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

Yue Cao\textsuperscript{1}, Zhaolin Wan\textsuperscript{2}, Dongwei Ren\textsuperscript{1(✉)}, Zifei Yan\textsuperscript{1}, Wangmeng Zuo\textsuperscript{1,3}

\textsuperscript{1}School of Computer Science and Technology, Harbin Institute of Technology, China
\textsuperscript{2}College of Artificial Intelligence, Dalian Maritime University, China
\textsuperscript{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)\textsuperscript{[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.

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\textsuperscript{[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

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.
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 pretrained 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 as a new IQA network for leveraging unlabeled data and alleviating the spatial misalignment issue.

3. Proposed Method

3.1. Problem Setting

Denote by $x = (I_{Ref}, I_{Dis})$ a two-tuple of pristine-quality reference image $I_{Ref}$ and distorted image $I_{Dis}$, and $y$ the ground-truth MOS. Learning-based FR-IQA aims to find a mapping $f(x)$ parameterized by $\Theta$ to predict the quality score $y$ 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 $P = \{x_i, y_i\}_{i=1}^{N_p}$ denote the positive labeled data and $U = \{x_j\}_{j=1}^{N_u}$ 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 $f(x)$, our JSPL also learns a binary classifier $h(x_j)$ parameterized by $\Theta^b$ for determining an unlabeled two-tuple is a negative (i.e., outlier) or a positive sample.
3.2. JSPL Model

A joint semi-supervised and PU learning (JSPL) model is presented to learn IQA model \( f(x) \) and binary classifier \( h(x) \) from the labeled data \( \mathbb{P} \) and the unlabeled data \( \mathbb{U} \). Particularly, PU learning is utilized to learn \( h(x) \) for identifying positive unlabeled samples. And SSL is used to learn \( f(x) \) 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 \( h(x) \), we treat all samples in \( \mathbb{P} \) as positive samples, and all samples in \( \mathbb{U} \) as unlabeled samples. For a positive sample \( x_i \), we simply adopt the cross-entropy (CE) loss,

\[
CE(h(x_i)) = -\log h(x_i). \quad (1)
\]

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

\[
H(h(x_j)) = -h(x_j) \log h(x_j) - (1-h(x_j)) \log (1-h(x_j)). \quad (2)
\]

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

\[
NE(\mathbb{B}_u) = -\log (1 - \min_{x_j \in \mathbb{B}_u} h(x_j)) . \quad (3)
\]

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

\[
L_{PU} = \sum_i CE(h(x_i)) + \sum_j H(h(x_j)) + \sum_{\mathbb{B}_u} NE(\mathbb{B}_u). \quad (4)
\]

**SSL.** FR-IQA is a regression problem. For labeled sample \( x_i \) with ground-truth MOS \( y_i \), we adopt the mean squared error (MSE) loss defined as,

\[
\ell(f(x_i), y_i) = \| f(x_i) - y_i \|^2 . \quad (5)
\]

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

\[
\ell(f(x_j), y_j^*) = \| f(x_j) - y_j^* \|^2 , \quad (6)
\]

where \( y_j^* \) denotes the pseudo MOS for \( x_j \). 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,

\[
y_j^*(t) = \alpha \cdot y_j^*(t-1) + (1-\alpha) \cdot f(t)(x_j) , \quad (7)
\]

where \( \alpha \) (\(=0.95\)) is the momentum. \( y_j^*(t) \) denotes the pseudo MOS after \( t \) iterations of training, and \( f(t)(x_j) \) denotes the network output after \( t \) iterations of training. Therefore, we define the SSL loss as,

\[
L_{SSL} = \sum_i \ell(f(x_i), y_i) + \sum_j \mathbb{I}_{h(x_j) \geq \tau} \ell(f(x_j), y_j^*) . \quad (8)
\]

\( \mathbb{I}_{h(x_j) \geq \tau} \) is an indicator function, where it is 1 if \( h(x_j) \geq \tau \) 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,

\[
\min_{\Theta, \theta_h} \mathcal{L} = \mathcal{L}_{SSL} + \mathcal{L}_{PU} . \quad (9)
\]

We note that our JSPL is a joint learning model, where both the FR-IQA network \( f(x) \) and binary classifier \( h(x) \) 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 \( \mathcal{L}_{SSL} \).
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 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 $L_2$ to avoid aliasing when down-sampling 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 $I_{Ref}$ and distorted image $I_{Dis}$ are fed into the feature extraction network to obtain the reference feature $f_{Ref}^s$ and distortion feature $f_{Dis}^s \ (s = 1, 2, 3)$, respectively. Then, local sliced Wasserstein (LocalSW) distance is presented to produce distance map $f_{Dist}$, and a spatial attention module is deployed for reweighting distance map to generate calibrated difference map $f_{Diff}^s$ for each scale $s$. As shown in Fig. 2, the score prediction network has three branches, where each branch involves two $1 \times 1$ convolutional layers and a spatial-wise global average pooling layer. $f_{Diff}^s$ 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 $f_{Ref}^s$ and distortion feature $f_{Dis}^s$, one direct solution is the element-wise difference, i.e., $|f_{Ref}^s - f_{Dis}^s|$. 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 misalign-

![Diagram](Image)

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.
and local average pooling is deployed to generate spatial
weighting map \( f_W \in \mathbb{R}^{H \times W} \), where the size of the local
average pooling region is set to \( p \times p \). Calibrated difference
map \( f_{\text{Diff}} \) can then be obtained by using \( f_W \) for reweighting
each channel of distance map \( f_{\text{Dist}} \) in an element-wise
manner, while final score can be predicted by feeding \( f_{\text{Diff}} \)
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

4.1.1. 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 dis-
torted 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 parti-
tion 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 \( \times \) 288. The validation set
consists of 25 reference images and 1,000 distorted images.
Since the testing set of PIPAL is not publicly avail-
able, we in this paper report the evaluation results on vali-
dation set via the online server\(^1\). The distorted images in PI-
PAL dataset include traditional distorted images and images
restored by multiple types of image restoration algorithms
(e.g., denoising, super-resolution, deblurring, 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 \( \times \) 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 im-
ages 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 lev-
els 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 re-
ported 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 \( \times \) 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 \( 1e^{-5} \) af-
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
model. Hyper-parameter \( p_i \), i.e., the region size in local Sliced Wasserstein distance (LocalSW), is set to \( \delta \) and 2 for PIPAL and traditional IQA datasets, respectively. The momentum parameter \( \alpha \) is set to 0.95. Hyperparameter \( \tau \) changes with iterations, i.e., \( \tau = \max\{\tau_0, \tau_\min\} \) for \( t \)-th iteration, where parameters \( \tau_0, \tau_0 \), 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.

### 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], WaDlQaM-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 unlabeled data while mitigating the adverse effects.
### 4.4.2 Evaluation on Traditional Datasets

Our methods with two learning manners, \textit{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, it 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.

### 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, \textit{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


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


