Region-Adaptive Deformable Network for Image Quality Assessment

Shuwei Shi[1*] Qingyan Bai[1*] Mingdeng Cao[2] Weihao Xia[1]
Jiahao Wang[2] Yifan Chen[1] Yujiu Yang[1†]
[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]Department of Automation, Tsinghua University
{ssw20, bqy20, cmd19, wang-jh19, chenyf20}@mails.tsinghua.edu.cn
xiawh3@outlook.com yang.yujiu@sz.tsinghua.edu.cn

Abstract

Image quality assessment (IQA) aims to assess the perceptual quality of images. The outputs of the IQA algorithms are expected to be consistent with human subjective perception. In image restoration and enhancement tasks, images generated by generative adversarial networks (GAN) can achieve better visual performance than traditional CNN-generated images, although they have spatial shift and texture noise. Unfortunately, the existing IQA methods have unsatisfactory performance on the GAN-based distortion partially because of their low tolerance to spatial misalignment. To this end, we propose the reference-oriented deformable convolution, which can improve the performance of an IQA network on GAN-based distortion by adaptively considering this misalignment. We further propose a patch-level attention module to enhance the interaction among different patch regions, which are processed independently in previous patch-based methods. The modified residual block is also proposed by applying modifications to the classic residual block to construct a patch-region-based baseline called WResNet. Equipping this baseline with the two proposed modules, we further propose Region-Adaptive Deformable Network (RADN). The experiment results on the NTIRE 2021 Perceptual Image Quality Assessment Challenge dataset show the superior performance of RADN, and the ensemble approach won fourth place in the final testing phase of the challenge.

1. Introduction

Image quality assessment tasks have gained increasing research attention for decades, and its goal is to assess image perceptual quality like humans. In the past decades, researchers have proposed some IQA algorithms based on deep learning [2, 16, 8, 33]. Although these IQA methods can maintain consistency with human subjective evaluation to some extent, they still show limitations in evaluating the results of image restoration and image super-resolution. As introduced in [8], some GAN-based image restoration (IR) algorithms usually produce fake textures and other details. However, the existing algorithms cannot distinguish GAN-generated image textures from noises and natural details,

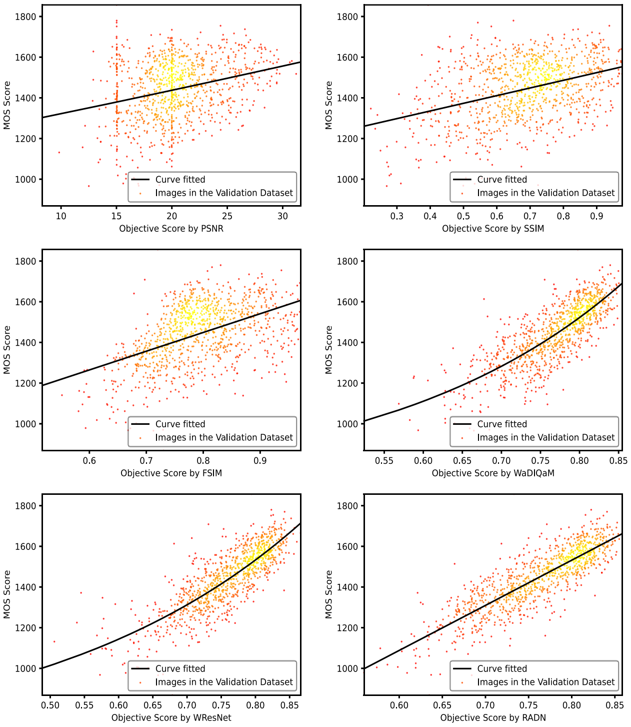


Figure 1. The scatter plots of various IQA models on the NTIRE 2021 Perceptual Image Quality Assessment Challenge validation dataset, which show the relationship between the predicting scores and the MOS labels.

[*] The first two authors contribute equally to this work.

[†] This work was supported partially the Major Research Plan of National Natural Science Foundation of China (Grant No. 61991450) and the Shenzhen Science and Technology Project under Grant (ZDYBH201900000002, JCYJ20180508152042002).

which deteriorates the performance of existing IQA algorithms. To deal with GAN-based distortion, Gu *et al.* [8] proposed a novel IQA benchmark characterized by including a proportioned GAN-based distortion dataset, and most previously proposed IQA methods have shown unsatisfying performance on the dataset (see Fig. 1). They also propose a Space Warping Difference (SWD) layer to compare the features on a small range around the corresponding position. The operation is robust to spatial shifts. We observe that although the method has a specific effect, it cannot be used in all scenarios because the range of the field defined in [8] is a hyper-parameter that varies in different distortion scenarios. Therefore, it is limited to specific circumstances and is not general and flexible enough.

Considering the drawbacks above, we propose the Region-Adaptive Deformable Network (RADN). The proposed method consists of three components: modified residual block, patch-level attention block, and reference-oriented deformable convolution. We first revisit the classic residual blocks from IQA tasks and propose the modified residual block. Some modifications of the classical residual block are made to adapt the characteristics of IQA tasks, including removing the Batch Normalization (BN) layers, employing only 3×3 convolutional layers, and adjusting the numbers of the convolutional layers. We then build our baseline WResNet using the modified residual blocks. Experiments in Sec. 4.6 show that the WResNet (without the other two modules) has already outperformed WaDIQaM [2] in both metric performance and convergence.

For adaptation to images of significant differences, we propose a novel module dubbed reference-oriented deformable convolution [4] which can select the region of interest adaptively according to the shape of the object. Furthermore, it changes adaptively with the size of the target and selects critical information around it. We believe that the offset predicted in the reference image can capture the object's actual shape, similar to the result observed by humans, which cannot be affected by the GAN-based distortion. Applying such deformable convolution to the distorted images can make the reference images interact with the distorted images and be robust to the GAN-based distortion.

To boost the information interactions among the patch regions, we further propose a patch-level attention mechanism. Despite good performance on synthetically distorted images of the patch-based algorithms, it may destroy both the high-level semantic information of the image and the relationship among patches. In other words, dividing a complete image into patches can affect the performance of the IQA methods. To alleviate this problem, we propose patch-level attention. The critical assumption is that the characteristics of an image patch depend not only on itself but also on other image patches. Therefore, we introduce patch-level attention after feature extraction of reference and distorted images to capture dependencies by computing interactions between two patch regions, respectively. This operation will strengthen the information interaction among different patches to obtain more accurate feature expressions for the two images. This plug-and-play module can be incorporated into any patch-based IQA method to improve performance.

Experiments on the newly-proposed PIPAL dataset and cross-dataset evaluation on TID2013 and LIVE show the competitive performance of our method on the above datasets. Our method ranked fourth in the NTIRE 2021 Perceptual Image Quality Assessment Challenge (NTIRE 2021 IQA Challenge) [10].

2. Related Work

Image Quality Assessment. The IQA algorithms are used to evaluate the quality of images that may be degraded during transmission, compression, and algorithm processing. Researchers have worked hard to develop general quality assessment algorithms close to human subjective evaluation in the past decades. According to different scenarios, IQA algorithms can be divided into full-reference (FR-IQA) and no-reference methods (NR-IQA). FR-IQA methods commonly include SSIM [30], MS-SSIM [31] and PSNR, *etc.* Inspired by them, FSIM [39], SR-SIM [36], and GMSD [34] are proposed. These hand-crafted methods assess the image quality by comparing the feature difference between the distorted image and the reference image. Recently, the deep learning-based FR-IQA methods [20, 2] get superior prediction performance over hand-crafted methods. Apart from FR-IQA methods, NR-IQA methods are developed to assess image quality without the reference image. Liu *et al.* [16] used comparative learning to assess image quality. In [41], meta-learning is introduced in IQA to learn meta-knowledge shared with humans. In some recent works, IQA methods are applied to measure image restoration algorithms [1]. Researchers hope to improve the performance of image restoration algorithms by developing better IQA methods. Gu *et al.* [8] proposed a new dataset named PIPAL. They test several existing algorithms to demonstrate that these algorithms' low tolerance toward spatial misalignment may be a key reason for their dropping performance. Unlike these approaches, we adopted patch-level attention and reference-oriented deformable convolution in our model to handle the GAN-based distorted images in PIPAL.

Deformable Convolution. Dai *et al.* [4] first proposed deformable convolution and proved that it is effective for sophisticated vision tasks such as object detection [42] and semantic segmentation [4]. Offsets learned in deformable convolution blocks can obtain the information in the light of the object's shape, improving the capability of the fea-

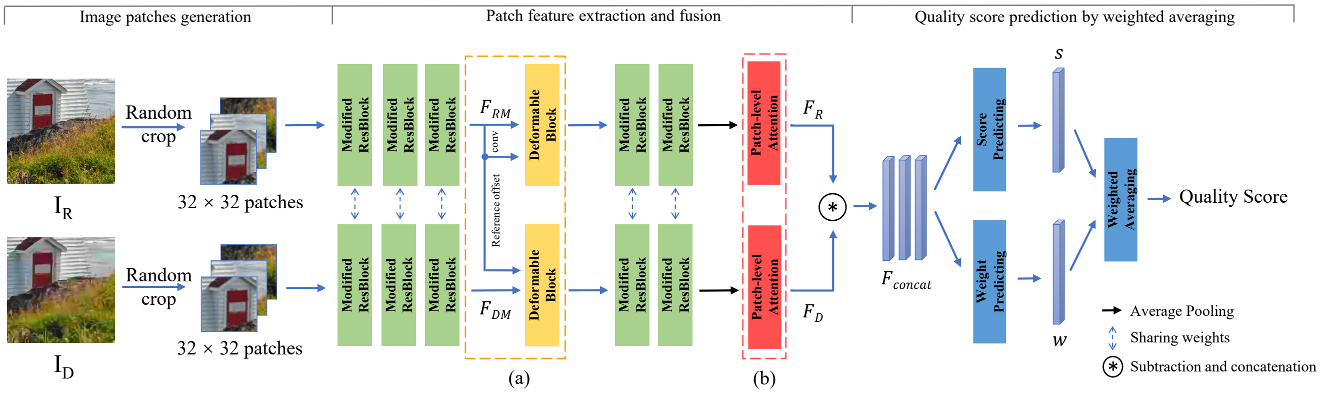


Figure 2. The architecture of the proposed approach - Region Adaptive Deformable Network (RADN). A distorted image I_D and its corresponding reference image I_R are randomly cropped into 32×32 sized patches. Then the patches are inputted into the feature extracting module, and the final quality scores are obtained by weighted averaging. The yellow dashed box (a) and the red one (b) indicate our reference-oriented deformable convolution and patch-level attention module, respectively, which can be found in Sec. 3.3 and Sec. 3.4. Our baseline model WResNet can be obtained by removing the two proposed modules (a) and (b). The details of our modified residual block can be found in Sec. 3.1.

ture extraction. It also performs well in low-level vision tasks such as video super-resolution [24] and video deblurring [26]. However, it has not been introduced in the IQA task. Inspired by the methods mentioned above, we adopt reference-oriented deformable convolution for FR-IQA.

Attention Mechanism. Attention mechanisms have been widely used in various tasks [14, 15, 25, 32]. For instance, in NR-IQA, Yang *et al.* [35] proposed an end-to-end saliency-guided architecture and applied spatial and channel attention in their model. Their method got a good performance in the NR-IQA task. Non-local operations [27] compute the response at a position as a weighted sum of the features at all positions for capturing long-range dependencies. Motivated by these methods, we proposed patch-level attention to capture dependencies between any two patches of one image to obtain more accurate feature maps from both reference and distorted images.

3. Proposed Method

The structure of the proposed Region Adaptive Deformable Network (RADN) is shown in Fig. 2. For a pair of images, we first crop the reference image I_R and the distorted image I_D into patches with spatial size 32 × 32. During feature extraction, first, the modified residual blocks (green) are proposed and employed. Then with the intermediate feature maps of the reference and the distorted image (*i.e.*, F_{RM} and F_{DM}), we use the proposed reference-oriented deformable convolution (yellow) to select the region of interest and capture the object's actual shape adaptively. Next, the patch-level attention module (red) is employed to boost the interaction among the image patches. More details of the two proposed modules can refer to the Sec. 3.3 and Sec. 3.4. The final feature maps (*i.e.*, F_D and F_R) and their difference ($F_{diff} = F_D - F_R$) are then concatenated in the channel dimension to serve as a new feature, *i.e.*$F_{concat} = concat(F_{diff}, F_D, F_R)$. Finally as in Eq.1:

$$\hat{q} = \frac{\sum_{0<i<N_{patch}} w_i \times s_i}{\sum_{0<i<N_{patch}} w_i} \tag{1}$$

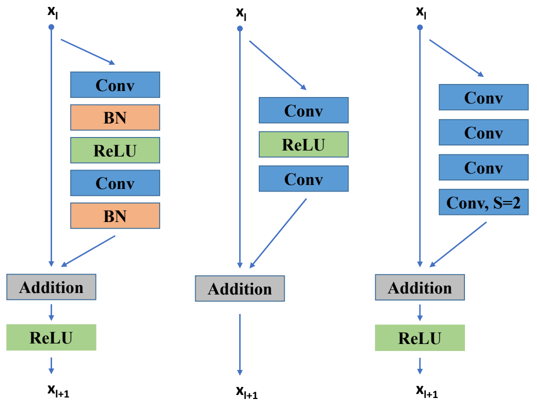
The combined feature F_{concat} will be sent to the fully-connected layers to predict the weight w_i and the score s_i of per patch and get the final quality score $\hat{q}$ by weighted averaging and N_{patch} in the equation means the amount of the patches for one image.

Besides the aforementioned modules, we also propose a contrastive pretraining strategy to further improve the model's ability to distinguish the image quality rather than direct regression of the quality score.

3.1. Modified Residual Block

Considering the characteristics of IQA tasks and the PIPAL dataset [8], we improve the performance of the classic residual block with a more reasonable structure (shown in Fig. 3). The low-quality images vary greatly in content and distortion types. As batch normalization operations will result in over-smoothness of special features in different samples [12, 28] which makes the model performance degrade greatly, we remove all the batch- normalization layers in the original residual block.

Furthermore, we adopt the 3 × 3 convolution without any 7 × 7 employed in the original ResNet. The 3 × 3 convolution has been proven to be more hardware-friendly [7] and more effective than the 7 × 7 convolution layer followed by



(a) Original (b) EDSR (c) Proposed

Figure 3. The architectures of various residual blocks. Note that we use 2-stride convolution layers of the residual blocks to perform down-sampling operations.

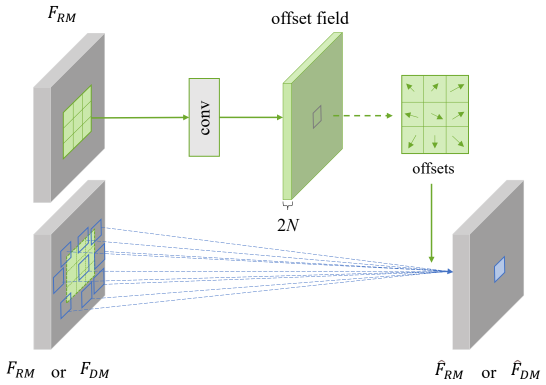


Figure 4. Reference-oriented deformable convolution. F_{RM} and F_{DM} indicate the feature map of the reference image and the distorted image respectively. The number of offsets is defined as: $N = |\mathbf{p}_k|$.

a pooling layer in the earliest feature extraction stage. We only add a shortcut every four convolutional layers, which can generate more complex representations and show better performance in our experiment than the two convolution layers adopted in the original residual block.

3.2. WResNet

We use the modified residual blocks to build our baseline method named WResNet (W stands for weighted averaging). The architecture of our baseline is shown in Fig. 2, which can be obtained by removing the two proposed modules (a) and (b). For WResNet, we apply the l_2 loss function to regress the quality scores during training.

3.3. Reference-Oriented Deformable Convolution

Considering that humans are less sensitive to the error and misalignment of the edges in distorted images generated by GAN, Gu *et al.* [8] proposed SWD to deal with the GAN-based distortion. In fact, the fixed design limits its adaptation to different GAN-based distortion. Therefore, we adopt the deformable convolution module in our residual blocks to adapt to the mismatched regions of GAN-based distortion between reference and distorted images. To make better use of the reference information, different from the original deformable convolution, we introduce reference-oriented one shown in Fig. 4, where F_{DM} and F_{RM} are the intermediate distorted and reference feature maps (see Fig. 2). For conventional 2D convolution with a kernel of N sampling locations, it first samples the values from the regular offsets $\mathbf{p}_k, k \in 1, 2, \cdots, N$, then sums the sampled values weighted by W_k corresponding to the kth location. For example, a 3×3 kernel has 9 regular sampling locations which are defined as $\mathbf{p}_k \in \{(-1,-1), (-1,0), \cdots, (1,1)\}$. In terms of our reference-oriented deformable convolution, we first generate the offsets $\Delta\mathbf{p}_k$ according to kth sampling location from reference feature maps:

$$\Delta\mathbf{p}_k = f(\mathbf{F}_{RM}) \tag{2}$$

where f means a 3×3 convolution, the output channel number of which is $2N$. Then the learned offsets $\Delta\mathbf{p}_k$ are used for sampling the values in both reference and distorted feature maps, rather than the regular sampling with $\mathbf{p}_k$. For each location $\mathbf{p}_0$ on the output reference and distorted feature maps $\hat{F}_{RM}$ and $\hat{F}_{DM}$, the value in it is aggregated by the following process:

$$\begin{aligned}\hat{\mathbf{F}}_{RM}(\mathbf{p}_0) &= \sum_{k=1}^{N} w_k^r \cdot \mathbf{F}_{RM}(\mathbf{p}_0 + \mathbf{p}_k + \Delta\mathbf{p}_k), \\ \hat{\mathbf{F}}_{DM}(\mathbf{p}_0) &= \sum_{k=1}^{N} w_k^d \cdot \mathbf{F}_{DM}(\mathbf{p}_0 + \mathbf{p}_k + \Delta\mathbf{p}_k),\end{aligned} \tag{3}$$

where w^r and w^d are the convolution weights for reference and distortion feature maps. Owing to the deformable sampling locations $\mathbf{p}_0 + \mathbf{p}_k + \Delta\mathbf{p}_k$ are fractional, bilinear interpolation is applied [4] to sample the values.

With the application of this reference-oriented deformable convolution in both reference and distorted branches (shown in Fig. 2), our model can deal with the GAN-based distortion better and learn the spatial shift-invariant features from the paired images adaptively. We demonstrate the effectiveness of reference-oriented deformable convolution in Sec. 4.6.

3.4. Patch-Level Attention

The previous patch-level-based IQA methods predict the quality of each patch individually, which leads to a specific

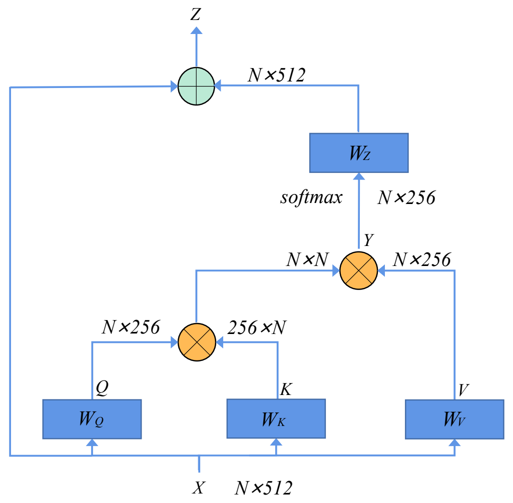


Figure 5. The patch-level attention block. The feature maps are shown as the shape of their tensors, and N is the number of patches.

performance drop due to the lack of interaction among the patch regions. We believe the quality of a patch not only depends on its own feature but also affected by other patches in the same image. Recently, the self-attention mechanism exhibits excellent relation modeling in computer vision tasks from low-level to high-level tasks. However, how to introduce it into the FR-IQA task is a challenge to be explored. To this end, we introduce a self-attention module to handle the feature maps of reference and distorted images to boost the interaction between any two patches in one image, as shown in the red dashed box in Fig. 2. Different from the non-local block proposed in [27], we compute the response at a patch as a weighted sum of the features at all patches in one image, namely patch-level attention.

Since dot-product attention is much faster and more space-efficient in practice, we use it scaled by $\frac{1}{\sqrt{d_k}}$ in our patch-level attention block. Attention block generates corresponding **Q**ueries, **K**eys and **V**alues of dimension d_k by performing linear projection on the patch-level feature vectors X with W_Q, W_K, W_V:

$$Q = W_Q X, \quad K = W_K X, \quad V = W_V X \tag{4}$$

where X is reshaped from $N \times 512 \times 1 \times 1$ to $N \times 512$ (N is the number of patches) before it is input into the patch-level attention block. Attention operations are defined as following:

$$Y = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{5}$$

We involve the attention operation described in Eq. 5 into a patch-level attention block that can be integrated into our model. It is defined as:

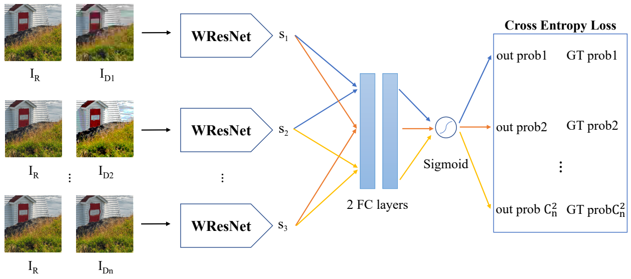


Figure 6. The architecture of our efficient contrastive model. Arrows with the same color stand for a contrast pair and its result. Every two scores are compared as a contrast pair so if there are n pairs of reference and distorted images, there will be C_n^2 contrast pairs and preference probabilities.

$$Z = W_Z Y + X \tag{6}$$

where Y is given in Eq. 5, W_Z is a weight matrix to be learned and Z is computed by a residual connection. An example of patch-level attention block is shown in Fig. 5.

Through the processing of the patch-level attention block, the patches in the same image can be better interacted with each other and have more accurate feature representation.

3.5. Contrastive Pretraining Strategy

Contrastive training is an acceptable way to take advantage of the labels from the side, given that the MOS labels are obtained by manually comparing the image pairs. Most current IQA models tend to use l_1 and l_2 to regress the quality score, which merely concentrates on the accuracy of the values and ignores the ranking relationships between the samples. The contrastive models [8, 16] can alleviate this problem by comparing the samples. Based on SWDN [8], we propose a contrastive pretraining strategy to make the model learn how to distinguish the image quality rather than directly regress the quality score. The difference is that our strategy is to pre-train the model by comparing the labels to get the preference probabilities and then make use of the MOS labels directly by l_2 regression. Siamese Network is an indispensable part of contrastive learning, and here we use our WResNet proposed in Sec. 3.2. As in Fig. 6, given a set of n quality scores $\{s_1, s_2, \cdots, s_n\}$ obtained by the Siamese Network, we make a comparison between every two samples (s_i, s_j) using the fully-connected layers $f(s_i, s_j)$ to get the preference probabilities. Then we apply the cross-entropy loss to regress them with the ground truth preferring probability p_{ij}. The final cross-entropy loss is shown as follows:

Table 1. Performance of different methods on the NTIRE 2021 IQA Challenge validation and testing datasets. The results of our ensemble model is bolded.

Method	Validation		Test	
	SROCC	PLCC	SROCC	PLCC
PSNR	0.2548	0.2917	0.2493	0.2769
NQM [5]	0.3458	0.4164	0.3644	0.3954
UQI [29]	0.4859	0.5476	0.4195	0.4500
SSIM [30]	0.3400	0.3984	0.3614	0.3936
MS-SSIM [31]	0.4864	0.5633	0.4618	0.5007
IFC [22]	0.5936	0.6767	0.4851	0.5549
VIF [21]	0.4335	0.5236	0.3970	0.4795
VSNR [3]	0.3213	0.3750	0.3682	0.4107
RFSIM [38]	0.2656	0.3045	0.3037	0.3284
GSM [13]	0.4181	0.4688	0.4094	0.4646
SRSIM [36]	0.5658	0.6541	0.5728	0.6360
FSIM [39]	0.4672	0.5606	0.5038	0.5709
FSIMc [39]	0.4679	0.5587	0.5057	0.5727
VSI [37]	0.4501	0.5162	0.4584	0.5169
MAD [11]	0.6078	0.6263	0.5434	0.5804
NIQE [18]	0.0644	0.1018	0.0341	0.1317
MA [17]	0.2006	0.2034	0.1405	0.1469
PI [1]	0.1690	0.1662	0.1036	0.1454
LPIPS-Alex [40]	0.6276	0.6463	0.5658	0.5711
LPIPS-VGG [40]	0.5915	0.6471	0.5947	0.6331
PieAPP [20]	0.7063	0.6972	0.6074	0.5974
WaDIQaM [2]	0.6779	0.6543	0.5533	0.5480
DISTS [6]	0.6743	0.6860	0.6548	0.6873
SWD [8]	0.6611	0.6680	0.6243	0.6342
WResNet-Classic	0.4881	0.4868	0.4891	0.5056
WResNet-EDSR	0.6920	0.6856	0.6789	0.6877
WResNet	0.8137	0.8177	0.7501	0.7542
Ours	**0.8655**	**0.8666**	**0.7770**	**0.7709**

$$L(s_i, s_j, p_{ij}) = \sum_{\substack{0<i<n \\ i<j<n}} -p_{ij} \times f(s_i, s_j) \\ -(1 - p_{ij}) \times (1 - f(s_i, s_j)) \tag{7}$$

We implement the efficient back-propagation in Siamese networks proposed by RankIQA [16] to remarkably improve the training efficiency *i.e.*, from 5-7 hours per epoch to 20 minutes per epoch, and improve the performance of the pretrained model. We validate the effectiveness of our contrastive pretraining strategy in Sec. 4.6.

4. Experiments

4.1. Datasets

PIPAL [9] is a recently proposed IQA dataset, which contains images processed by image restoration and enhancement methods (particularly the deep learning-based methods) besides the traditional distorting methods. The dataset contains 250 reference images, 29k distorted images of 40 distortion types, and 1.13m human judgment quality scores. We also conduct cross-dataset evaluation on TID2013 [19] and LIVE [23].

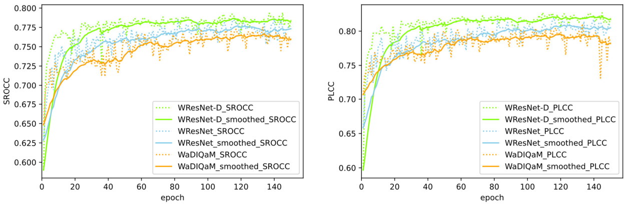


Figure 7. SROCC and PLCC performance of various models on our validation dataset. The solid curves are the smoothed ones with inertial filtering while the dotted curves are the original ones without smoothing. WResNet-D indicates WResNet with reference-oriented deformable convolutions.

Table 2. Cross-dataset evaluation on TID2013 and LIVE.

Method	TID2013		LIVE	
	SROCC	PLCC	SROCC	PLCC
PSNR	0.687	0.677	0.873	0.865
WaDIQaM	0.698	0.741	0.883	0.837
RADN	0.747	0.796	0.905	0.878

4.2. Implementation Details

Training Details. We train the proposed model with the patch-based training strategy in WaDIQaM [2], which has been proved to be able to augment the data effectively and improve the performance. For training, we randomly sample 32 patches from per distorted image and its corresponding reference image rather than the whole image. These sampled patches could be overlapped. We set the mini-batch size as 2 for the consideration of the remarkable difference between distorted images. We found in experiments that the model can better learn and adapt to such differences with small batch sizes. We use ADAM optimizer with the parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$. For our models, the learning rate is initialized as 10^{-4} and will decay to 0.8 of itself at every 100 epochs. For testing, each pair of images are cropped into a certain number M of 32 × 32 *non-overlapping* patches. These patches are then fed into the network to predict the weight w_i and the score of each patch s_i. The final quality score s for the distorted image is calculated by $s = \sum_{i=1}^{M} w_i s_i$.

Data Arrangement for Contrastive Pretraining. As mentioned in Sec. 4.1, the PIPAL dataset contains 7 distortion categories and 40 distortion subtypes. Each reference image corresponds to 116 distorted images of different sub-types from the seven distortion categories, but the specific subtype is unknown. Considering the gap among various images, we collect distorted images corresponding to the same reference and from the same category for contrastive pretraining. We call these image collections ‘contrast groups’. To implement the efficient Siamese Network, we arrange the images from the same contrast group into one batch to avoid the duplicate score computing process. Thus, the training time

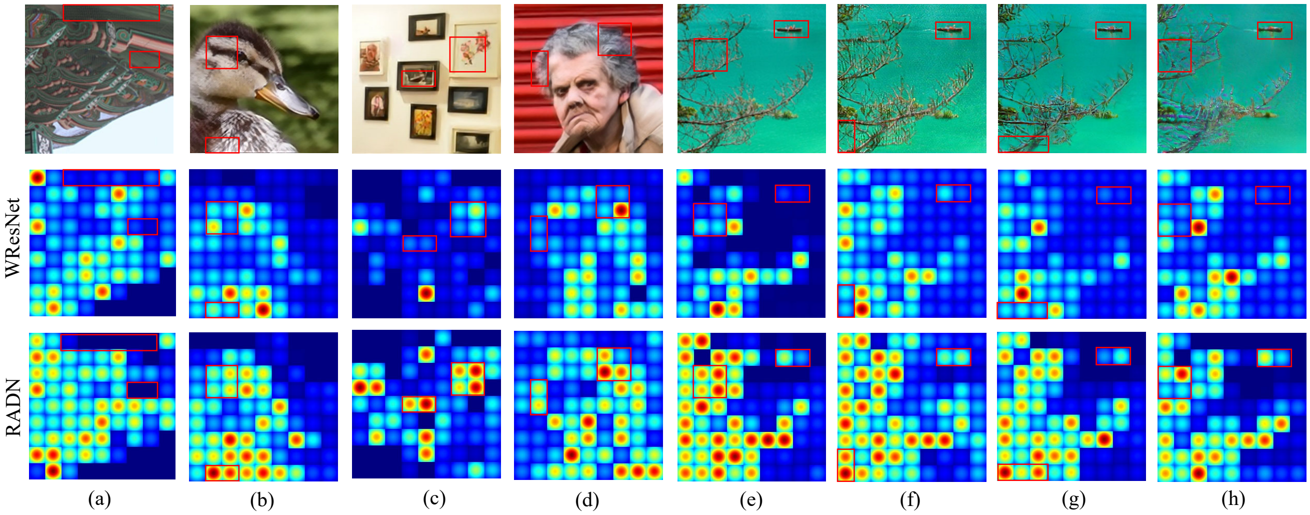


Figure 8. The attention maps of the images obtained by visualizing the weights of the patches. The distortion types of (a)-(e) are traditional, and the rest are GAN-based. The comparison between our proposed RADN and WResNet shows the effectiveness of our proposed modules, which contributes to focusing on the noteworthy regions for IQA tasks, especially for the GAN-based distortions.

would be essentially saved.

4.3. Evaluation Criteria

To evaluate the performance, we use Spearman's rank order correlation coefficient (SROCC) and Pearson's linear correlation coefficient (PLCC), following the prior works [8, 2, 16]. For N testing images, the PLCC is defined as follows:

$$\text{PLCC} = \frac{\sum_{i=1}^{N}(s_i - \mu_{s_i})(\hat{s}_i - \mu_{\hat{s}_i})}{\sqrt{\sum_{i=1}^{N}(s_i - \mu_{s_i})^2}\sqrt{\sum_{i=1}^{N}(\hat{s}_i - \mu_{\hat{s}_i})^2}} \tag{8}$$

Where s_i and $\hat{s}_i$ respectively indicate the ground-truth and predicted quality scores of i-th image, and μ_{s_i} and $\mu_{\hat{s}_i}$ indicate the mean of them. Let d_i denote the difference between the ranks of i-th test image in ground-truth and predicted quality scores. The SROCC is defined as

$$\text{SROCC} = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)} \tag{9}$$

Both metrics, PLCC and SROCC, are in [-1, 1], and higher values indicate better performance.

4.4. Comparison with the State-of-the-arts

Evaluation on PIPAL. We compare our models with the state-of-the-art FR-IQA methods on the NTIRE 2021 IQA challenge validation and testing datasets. The quantitative comparisons on both datasets are shown in Tab. 1 and the 'ours' term indicates our ensemble approach. We divide the general-purpose IQA methods into traditional methods and deep learning-based methods. In general, deep learning-based methods achieved better performance than the traditional methods. As can be seen, our method is superior to WaDIQaM [2] which is also a patch-based approach and widely used in the assessment of synthetically distorted images. Our method especially achieves superior results over PieAPP [20] which is considered the most effective approach in [9] on both datasets by a large margin.

Cross-dataset Evaluation on TID2013 and LIVE. To validate the generalization of our proposed RADN, we conduct the cross-dataset evaluation on TID2013 and LIVE. We trained RADN on the training set of PIPAL and test it on the full set of TID2013 and LIVE. As shown in Tab. 2, RADN outperforms WaDIQaM [2] and PSNR with large margins on both datasets, which indicates the effectiveness of our proposed modules.

4.5. NTIRE2021 IQA Challenge

Our methods are originally proposed for participating in the NTIRE 2021 Perceptual Image Quality Assessment Challenge[10], which aims to establish an algorithm to measure the visual quality of the images fairly and focuses on the PIPAL dataset. Our ensemble approach with the strategies mentioned above ranked 3rd place in the public validation phase (also called the development phase) and ranked 4th place in the final private test phase (as shown in Tab. 1).

4.6. Ablation Study

To further investigate the effectiveness of our proposed components, we conduct ablation studies on the validation

Table 3. Ablation study on the validation dataset of the NTIRE 2021 IQA Challenge. *Contrastive* refers to our contrastive pretraining strategy. *Deform* indicates our reference-oriented deformable module and *PatchAttn* indicates our patch-level attention module.

Contrastive	Deform	PatchAttn	SROCC	PLCC
			0.8137	0.8177
√			0.8244	0.8252
√	√		0.8329	0.8337
√		√	0.8343	0.8345
√	√	√	**0.8438**	**0.8435**

dataset of the NTIRE 2021 IQA Challenge. Both SROCCs and PLCCs are shown in Tab. 3.

Modified Residual Block. Fig. 7 depicts the performance of different models on the validation dataset during training. The dotted curves are the original ones without smoothing, while the solid curves are the smoothed ones obtained by inertial filtering. The orange curves indicate WaDIQaM, and the blue ones indicate WResNet with our modified residual blocks. We can easily conclude that WResNet outperforms WaDIQaM on SROCC and PLCC. During the training process, our WResNet ascend more steadily compared with the vibrated curve of WaDIQaM. Also as in Tab. 1, WResNet-Classic and WResNet-EDSR indicate WResNet with classic and EDSR-like residual blocks respectively. For a fair comparison, all three models adopt 20 convolution layers, and the superior performance of WResNet demonstrates the effectiveness of our modified residual block. The architectures of various residual blocks can refer to Fig. 3.

Contrastive Pretraining Strategy. The contrastive training strategy is adopted for pretraining our models. Tab. 3 shows that the proposed pretraining strategy can learn the contrastive knowledge priors, which can improve the model's performance.

Reference-oriented Deformable Convolution. We only add reference-oriented deformable convolution to our baseline (*i.e.*, WResNet+Deform), a major improvement on SROCC and PLCC are shown in Tab. 3. The training process (indicate by the green curves) in Fig. 7 also shows superior performance compared with our baseline model, which indicates the proposed reference-oriented deformable convolution module can adapt to the GAN-based distortion scenarios in PIPAL.

Patch-Level Attention. The patch-level attention mechanism is proposed to enhance interactions among all patches. The results in Tab. 3 show such interactions are indispensable for the patch-based IQA algorithms. To be specific, the *patch-level attention module* gains another 0.01 improvements on SROCC and PLCC, respectively.

Combined with all the components we proposed, RADN significantly improves the evaluation performance, especially compared to our baseline.

4.7. Visualization and Discussion

To intuitively illustrate the effectiveness of our method, we visualize the weights of patches in some images, as shown in Fig. 8. According to the weight from the lowest to the highest, the color of the patch is displayed as blue, green, yellow, and red. A higher weight means the model pays more attention to the patch region. The distortion types of (a)-(e) are traditional, and the rest are GAN-based. The second and third-row visualization results come from the WResNet (without the deformable and patch-level attention modules) and the whole method RADN.

As shown, both methods can adaptively give higher weights to noteworthy regions. These regions usually contain complex textures or salient subjects, which are essential in assessing the image quality because human tends to pay more attention to such regions. The highlighted regions are indicated with red boxes. Compared with WResNet, our whole model RADN pays less attention to the regions with less informative or the texture-less regions as shown in Fig. 8 (a). As shown in Fig. 8 (b)-(e), for images of different content, RADN perceives the images better and observably stays under human's perception - for example, the contours of the duck's head, the paintings, and the elder's head are clearly outlined, which strongly shows the effectiveness of our proposed modules.

To further illustrate the effectiveness of our method on GAN-based distortion, we show the visualized results for different GAN-based distortion images according to the same reference in Fig. 8 (f)-(h). Despite the diversity and severe GAN-based distortions, our RADN can capture the actual outline of the attracting targets like the boat and the branches of the tree because of the reference-oriented deformable convolution. Also, RADN distributes fewer attention weights in flat areas like the sea surface due to the effectiveness of the patch-level attention module.

5. Conclusion

We propose a full-reference image quality assessment approach called Region-Adaptive Deformable Network (RADN). We first revisit the classic residual blocks and propose the modified residual blocks from the viewpoint of IQA tasks, which are used to build our baseline. We introduce the patch-level attention mechanism for information interaction among the patch regions and the reference-oriented deformable convolution for adaptation to images of significant differences. We also propose a contrastive pretraining strategy to further improve the model's capability to truly distinguish the image quality rather than directly learning the regression of the quality score. The experimental results reveal the excellent effectiveness of the proposed method. Our ensemble method ranked fourth in the NTIRE 2021 Perceptual Image Quality Assessment Challenge.

References

[1] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *CVPR*, pages 6228–6237, 2018. 2, 6

[2] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *TIP*, 27(1):206–219, 2017. 1, 2, 6, 7

[3] Damon M Chandler and Sheila S Hemami. Vsnr: A wavelet-based visual signal-to-noise ratio for natural images. *TIP*, 16(9):2284–2298, 2007. 6

[4] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, pages 764–773, 2017. 2, 4

[5] Niranjan Damera-Venkata, Thomas D Kite, Wilson S Geisler, Brian L Evans, and Alan C Bovik. Image quality assessment based on a degradation model. *TIP*, 9(4):636–650, 2000. 6

[6] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Comparison of full-reference image quality models for optimization of image processing systems. *IJCV*, pages 1–24, 2021. 6

[7] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *CVPR*, 2021. 3

[8] 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. 1, 2, 3, 4, 5, 6, 7

[9] Jinjin Gu, Haoming Cai, Haoyu Chen, Xiaoxing Ye, Ren Jimmy S, and Chao Dong. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In *ECCV*, pages 633–651. Springer, 2020. 6, 7

[10] Jinjin Gu, Haoming Cai, Chao Dong, Jimmy S. Ren, Yu Qiao, Shuhang Gu, Radu Timofte, et al. NTIRE 2021 challenge on perceptual image quality assessment. In *CVPR Workshops*, 2021. 2, 7

[11] Eric Cooper Larson and Damon Michael Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of electronic imaging*, 19(1):011006, 2010. 6

[12] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPR Workshops*, July 2017. 3

[13] Anmin Liu, Weisi Lin, and Manish Narwaria. Image quality assessment based on gradient similarity. *TIP*, 21(4):1500–1512, 2011. 6

[14] Ding Liu, Zhaowen Wang, Yuchen Fan, Xianming Liu, Zhangyang Wang, Shiyu Chang, and Thomas Huang. Robust video super-resolution with learned temporal dynamics. In *ICCV*, pages 2507–2515, 2017. 3

[15] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang. Non-local recurrent network for image restoration. *arXiv preprint arXiv:1806.02919*, 2018. 3

[16] Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. Rankiqa: Learning from rankings for no-reference image quality assessment. In *ICCV*, pages 1040–1049, 2017. 1, 2, 5, 6, 7

[17] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for single-image super-resolution. *CVIU*, 158:1–16, 2017. 6

[18] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a “completely blind” image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 6

[19] 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. 6

[20] Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *CVPR*, pages 1808–1817, 2018. 2, 6, 7

[21] Hamid R Sheikh and Alan C Bovik. Image information and visual quality. *TIP*, 15(2):430–444, 2006. 6

[22] Hamid R Sheikh, Alan C Bovik, and Gustavo De Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *TIP*, 14(12):2117–2128, 2005. 6

[23] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *TIP*, 15(11):3440–3451, 2006. 6

[24] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *CVPR*, pages 3360–3369, 2020. 3

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 3

[26] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *CVPR Workshops*, pages 0–0, 2019. 3

[27] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 3, 5

[28] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *ECCV Workshops*, 2018. 3

[29] Zhou Wang and Alan C Bovik. A universal image quality index. *IEEE signal processing letters*, 9(3):81–84, 2002. 6

[30] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 13(4):600–612, 2004. 2, 6

[31] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003. 2, 6

[32] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, pages 3–19, 2018. 3

[33] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Jing Xiao. Domain fingerprints for no-reference image quality assessment. *TCSVT*, 2020. 1

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

[35] Sheng Yang, Qiuping Jiang, Weisi Lin, and Yongtao Wang. Sgdnet: An end-to-end saliency-guided deep neural network for no-reference image quality assessment. In *ACM MM*, pages 1383–1391, 2019. 3

[36] Lin Zhang and Hongyu Li. Sr-sim: A fast and high performance iqa index based on spectral residual. In *ICIP*, pages 1473–1476. IEEE, 2012. 2, 6

[37] Lin Zhang, Ying Shen, and Hongyu Li. Vsi: A visual saliency-induced index for perceptual image quality assessment. *TIP*, 23(10):4270–4281, 2014. 6

[38] Lin Zhang, Lei Zhang, and Xuanqin Mou. Rfsim: A feature based image quality assessment metric using riesz transforms. In *ICIP*, pages 321–324. IEEE, 2010. 6

[39] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *TIP*, 20(8):2378–2386, 2011. 2, 6

[40] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 6

[41] Hancheng Zhu, Leida Li, Jinjian Wu, Weisheng Dong, and Guangming Shi. Metaiqa: deep meta-learning for no-reference image quality assessment. In *CVPR*, pages 14143–14152, 2020. 2

[42] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2