldsymbol{f}^s_{Dist}$. Considering the superiority of extracting features from reference in Table E, individual spatial attention on reference features is finally adopted in our method, while in ASNA [3], spatial attention and channel attention are directly adopted on distance map. In Table F, ablation studies on attention mechanism are reported, where individual spatial attention on reference features performs best. In IW-SSIM [60], spatially local information is suggested as one key factor for assessing distortions, which motivates us to only adopt spatial attention.

Hyper-parameter τ_{min}. We study the effects of threshold parameter, *i.e.*, τ_{min} on PIPAL [19] and KADID-10k [35]. From Table G, the best performance is achieved on both two datasets when τ_{min} is set to 0.5.

LocalSW. As for LocalSW, we suggest that local regions with proper size are more suitable for assessing distortions. As shown in Table H, region size $p = 8$ is the best choice on PIPAL, while original sliced Wasserstein (Global) yields significant performance drop. We further study the effects of hyper-parameter p on PIPAL [19] and KADID-10k [35], because the distortion types of these two datasets are very different. Due to the spatial misalignment properties of GAN-based distorted images in PIPAL, when the region size p is set to 8, the proposed LocalSW can compare the features within the most appropriate range around the corresponding position as shown in Table H. When applied to traditional dataset, *i.e.*, KADID-10k, the LocalSW with the hyper-parameter $p = 2$ achieves the best results.

Applying JSPL to Different FR-IQA models. To verify the generalization capability of JSPL, we apply the proposed JSPL to 6 different FR-IQA models, and use the PIPAL training set to retrain the 6 different FR-IQA models. From Table I, the pioneering CNN-based FR-IQA models, *e.g.*, WaDIQaM-FR [6], DISTS [13], PieAPP [46] and LPIPS [73] trained with PIPAL in supervised learning manner perform better than the original models (Table 4 in the manuscript) on PIPAL validation set. In terms of the SRCC metric, the proposed FR-IQA achieves the best performance with the help of LocalSW and spatial attention. Compared to the supervised learning, the proposed JSPL can further boost the performance of all six FR-IQA models, which indicates that the proposed learning strategy has good generalization ability.

Table J. Total number of distortion images (# U), number of positive samples (# PU) and number of negative samples (# NU) in the different distortion types.

Distortion Types	# U	# PU	# NU
DnCNN denoising algorithm	2,000	1,996	4
Gaussian blur	2,000	1,996	4
Additive white Gaussian noise	2,000	1,979	21
Color over-saturation	2,000	0	2,000
Color blocking	2,000	10	1,990
Sharpness	2,000	12	1,988

(a) Reference (b) GB (c) WN (d) DN

Figure A. Visualization of the excluded outliers, *i.e.*, the corresponding reference images, DnCNN denoising (DN) distorted images, Gaussian blur (GB) distorted images and additive white Gaussian noise (WN) distorted images.

E. Discussion

More Analysis on Binary Classifier. The labeled IQA datasets [19, 35] selected reference images which are representative of a wide variety of real-world textures, and should not be over-smooth or monochromatic. The reference images in unlabeled data are chosen randomly from DIV2K [1] validation set and Flickr2K [56], hence a small number of images may not meet the requirements. The unlabeled data may also contain distorted images which differ significantly from the distribution of the labeled data.

To verify that the binary classifier can eliminate the outliers mentioned above, we conduct the experiment to analyze the positive unlabeled data and outliers selected by the classifier. Take our FR-IQA as an example, the PIPAL training samples are selected as labeled data and the unlabeled data are considered to use the KADID-10k Synthesis, which contain multiple distortion types and are more useful for analysis than ESRGAN Synthesis and DnCNN

Table K. SRCC comparison on different numbers of reference images and distortion types.

Distortion \ # Reference image	1,000	500	100
Full 25 types	**0.776**	0.766	0.739
10 types with top-10 ratios	0.770	0.759	0.735
10 types with bottom-10 ratios	0.743	0.736	0.719

Synthesis. We choose the 6 distortion types out of a total of 25 for analysis, *i.e*., DnCNN denoising algorithm, Gaussian blur, additive white Gaussian noise, color oversaturation, color blocking and sharpness. As shown in Table J, each distortion type contains 2,000 distorted images. The three types of distortion, *i.e*., DnCNN denoising algorithm, Gaussian blur and additive white Gaussian noise, are present on both PIPAL and KADID-10k Synthesis and are therefore heavily selected as positive unlabeled data by the classifier for semi-supervised learning of IQA models. In contrast, the other three types of distortion are unseen for PIPAL, and the corresponding distortion images differ significantly from the distribution of the labeled data in PIPAL, which are excluded by the classifier. Furthermore, we find that the 4 outliers in the DnCNN denoising algorithm or Gaussian blur settings are synthesized based on the same two reference images, as shown in Fig. A. We consider the reason is that those two reference images are over-smooth or monochromatic, which lack real-world textures and not meet the requirement for reference images. In summary, the proposed JSPL is leveraged to identify negative samples from unlabeled data, *e.g*., reference images that lack real-world textures or distorted images that differ significantly from the labeled data.

More discussion on how much unlabeled data and number of distortions. We use the PIPAL training set as labeled set, and use several representative distortion models to synthesize unlabeled samples. Specifically, there are total 25 distortion types in KADID-10k and 1,000 reference images. Based on the trained classifier, the ratios $\rho = \frac{\text{positive unlabeled samples}}{\text{outliers}}$ can be computed for 25 distortion types. In Table J, distortion types with top-3 and bottom-3 ratios are presented. Taking KADID-10k as testing bed, we discuss the sensitivity of our JSPL with different numbers of unlabeled samples and distortion types. As for the number of reference images, we set it as 1,000, 500 and 100. As for distortions, we adopt three settings, *i.e*., full 25 types, 10 types with top-10 ρ ratios and 10 types with bottom-10 ρ ratios. The results are summarized in Table K. We can observe that: (i) Benefiting from unlabeled samples, our JSPL contributes to performance gains for any setting, *i.e*., the models in Table K are all superior to the model trained on only labeled data (SRCC = 0.717 by Our(SL) in Table B). (ii) When reducing the number of reference images from 1,000 to 500, our JSPL slightly degrades for all the three distortion settings. And it is reasonable that the performance of JSPL is close to Our(SL) when few unlabeled samples are exploited. (iii) As for distortions, the IQA models with bottom-10 ρ ratios are notably inferior to Our(JSPL), indicating that JSPL can well exclude outliers.

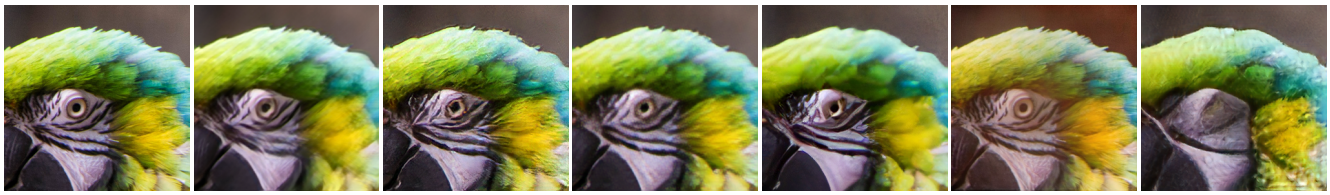
F. More Details on IQA Datasets

Details of the different IQA datasets containing the distortion types can be viewed in Table L. Among them, the KADID-10k contains the richest traditional distortion types and the PIAPL contains the richest distortion types of the recovery results.

As shown in Fig. B, we take an example image from validation set of PIPAL to visually show the consistency between various methods and subjective perception, inlcuding PSNR, SSIM [58], MS-SSIM [61], LPIPS [73], IQT [9] and our method. One can see that the proposed FR-IQA with JSPL achieves the closest rank agreement with the human annotated MOS.

Table L. Descriptions of the five IQA databases.

Database	# Ref.	# Dis.	Distortion Types
TID2013 [45]	25	3,000	(1) Additive Gaussian noise; (2) Additive noise in color components; (3) Spatially correlated noise; (4) Masked noise; (5) High frequency noise; (6) Impulse noise; (7) Quantization noise; (8) Gaussian blur; (9) Image denoising; (10) JPEG compression; (11) JPEG2000 compression; (12) JPEG transmission errors; (13) JPEG2000 transmission errors; (14) Non eccentricity pattern noise; (15) Local block-wise distortions of different intensity; (16) Mean shift (intensity shift); (17) Contrast change; (18) Change of color saturation; (19) Multiplicative Gaussian noise; (20) Comfort noise; (21) Lossy compression of noisy images; (22) Image color quantization with dither; (23) Chromatic aberrations; (24) Sparse sampling and reconstruction
LIVE [47]	29	982	(1) JPEG compression; (2) JPEG2000 compression; (3) Additive white Gaussian noise; (4) Gaussian blur; (5) Rayleigh fast-fading channel distortion
CSIQ [33]	30	866	(1) JPEG compression; (2) JP2K compression; (3) Gaussian blur; (4) Gaussian white noise; (5) Gaussian pink noise; (6) Contrast change
KADID-10k [35]	81	10,125	(1) Gaussian blur; (2) Lens blur; (3) Motion blur; (4) Color diffusion; (5) Color shifting; (6) Color quantization; (7) Color over-saturation; (8) Color desaturation; (9) JPEG compression; (10) JP2K compression; (11) Additive white Gaussian noise; (12) White with color noise; (13) Impulse noise; (14) Multiplicative white noise; (15) DnCNN denoising algorithm; (16) Brightness changes; (17) Darken; (18) Shifting the mean; (19) Jitter spatial distortions; (20) Non-eccentricity patch; (21) Pixelate; (22) Quantization; (23) Color blocking; (24) Sharpness; (25) Contrast
PIPAL [19]	250	25,850	(1) Median filter denoising; (2) Linear motion blur; (3) JPEG and JPEG 2000; (4) Color quantization; (5) Gaussian noise; (6) Gaussian blur; (7) Bilateral filtering; (8) Spatial warping; (9) Comfort noise; (10) Interpolation; (11) A+; (12) YY; (13) TSG; (14) YWHM; (15) SRCNN; (16) FSRCNN; (17) VDSR; (18) EDSR; (19) RCAN; (20) SFTMD; (21) EnhanceNet; (22) SRGAN; (23) SFTGAN; (24) ESRGAN; (25) BOE; (26) EPSR; (27) PESR; (28) EUSR; (29) MCML; (30) RankSRGAN; (31) DnCNN; (32) FFDNet; (33) TWSC; (34) BM3D; (35) ARCNN; (36) BM3D + EDSR; (37) DnCNN + EDSR; (38) ARCNN + EDSR; (39) noise + EDSR; (40) noise + ESRGAN;



	Ref.	Dis.1	Dis.2	Dis.3	Dis.4	Dis.5	Dis.6
MOS↑		1359.45(1)	1327.90(2)	1261.15(3)	1213.73(4)	1206.27(5)	868.30(6)
PSNR↑		24.18(2)	22.99(4)	26.32(1)	23.61(3)	20.67(5)	19.91(6)
SSIM↑		0.679(3)	0.572(5)	0.720(2)	0.620(4)	0.863(1)	0.450(4)
MS-SSIM↑		0.893(3)	0.882(5)	0.934(2)	0.883(4)	0.938(1)	0.703(6)
LPIPS↓		0.198(4)	0.161(2)	0.174(3)	0.252(5)	0.110(1)	0.327(6)
IQT↑		1364.39(1)	1327.20(3)	1135.62(2)	1282.94(5)	1316.89(4)	1069.47(6)
Ours(SL)↑		0.765(1)	0.757(3)	0.758(2)	0.734(5)	0.752(4)	0.689(6)
Ours(JSPL) ↑		0.765(1)	0.759(2)	0.756(3)	0.736(5)	0.754(4)	0.688(6)

Figure B. An evaluation example from validation set of PIPAL. The quality is measured by MOS and 7 IQA methods. The numbers in brackets indicate the ranking of the corresponding distortion image.