

# 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  $128 \times 128$  and  $64 \times 64$  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 ( $288 \times 288$ ) 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

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

Database Type	LIVE				
	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

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 regarded as the corresponding distorted images in unlabeled data.

### 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 1<sup>st</sup> and 2<sup>nd</sup> 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	<b>0.880</b>	0.796	<b>0.718</b>
our(SL)	PIPAL	<b>0.919</b>	0.873	<b>0.804</b>	0.717
our(JSPL)	PIPAL & KADID-10k Synthesis	<b>0.930</b>	<b>0.894</b>	<b>0.812</b>	<b>0.776</b>

Table C. SRCC comparisons on different cross-dataset with the KADID10k as training set. **Red** and **blue** are utilized to indicate top 1<sup>st</sup> and 2<sup>nd</sup> 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	<b>0.973</b>	<b>0.951</b>	<b>0.908</b>	<b>0.770</b>
our(JSPL)	KADID-10k & KADID-10k Synthesis	<b>0.974</b>	<b>0.953</b>	<b>0.910</b>	-
our(JSPL)	KADID-10k & ESRGAN Synthesis	-	-	-	<b>0.801</b>

Table D. SRCC comparisons on different cross-dataset with the TID2013 as training set. **Red** and **blue** are utilized to indicate top 1<sup>st</sup> and 2<sup>nd</sup> 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	<b>0.775</b>	0.639
our(SL)	TID2013	<b>0.944</b>	<b>0.932</b>	0.762	<b>0.651</b>
our(JSPL)	TID2013 & KADID-10k Synthesis	<b>0.948</b>	<b>0.934</b>	<b>0.795</b>	-
our(JSPL)	TID2013 & ESRGAN Synthesis	-	-	-	<b>0.699</b>

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 in-

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

Based on	PIPAL Val.
Reference feature $f_{Ref}^s$	<b>0.868 / 0.868</b>
Distortion feature $f_{Dis}^s$	0.861 / 0.860
Distance map $f_{Dist}^s$	0.864 / 0.864

Table F. Performance on different attention mechanism on PIPAL.

Attention Mechanism		SRCC
Spatial	Channel	
✗	✗	0.857
✓	✗	<b>0.868</b>
✗	✓	0.840
✓	✓	0.859

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

$\tau_{min}$	PIPAL	KADID-10k
	PLCC / SRCC	PLCC / SRCC
0.4	0.872 / 0.870	0.951 / 0.949
0.5	<b>0.877 / 0.874</b>	<b>0.963 / 0.961</b>
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
	$p = 16$	0.862
	$p = 8$	<b>0.868</b>
	$p = 4$	0.866
	$p = 2$	0.864
	$p = 1$	0.857
		0.940

formative 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 I. PLCC / SRCC comparisons on different FR-IQA with SL or JSPL training on PIPAL. Red and blue are utilized to indicate top 1<sup>st</sup> and 2<sup>nd</sup> 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  $f_{Ref}^s$ , the distortion feature  $f_{Dis}^s$  and the distance map  $f_{Dist}^s$ . 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 IWSSIM [60], spatially local information is suggested as one key factor for assessing distortions, which motivates us to only adopt spatial attention.

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

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

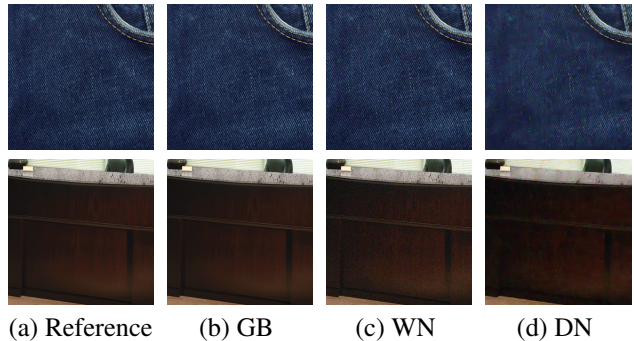


Figure A. Visualization of the excluded outliers, *i.e.*, the corresponding reference images, DnCNN denoising (DN) distorted images, Gaussian blur (GB) distorted images and additive white Gaussian noise (WN) distorted images.

ner 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.

## 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.

# Reference image	1,000	500	100
Distortion			
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 over-saturation, 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  $\rho$  ratios and 10 types with bottom-10  $\rho$  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  $\rho$  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, including 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;

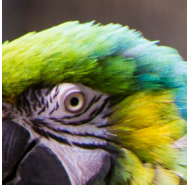
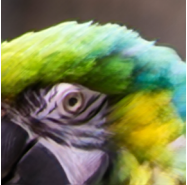
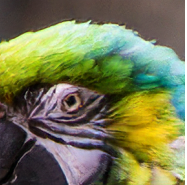
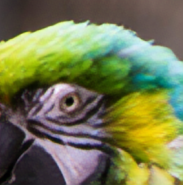

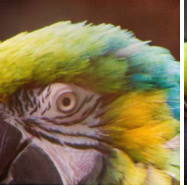

						
Ref.	Dis.1	Dis.2	Dis.3	Dis.4	Dis.5	Dis.6
MOS↑	1359.45(1)	1327.90(2)	1261.15(3)	1213.73(4)	1206.27(5)	868.30(6)
PSNR↑	24.18(2)	22.99(4)	26.32(1)	23.61(3)	20.67(5)	19.91(6)
SSIM↑	0.679(3)	0.572(5)	0.720(2)	0.620(4)	0.863(1)	0.450(4)
MS-SSIM↑	0.893(3)	0.882(5)	0.934(2)	0.883(4)	0.938(1)	0.703(6)
LPIPS↓	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.