IQMA Network: Image Quality Multi-scale Assessment Network

Haiyang Guo Yi Bin Yuqing Hou Qing Zhang Hengliang Luo
Meituan
Beijing, China
{guohaiyang02, binyi, houyuqing, zhangqing31, luohengliang}@meituan.com

Abstract

Image Quality Assessment (IQA), which aims to provide computational models for automatically predicting perceptual image quality, is an important computer vision task with many applications. In recent years, a variety of IQA methods have been proposed based on different metric designs, which measure the quality of images affected by various types of distortion. However, with the rapid development of Generative Adversarial Networks (GAN), a new challenge has been brought to the IQA community. Especially, the GAN-based Image Reconstruction (IR) methods overfit the traditional PSNR-based IQA methods by generating images with sharper edges and texture-like noises, leading the outputs to be similar to the reference image in appearance but with loss of details. In this paper, we propose a bilateral-branch multi-scale image quality estimation network, named IQMA network. The two branches are designed with Feature Pyramid Network (FPN)-like architecture, extracting multi-scale features for patches of the reference image and corresponding patches of the distorted image separately. Then features of the same scale from both branches are sent into several scale-specific feature fusion modules. Each module performs feature fusion and a novelly designed pooling operation for corresponding features. Then several score regression modules are used to learn a quality score for each scale. Finally, image scores for different scales are fused as the quality score of the image. IQMA network has achieved 1st place on the NTIRE 21 IQA public leaderboard and 2nd place on the NTIRE 21 IQA private leaderboard, and consistently outperforms existing state-of-the-art (SOTA) methods on LIVE and TID2013.

1. Introduction

Image quality assessment methods are developed to measure the perceptual quality of images after distortion or post-processing operations [15]. It is vital for many visual tasks, such as benchmarking image processing algorithms, i.e. image restoration [4]. With the rapidly growing need for visual analysis, the IQA task has received lots of attention. Current IQA methods can be categorized as Full-Reference methods (FR-IQA) and No-Reference ones (NR-IQA). FR-IQA methods measure the quality difference by comparing the target (distorted) image with the reference (ground truth) image. In this paper, we focus on the FR-IQA, which is the same as the setting of NTIRE 21 IQA challenge.

The most widely-used traditional metrics for FR-IQA are Peak Signal-to-Noise Ratio (PSNR) [2] and Structural SIMilarity (SSIM) [25]. These two metrics are of simple mathematical formulations and have achieved remarkable performances on different benchmarks. Existing works have revealed the theoretical connection between PSNR and SSIM, and their advantages for different kinds of traditional noises, such as PSNR for additive Gaussian noise and SSIM for JPEG-based compression noise [13].

During recent years, many deep learning-based IQA methods have been proposed [5, 8, 10, 20, 30]. Although deep learning methods have achieved SOTA performance, most of them aim to optimize the PSNR or SSIM metric. On the other hand, with the rapid development of GAN, new types of distortion are brought to the IQA field. The GAN-generated images usually have seemingly realistic yet fake details and textures, bringing in texture-like noises [15]. The GAN-based distortion can be easily perceived by human eyes but is challenging for the widely used PSNR or SSIM metrics. Thus it is necessary to design new methods to narrow the gap.

The NTIRE 21 IQA challenge provides dataset (PIPAL) contains 4 categories of distortions: Traditional, Super-Resolution (where GAN-based distortion is included), Denoizing and Mixture Restoration [15]. This setting is similar to real-world scenarios where different types of distortion coexist, and distortion at different scales coexist. For traditional distortions, methods which extract structural information and evaluate image at the global scale usually perform the best. However, to capture the GAN-generated texture-like noises, it is necessary to evaluate the image at a fine-grained texture level. In other words, small-scale image analysis is vital for GAN-based distortion. Thus we de-

sign a multi-scale image evaluation architecture to measure distortions at different scales. Inspired by the two-stream approach [5], our network is designed as a bilateral-branch architecture. One branch is used to extract multi-scale features of the distorted image, the other for extracting features of the corresponding reference image. The two branches share the same architecture and parameters. FPN [16] is used as the backbone for each branch, which can extract and fuse multi-scale features efficiently. To further capture the features at the fine-grained level, images are divides into small patches for analysis.

Features of the same patch position at the same scale from two branches are sent in pairs to a feature fusion module for quality prediction. The module performs early feature fusion, max pooling, and mean pooling, and then outputs the fused feature. Then all the fused features for the same scale are sent to a score regression module to predict the image quality score for that scale. At last, a final image score is calculated by averaging quality scores for different scales. Besides the network design, we also perform data augmentation to alleviate the imbalanced distribution of image scores.

To summarize, our major contributions are as follows:

1. Image quality is evaluated on different scale levels to capture different kinds of distortion. An FPN-based bilateral-branch multi-scale quality assessment network is designed to estimate image quality.

2. Features of the same scale are sent to corresponding feature fusion and score regression modules. Early fusion, a novel pooling operation, and a novel attention-based score prediction network are utilized to learn the image score of a specific scale.

3. Data augmentation is used to make quality score distribution be more balanced.

4. Our method has achieved the 1st place on the NTIRE 21 IQA public leaderboard (the 1st on both SRCC and PLCC) and the 2nd place on the private leaderboard (the 1st on SRCC and the 2nd on PLCC) and significantly outperforms existing methods on benchmark datasets LIVE [21] and TID2013 [19].

The remaining part of this paper is organized as follows. Section 2 introduces the related work. Section 3 introduces the proposed IQMA network in detail. Section 4 provides our remarkable performance on the training set, validation set , and testing set, as well as extensive ablation study results. The conclusion is given in Section 5.

2. Related Work

Our method is motivated by the fact that traditional FR-IQA metrics may fail when evaluating GAN-based distortion. To capture the GAN-based distortion as well as traditional distortion, we need to design a multi-scale network for feature extraction. And the bilateral-branch network is a natural fit for FR-IQA since pair of distorted image and reference can be processed simultaneously. And pair of features should be combined to predict the image quality. In this section, we will briefly introduce the related works.

2.1. Metrics for Image Quality Assessment

Quality assessment for an image can be performed subjectively or objectively. The most accurate way to judge image quality is by judging through human eyes subjectively, which inspired the notion of Mean Opinion Score (MOS), a numerical measure of human judgement. However, getting the MOS can be inconvenient or slow in some cases [25], leading the research field to find more automatic and objective methods for IQA.

Various objective methods have been proposed for FR-IQA [4, 18, 20, 22, 23, 26, 27]. Some early works simply use L_p norm to measure the similarity between target and reference, at the cost of accuracy [26]. Another line of research utilizes knowledge from Human Visual System (HVS) to guide the design of the similarity metrics. However, all the HVS-inspired models have failed on real-world tasks owing to the difficulty of accurately modeling the complexity of the HVS [20].

2.2. Multi-scale Feature Extraction

Multi-scale feature extraction is vital for IQA since the perceptual quality of an image depends on the distance from the image plane to the observer and the perceptual capability of the observer's visual system. Early works have proposed Multi-Scale SSIM (MS-SSIM) as an enhanced metric for single-scale SSIM [27]. However, MS-SSIM depends on handcrafted features and needs parameter tuning. With the development of deep learning, FPN [16] is proposed which can generate multi-scale feature maps end-to-end. FPN extends the featurized image pyramid [1] design with upsampling operations to extract semantically stronger features. It combines low-resolution, semantically strong features with high-resolution, semantically weak features via a top-down pathway and lateral connections. FPN has achieved remarkable performance on SOTA benchmarks, and similar network designs have been applied for different fields [14, 24].

2.3. Bilateral-branch Feature Fusion

Bilateral-branch networks, especially Siamese networks [3], have been used to learn similarity between two inputs. The inputs are processed in parallel by two networks sharing their synaptic connection weights. The architecture is naturally fit for feature extraction of FR-IQA tasks, and [5]

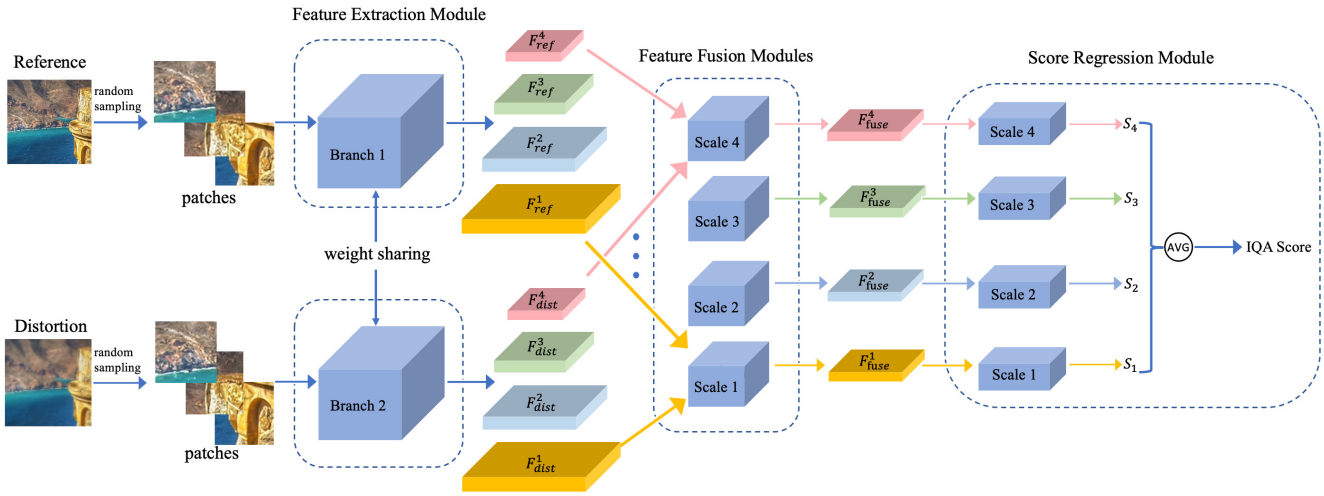

Figure 1: The overall framework of IQMA network. Random sampled pairs of patches are sent to the network. Features corresponding to four scales are extracted by the Feature Extraction Module. Then four Feature Fusion Modules fuse corresponding features of the same scale from two branches and output four fused features. Four Score Regression Block predict four scores. And the four scores are averaged as the final IQA score.

has proposed a bilateral-branch for both FR-IQA and NR-IQA. In their work, feature extraction is followed by a feature fusion step, where features from both branches and their difference are all considered. They are concatenated before pooling. Besides concatenation, adding, stacking and outer product [9] are also frequently used for feature fusion.

Features can be fused at different stages. Early fusion combines features before classification (regression), and late fusion combines the outputs after classification (regression). Early fusion is performed on feature-level, where the feature vectors from different sources are concatenated into a new feature vector which will then be used for classification. For IQMA network, early fusion is utilized in the Feature Fusion Modules.

For the pooling of the fused feature, bilinear pooling is widely used when features coming from different modalities [9]. However, for our network, we adopted a much simpler solution proposed by [31] to concatenate the output of max pooling and mean pooling.

2.4. Class-Imbalanced Learning

Real-world data usually has class-imbalanced distribution. Several learning strategies are developed to cope with this kind of task, such as re-sampling, class-based re-weighted loss, and data augmentation. Re-sampling tries to reshape the original imbalanced dataset into a class-based uniform one by down-sampling or up-sampling [11]. Re-weighting assigns different weights to samples of different classes, i.e., multiplying large (small) weights for training samples of minority (majority) classes in loss function [17].

Data augmentation shares the same spirit of Re-sampling by generating synthetic samples for minority classes to make the distribution of the dataset more balanced [6, 7]. The synthetic samples can be generated based on original samples by various operations, such as image shift, flips, rotation, brightness adjustment, random cropping, zooming, and so on. For the PIPAL dataset, we can observe a severe imbalanced distribution of image score and utilize the simple data augmentation technique to balance their distribution.

3. Image Quality Multi-scale Assessment network (IQMA network)

In this section, we introduce the proposed IQMA network. As shown in Fig. 1, the IQMA network takes pairs of image patches as input. It mainly consists of 3 parts: an FPN-based **Feature Extraction Module** which extracts features at four different scales for each pair of patches (reference, distortion); Four **Feature Fusion Modules**, each takes a pair of features with the corresponding scale and outputs fused features; **Score Regression Module** consists of four Score Regression Blocks, each of which takes fused features as input and predicts a quality score for that scale. Then the four scores for different scales are averaged as the predicted score.

3.1. A Scale-wise Patch-based Framework

Compared with the whole image, a patch can better represent fine-grained information. Thus IQMA network takes pairs of patches as input using bilateral-branch. Each reference image is randomly cropped into K different patches

Algorithm 1: IQMA network

input : A pair of images ref and $dist$
output: A predicted IQA score $IQA\ Score$

Denote RandomCropping as RC();
Denote FeatureExtraction as FE();
Denote FeatureFusion as FF();
Denote ScoreRegression as SR();
 # Random Cropping into K patches;
$\{P_{ref_1}, P_{ref_2}, \cdots, P_{ref_K}\} \leftarrow \mathrm{RC}\ (ref)$;
$\{P_{dist_1}, P_{dist_2}, \cdots, P_{dist_K}\} \leftarrow \mathrm{RC}\ (dist)$;
for $i \leftarrow 1$ **to** K **do**
  for $j \leftarrow 1$ **to** 4 **do**
    # Feature Extraction;
    $F^j_{ref_i} \leftarrow \mathrm{FE}\ (P_{ref_i}, scale = j)$;
    $F^j_{dist_i} \leftarrow \mathrm{FE}\ (P_{dist_i}, scale = j)$;
    # Feature Fusion;
    $F^j_{fuse_i} \leftarrow \mathrm{FF}\ (F^j_{ref_i}, F^j_{dist_i})$
  end
end
 # Score Regression for each scale;
for $j \leftarrow 1$ **to** 4 **do**
  $S_j \leftarrow \mathrm{SR}\ (\{F^j_{fuse_1}, F^j_{fuse_2}, \cdots, F^j_{fuse_K}\})$;
end
 # Get final IQA score;
$IQA\ Score \leftarrow \frac{\sum_{j=1}^{4} S_j}{4}$

with the size of $32*32$ or $64*64$, while its distortion image is cropped in the same way.

For each patch, features corresponding to 4 different scales are extracted by the ResNet [12] or ResNeXt [28]-based Feature Extraction Module, where feature map sizes are $C*8*8, C*4*4, C*2*2$ and $C*1*1$, where C represents the number of channel.

To learn different types of distortion, the IQMA network calculates 4 image quality scores for each distortion image. Each quality score corresponds to the image quality evaluated at different resolutions. We define the scale level j as $j \in \{level1, level2, level3, level4\}$ where $\{level1 : C*8*8, level2 : C*4*4, level3 : C*2*2, level4 : C*1*1\}$. For example, the image quality score for $j = level3$ means that the score is calculated based on features with the size of $C*2*2$ for all the patches. The predicted score is calculated as the average of the 4 image quality scores.

We summarize the proposed method in Algorithm 1.

3.2. Feature Extraction Module

Existing methods have achieved good performance on images with traditional noises but often fail on images with GAN-generated noise. These methods usually put more emphasis on global structure analysis but ignore the texture-like noises, making them have low tolerance toward spatial misalignment [15]. To cope with this kind of distortion, it is vital to consider feature vectors for both large scale and small scale. This point of view motivates us to design a feature pyramid-based multi-scale feature extractor.

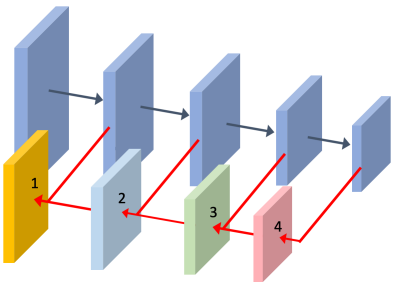

Figure 2: Structure of Feature Extraction Module. Different color represents feature maps with different scales.

ResNet or ResNeXt are utilized as the backbone, where output feature maps of different blocks correspond to different resolutions. To improve the alignment of features of different scale, we adopt the design of FPN [16], as shown in Fig. 2. Different color represents feature maps with different scales.

Denote patch i of one reference image and its corresponding patch of the distortion image as P_{ref_i} and P_{dist_i}. The feature extraction module will extract 4 features of different scales for each of them, denotes as $F^j_{ref_i}$ and $F^j_{dist_i}$, where j denotes scale as stated above.

3.3. Feature Fusion Modules

Each Feature Fusion Module contains 2 blocks, namely, Feature Smoothing Block and Feature Pooling Block, as shown in Fig. 3.

3.3.1 Feature Smoothing Block

Taking a pair of feature vectors of scale level j as input: $(F^j_{ref_i}, F^j_{dis_i})$, the Feature Smoothing Block concatenates them as $concat(F^j_{ref_i}, F^j_{dist_i}, F^j_{ref_i} - F^j_{dist_i})$, the last part of which models the difference between the patches [5].

The concatenated features go through 2 convolution layers for further fusion, with kernel size 3×3 and 2×2, respectively. The module outputs a concatenated feature map F_{concat_i}.

3.3.2 Feature Pooling Block

Different from existing work which improves the performance of IQA networks on GAN-based distortions by introducing anti-aliased pooling layers [15], the Feature Pooling Block pools the feature maps with combined pooling operations. It first performs mean pooling and max pooling, then

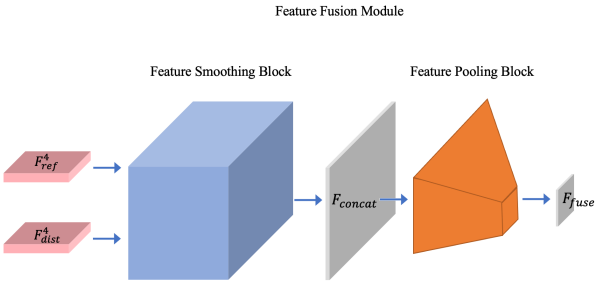

Figure 3: Structure of a Feature Fusion Module.

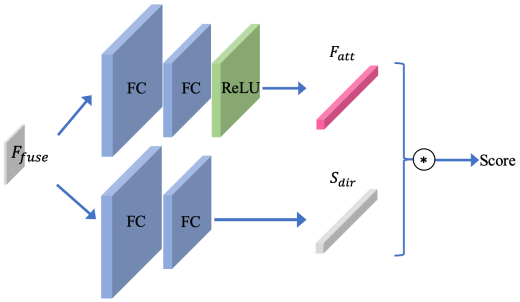

Figure 4: Structure of a Score Regression Block.

concatenates the results as output $F^j_{fuse_i}$, as shown below:

$$
\begin{aligned}
&F^j_{mean_i} = mean(F^j_{concat_i}, axis = [3, 4]),\\
&shape = (N * K, C, 1, 1);\\
&F^j_{max_i} = max(F^j_{concat_i}, axis = [3, 4]),\\
&shape = (N * K, C, 1, 1);\\
&F^j_{fuse_i} = concat(F^j_{mean_i}, F^j_{max_i}),\\
&shape = (N * K, 2 * C, 1, 1),
\end{aligned}
$$

where N denotes the batch size, K represents the number of patch. Comparing with the more complicated pooling operation conducted by [5], we discard the min pooling part since we observe that min pooling has a negative impact on learning fine-grained distortions.

3.4. Score Regression Module

The Score Regression Module consists of 4 Score Regression Blocks followed by an averaging operation. The Score Regression Block j predicts the quality score for scale j, outputs S_j. And the averaging operation gets the predicted score by $\frac{\sum_{j=1}^{4} S_j}{4}$

Spatial pooling does not consider the effect of spatially varying perceptual relevance of local quality. Inspired by the weighted average patch aggregation design [5], each Score Regression Block is designed with two branches, as shown in Fig. 4. A direct score prediction branch is used to directly predict the quality of a specific patch, and a patch-wise quality regression branch takes all patches of the same scale for consideration, outputs an attention vector. Then the two outputs are combined to obtain the quality score S_j for level j.

The 2 branches share the structure of 2 fully connected layers. The direct score prediction takes $F^j_{fuse_i}$ as input and outputs a quality score S_{dir}. For the patch-wise quality regression branch, it further maps all the direct features to positive ones by: $F^j_{pos_i} = ReLU(S^j_{dir_i}) + const$, where $const$ = 1e-8. Then the attention feature is calculated as:

$$
F^j_{att_i} = \frac{F^j_{pos_i}}{\sum_{t=1}^{K} F^j_{pos_t}}
$$

where the division operation is a bitwise division.

And we got the image quality score for level j as $S_j = \frac{\sum_i^K \left\langle F^j_{att_i}, S^j_{dir_i} \right\rangle}{K}$, and the predicted IQA score = $\frac{\sum_{j=1}^{4} S_j}{4}$.

4. Experiments

4.1. Datasets

Our experiments are conducted on three datasets. The first two are widely used benchmark datasets LIVE [21] and TID2013 [19]. LIVE includes 29 reference images and 779 distorted images corrupted by 5 types of distortions, i.e., JPEG compression (JPEG), JPEG2000 compression (JP2K), white noise (WN), Gaussian blur (GB), and simulated fast fading Rayleigh channel (FF). Each distortion type contains 5 or 4 distortion levels. TID2013 includes 25 reference images and 3,000 distorted images corrupted by 24 types of distortions, with 5 levels for each distortion type. The third one is the PIPAL dataset [15], which is provided by the NTIRE 21 IQA Challenge. The PIPAL training set contains 200 reference images, 23,200 distortion images, 40 distortion types, and 23,200 Elo scores for all distortion image. Especially, compared with other IQA datasets, it contains many GAN-based distortion types produced by GAN-based algorithms. PIPAL validation set contains 25 reference images and 1,000 distortion images. PIPAL testing set contains 25 reference images and 1,650 distortion images.

In this paper we mainly report IQMA's performance and ablation studies on PIPAL. Cross-Dataset evaluation are provided in Section 4.5 and supplymentary material.

4.2. Evaluation Metrics

Model performance is evaluated by two common measurements: Pearson linear correlation coefficient (PLCC) and Spearman rank order correlation coefficient (SRCC).

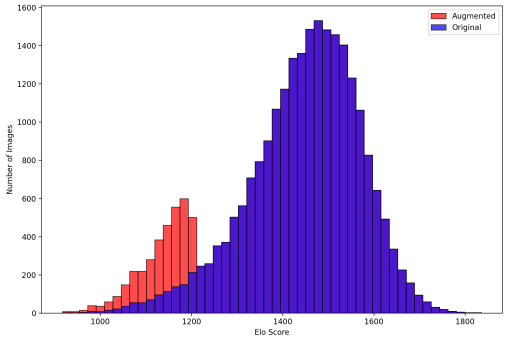

Figure 5: Image distribution of PIPAL training set on quality score before and after data augmentation. Blue: original distribution. Red: images we augmented.

PLCC is used to evaluate the linearity and consistency of the prediction, its definition is: PLCC = $\frac{\sum_{i=1}^{N}(p_i-\bar{p})(s_i-\bar{s})}{\sqrt{\sum_{i=1}^{N}(p_i-\bar{p})^2(s_i-\bar{s})^2}}$. Here s_i and p_i indicate the i-th image's subjective score and converted objective rating after nonlinear mapping, $\bar{s}$ and $\bar{p}$ are the mean of all s_i and p_i, N is the number of testing images. SRCC is used to evaluate the monotonicity of the performance, its definition is: SRCC = $1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)}$, where d_i represents the difference between the i-th images' ranks in subjective and objective evaluations [29]. In the NTIRE 21 IQA Challenge, the sum of PLCC and SRCC is the final score.

4.3. Data Augmentation

We analyzed the data distribution of the PIPAL training set and observed a severe imbalanced distribution on image score, as is shown in Fig.5. The blue histogram demonstrates the distribution of image numbers according to their image scores. Especially, only 3.78% images with Elo score < 1300. To alleviate the severe imbalance distribution, we performed data augmentation for images with Elo Score <1300. We up-sampled these images and used random horizontal flipping, vertical flipping, and $90°/180°/270°$ rotations to augment. After data augmentation, images with Elo score < 1300 were increased to 15.12%, as shown in the red part. We used 20,880 distortion-reference image pairs in PIPAL for training and the rest 2,320 image pairs for validation. ResNet pre-trained on ImageNet dataset was used as our backbone neural network to extract features.

The other two datasets have balanced distributions, and we divided their training set/validation set/testing set as 19/5/5 for LIVE and 17/4/4 for TID2013.

Figure 6: IQMA predicted scores vs. Elo scores on PIPAL validation set.

4.4. Implementation Details

During the training phase, we used the Adam optimizer with weight-decay α=1e-4, β_1 = 0.9 and β_2 = 0.999. We used smooth L_1 as loss function since compared with MSE loss it is more robust to outliers. The initial learning rate was set to 1e-4 and the cosine annealing learning rate scheduler was adopted with about 100K steps. The minimum learning rate was 4e-6. The mini-batch size was set to 16. We used PyTorch 1.4, NVIDIA V100 GPU with CUDA 10.0, and Multi-GPUs parallel training to accelerate training.

4.5. Comparison with SOTA Methods

We compare IQMA network with several SOTA methods on all three datasets. The methods include shallow methods PSNR [2] and SSIM [26], and deep learning-based methods such as LPIPS [30], PieAPP [20], DISTS [8], and SWD [10]. As shown in Tab. 1, the IQMA network achieves remarkable performance advantages over other methods on PIPAL validation set, PIPAL testing set, LIVE, and TID2013. On PIPAL, for the SRCC metric, the IQMA network outperforms the 2nd best method by a margin of 0.16 on the validation set and 0.14 on the testing set. For the PLCC metric, the IQMA network outperforms the 2nd best method by a margin of 0.17 on the validation set and 0.09 on the testing set. The IQMA network consistently outperform all the other methods on LIVE, and beats the 2nd best method with a large margin on TID2013. The scatter plot of the IQMA predicted scores vs. Elo scores on the PIPAL evaluation set is illustrated in Fig. 6. Some visual samples from PIPAL training set and scatter plots for LIVE and TID2013 are provided in supplementary material.

	PIPAL Validation Set		PIPAL Testing Set		LIVE		TID2013	
Model	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
IQMA	**0.87200**	**0.87590**	**0.78030**	**0.80090**	**0.97184**	**0.96881**	**0.94951**	**0.95030**
PSNR [2]	0.29165	0.25475	0.27693	0.24934	0.86500	0.87300	0.67700	0.68700
SSIM [26]	0.39842	0.33996	0.39356	0.36137	0.93700	0.94800	0.77700	0.72700
LPIPS-Alex [30]	0.64629	0.62758	0.57106	0.56580	0.93400	0.93200	0.74900	0.67000
LPIPS-VGG [30]	0.64712	0.59146	0.63306	0.59472	0.93400	0.93200	0.74900	0.67000
PieAPP [20]	0.69721	0.70627	0.59741	0.60741	0.90800	0.91900	0.85900	0.87600
DISTS [8]	0.68580	0.67428	0.68732	0.65483	0.95400	0.95400	0.85500	0.83000
SWD [10]	0.66802	0.66114	0.63419	0.62429	-	-	-	-

Table 1: Performance comparisons on the PIPAL, LIVE, and TID2013.

	Train patch detail	Test patch detail		
Backbone	(patch size) $*$ number of patches	(patch size) $*$ number of patches	PLCC	SRCC
Resnet50	(32$*$32)$*$512	(32$*$32)$*$1024	0.860	0.855
Resnet50	(64$*$64)$*$128	(64$*$64)$*$1024	0.857	0.845
Resnet101	(32$*$32)$*$512	(32$*$32)$*$1024	0.852	0.845
Resnet101	(64$*$64)$*$128	(64$*$64)$*$1024	0.852	0.844
Resnet152	(32$*$32)$*$512	(32$*$32)$*$1024	0.855	0.854
Resnet152	(64$*$64)$*$128	(64$*$64)$*$1024	0.856	0.848
Wide Resnet50	(32$*$32)$*$512	(32$*$32)$*$1024	0.852	0.854
Wide Resnet50	(64$*$64)$*$128	(64$*$64)$*$1024	0.848	0.841

Table 2: Comparison of different backbones.

FPN	Mean Pooling	Max Pooling	Min Pooling	PLCC	SRCC
	✓			0.762	0.760
✓	✓			0.821	0.823
✓	✓	✓		0.857	0.845
✓	✓	✓	✓	0.837	0.834

Table 3: Ablation study of FPN and Feature Pooling Block on PIPAL validation set.

	Train patch detail	Test patch detail		
Backbone	(patch size) $*$ number of patches	(patch size) $*$ number of patches	PLCC	SRCC
Resnet50	(16$*$16)$*$2048	(16$*$16)$*$2048	0.832	0.823
Resnet50	(32$*$32)$*$512	(32$*$32)$*$512	**0.860**	**0.855**
Resnet50	(64$*$64)$*$128	(64$*$64)$*$128	0.857	0.845

Table 4: Ablation study of patch size on PIPAL validation set.

4.6. Ablation Study

In this section, we discuss the ablation studies of the effectiveness of the FPN and the Feature Pooling Block, patch size, and network backbones on PIPAL. Note that except for the ablation study on the backbone, other experiments are conducted using the ResNet50 backbone.

FPN. The IQMA with FPN builds the top-down pathway for feature activations output at each stage's last residual block. And a model without FPN is used to verify the effect of the FPN. As shown in row 1 and row 2 in Tab. 3, FPN improves the PLCC metric from 0.762 to 0.821, improves SRCC metric from 0.760 to 0.823, which validate the effectiveness of FPN.

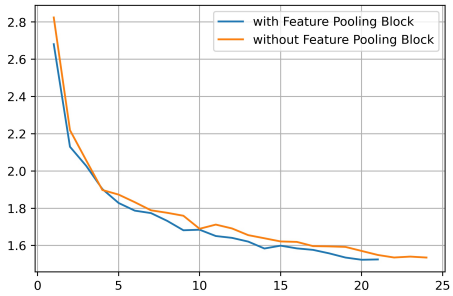

Figure 7: Convergence performance

Feature Pooling Block. In Feature Pooling Block, we design feature pooling fusion operation instead of only use mean pooling after feature extraction. We mainly compare the following 3 feature pooling methods:

$$\textbf{Method 1} : F_{mean},$$
$$\textbf{Method 2} : concat(F_{mean}, F_{max}),$$
$$\textbf{Method 3} : concat(F_{mean}, F_{max}, F_{min}),$$

where $F_{mean}, F_{max}, F_{min}$ denote mean pooling, max pooling and min pooling respectively.

As is shown in Tab. 3, feature pooling fusion demonstrates a significant performance improvement. Meanwhile, comparing method 2 with method 3, we can see that min pooling deteriorates performances but still performs better than method 1, further proving the importance of the feature pooling fusion operation. Furthermore, as shown in Fig. 7, we find that feature pooling fusion accelerates the convergence of training loss by around 20%.

Patch. We conduct ablation studies on different cropping patch sizes. Especially, we keep the number of total pixels the same on each patch size level (2^9). As shown in Tab. 4, patch size of 16×16 performs poorly in all 3 patch sizes setting. The reason may be that a tiny patch size will lose too much global information. Meanwhile, performance of 32×32 patch size is better than 64×64 , achieving 0.860 in PLCC and 0.855 in SRCC.

Backbone. Table 2 provides the comparison results on different ResNet backbones. It is clear that there is no significant performance difference among different backbones with the same setting of patch. In other words, it is not necessary to make network architecture deeper or wider for the IQMA network.

4.7. NTIRE 21 IQA Challenge Report

In both validation and testing phases, we use ensemble learning with multi-models trained by different backbones and different patch sizes to prevent the occurrence of overfitting and under-fitting. The IQMA network has achieved the 1st place (1st on both SRCC and PLCC) in the validation phase, and the 2nd place (1st on SRCC and 2nd on PLCC) in the final testing phase. The final ranking of the competition in the testing phase is shown in Tab. 5.

Method	PLCC	SRCC	SCORE
1st	**0.7896**	0.7990	**1.5885**
IQMA	0.7803	**0.8009**	1.5811
3rd	0.7707	0.7918	1.5625
4th	0.7709	0.7770	1.5480
5th	0.7615	0.7703	1.5317

Table 5: Result of NTIRE 21 IQA Challenge

5. Conclusion

In this paper, we propose a bilateral-branch Image Quality Multi-scale Assessment network (IQMA network) for image quality assessment. IQMA network takes pairs of images as input, and the pair of images is processed by two branches in parallel. The model learns image scores at 4 scales, and averages the 4 scores as the final output. The IQMA network adopts a bilateral-branch architecture, with 3 components: Feature Extraction Module, 4 Feature Fusion Modules, and 4 Score Regression Modules. A combination of ResNet and FPN is used for feature extraction, which outputs multi-scale features. Feature Fusion Modules read pair of features for corresponding patches of a certain scale and output a compactly fused feature vector for the pair of patches. Each Score Regression Module learns to predict an image score based on all the features of the corresponding scale. 4 image scores are averaged to provide a robust prediction of the predicted IQA score. With data augmentation, IQMA network has been validated on NTIRE 21 IQA Challenge dataset, LIVE, and TID2013. It achieves outstanding results on all the benchmarks.

References

[1] Edward H Adelson, Charles H Anderson, James R Bergen, Peter J Burt, and Joan M Ogden. Pyramid methods in image processing. *RCA engineer*, 29(6):33–41, 1984. 2

[2] Jochen Antkowiak, TDF Jamal Baina, France Vittorio Baroncini, Noel Chateau, France FranceTelecom, Antonio Claudio França Pessoa, FUB Stephanie Colonnese, Italy Laura Contin, Jorge Caviedes, and France Philips. Final report from the video quality experts group on the validation of objective models of video quality assessment march 2000. 2000. 1, 6, 7

[3] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pages 850–865. Springer, 2016. 2

[4] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6228–6237, 2018. 1, 2

[5] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on image processing*, 27(1):206–219, 2017. 1, 2, 4, 5

[6] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority oversampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002. 3

[7] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018. 3

[8] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Comparison of full-reference image quality models for optimization of image processing systems. *International Journal of Computer Vision*, pages 1–24, 2021. 1, 6, 7

[9] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016. 3

[10] 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, 6, 7

[11] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009. 3

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[13] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010. 1

[14] Shuqiang Jiang, Weiqing Min, Linhu Liu, and Zhengdong Luo. Multi-scale multi-view deep feature aggregation for food recognition. *IEEE Transactions on Image Processing*, 29:265–276, 2019. 2

[15] Gu Jinjin, Cai Haoming, Chen Haoyu, Ye Xiaoxing, Jimmy S Ren, and Dong Chao. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In *European Conference on Computer Vision*, pages 633–651. Springer, 2020. 1, 4, 5

[16] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2, 4

[17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 3

[18] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158:1–16, 2017. 2

[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. 2, 5

[20] Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1808–1817, 2018. 1, 2, 6, 7

[21] Hamid R Sheikh. Image and video quality assessment research at live. *http://live. ece. utexas. edu/research/quality*, 2003. 2, 5

[22] Hamid R Sheikh and Alan C Bovik. Image information and visual quality. *IEEE Transactions on image processing*, 15(2):430–444, 2006. 2

[23] Hamid R Sheikh, Alan C Bovik, and Gustavo De Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on image processing*, 14(12):2117–2128, 2005. 2

[24] Abhinav Shrivastava, Rahul Sukthankar, Jitendra Malik, and Abhinav Gupta. Beyond skip connections: Top-down modulation for object detection. *arXiv preprint arXiv:1612.06851*, 2016. 2

[25] Zhou Wang, Alan C Bovik, and Ligang Lu. Why is image quality assessment so difficult? In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages IV–3313. IEEE, 2002. 1, 2

[26] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 2, 6, 7

[27] 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

[28] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 4

[29] Guangtao Zhai and Xiongkuo Min. Perceptual image quality assessment: a survey. *Science China Information Sciences*, 63:1–52, 2020. 6

[30] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 1, 6, 7

[31] Wei Zhou and Zhibo Chen. Deep multi-scale features learning for distorted image quality assessment. *arXiv preprint arXiv:2012.01980*, 2020. 3