Region-Adaptive Deformable Network for Image Quality Assessment

Shuwei Shi1 Qingyan Bai1 Mingdeng Cao2 Weihao Xia1 Jiahao Wang2 Yifan Chen1 Yujiu Yang1†
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.
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 $3 \times 3$ 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-
### 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 $32 \times 32$. 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:

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

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 PI-PAL 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 $3 \times 3$ convolution without any $7 \times 7$ employed in the original ResNet. The $3 \times 3$ convolution has been proven to be more hardware-friendly [7] and more effective than the $7 \times 7$ convolution layer followed by...
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 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 \( p_k, k = 1, 2, \cdots, N \), then sums the sampled values weighted by \( W_k \) corresponding to the \( k \)th location. For example, a \( 3 \times 3 \) kernel has 9 regular sampling locations which are defined as \( p_k \in \{-1, -1\}, (-1, 0), \cdots, (1, 1)\} \). In terms of our reference-oriented deformable convolution, we first generate the offsets \( \Delta p_k \) according to \( k \)th sampling location from reference feature maps:

\[
\Delta p_k = f(F_{RM})
\]

where \( f \) means a \( 3 \times 3 \) convolution, the output channel number of which is \( 2N \). Then the learned offsets \( \Delta p_k \) are used for sampling the values in both reference and distorted feature maps, rather than the regular sampling with \( p_k \). For each location \( p_0 \) on the output reference and distorted feature maps \( F_{RM} \) and \( F_{DM} \), the value in it is aggregated by the following process:

\[
\hat{F}_{RM}(p_0) = \sum_{k=1}^{N} w^r_k \cdot F_{RM}(p_0 + p_k + \Delta p_k),
\]

\[
\hat{F}_{DM}(p_0) = \sum_{k=1}^{N} w^d_k \cdot F_{DM}(p_0 + p_k + \Delta p_k),
\]

where \( w^r \) and \( w^d \) are the convolution weights for reference and distortion feature maps. Owing to the deformable sampling locations \( p_0 + p_k + \Delta p_k \) 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
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 of each patch with each other and have more accurate feature representation.

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 Queries, Keys and Values of dimension \( d_k \) by performing linear projection on the patch-level feature vectors \( X \) with \( W_Q, W_K, W_V \):

\[
Q = W_QX, \quad K = W_KX, \quad V = W_VX
\]  

(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:

\[
Y = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V
\]  

(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:

\[
Z = W_ZY + X
\]  

(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 \( \{s_1, s_2, \ldots, s_n\} \) 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:

\[
\text{Cross Entropy Loss} = - \sum_{i=1}^{n} p_{ij} \log f(s_i, s_j)
\]
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       | SROCC Validation | PLCC Validation | SROCC Test | PLCC Test |
|--------------|------------------|-----------------|------------|-----------|
| 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    |
| FSIMe [39]   | 0.4679           | 0.5587          | 0.5057     | 0.5727    |
| VSI [37]     | 0.4501           | 0.5162          | 0.4584     | 0.5169    |
| MAD [14]     | 0.6078           | 0.6263          | 0.5434     | 0.5804    |
| NIQE [18]    | 0.6044           | 0.1016          | 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   | 0.6276           | 0.6463          | 0.5658     | 0.5711    |
| LPIPS-VGG    | 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    |

$$L(s_i, s_j, p_{ij}) = \sum_{0 < i < n}^\{n\} -p_{ij} \times f(s_i, s_j) - (1 - p_{ij}) \times (1 - f(s_i, s_j))$$ (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].

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 minibatch 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 $32 \times 32$ non-overlapping patches. These patches are then fed into the network to predict the weight $w_{ij}$ and the score of each patch $s_i$. The final quality score $s$ for the distorted image is calculated by $s = \sum_{i=1}^{M} w_{ij}s_i$.

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:

\[
PLCC = \frac{\sum_{i=1}^{N} (s_i - \mu_s)(\hat{s}_i - \mu_{\hat{s}})}{\sqrt{\sum_{i=1}^{N} (s_i - \mu_s)^2} \sqrt{\sum_{i=1}^{N} (\hat{s}_i - \mu_{\hat{s}})^2}}
\]  

(8)

Where \( s_i \) and \( \hat{s}_i \) respectively indicate the ground-truth and predicted quality scores of \( i \)-th image, and \( \mu_s \) and \( \mu_{\hat{s}} \) 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

\[
SROCC = 1 - \frac{6 \sum_{i=1}^{N} d_i^2}{N(N^2 - 1)}
\]  

(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            | 0.8137      | 0.8244 | 0.8329    |
| PLCC             | 0.8177      | 0.8252 | 0.8337    |
|                  | 0.8343      | 0.8345 |           |

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

[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.

[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.

[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.

[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.

[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.

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

[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.

[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.

[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.