Blindly Assess Image Quality in the Wild Guided by A Self-Adaptive Hyper Network

Shaolin Su†, Qingsen Yan†, Yu Zhu‡, Cheng Zhang, Xin Ge, Jinqiu Sun, Yanning Zhang *
School of Computer Science and Engineering, Northwestern Polytechnical University

https://github.com/SSL92/hyperIQA

Abstract

Blind image quality assessment (BIQA) for authentically distorted images has always been a challenging problem, since images captured in the wild include varies contents and diverse types of distortions. The vast majority of prior BIQA methods focus on how to predict synthetic image quality, but fail when applied to real-world distorted images. To deal with the challenge, we propose a self-adaptive hyper network architecture to blind assess image quality in the wild. We separate the IQA procedure into three stages including content understanding, perception rule learning and quality predicting. After extracting image semantics, perception rule is established adaptively by a hyper network, and then adopted by a quality prediction network. In our model, image quality can be estimated in a self-adaptive manner, thus generalizes well on diverse images captured in the wild. Experimental results verify that our approach not only outperforms the state-of-the-art methods on challenging authentic image databases but also achieves competing performances on synthetic image databases, though it is not explicitly designed for the synthetic task.

1. Introduction

The goal of image quality assessment (IQA) is to enable computers to perceive image quality like humans. In the past decades, huge efforts have been devoted and a variety of IQA methods have been proposed. Despite the success they have achieved for assessing laboratory generated synthetically distorted images, IQA for authentically distorted images remains a challenge. The challenge lies mainly in three aspects:

Firstly, IQA in the wild is limited to the field of blind IQA (BIQA) since there exists no access to a reference image. As widely accepted, the limitation of reference image has made BIQA the most difficult problem among the three IQA categories, *i.e.* full-reference IQA (FR-IQA), reduced-reference IQA (RR-IQA) and BIQA, also known as non-reference IQA (NR-IQA). Secondly, different from the common synthetic distortions (*e.g.* Gaussian blur, JPEG compression) added to the whole area of image, authentic distortions are more complicated. The captured images not only suffer from global uniform distortions (*e.g.* out of focus, low illumination), but also contain other kinds of non-uniform distortions (*e.g.* object moving, over lighting, ghosting) in local areas. As a result, algorithms are challenged to accurately capture both global and local distortions to merge them into a proper quality prediction. Thirdly, compared to synthetic IQA databases, image content variation, which is a typical challenge in IQA task, presents even more difficulty to authentic IQA databases. Existing synthetic IQA databases LIVE [34], TID2013 [32] and CSIQ [21] only include no more than 30 reference images as limited in the sense of image contents, while authentic IQA database LIVE Challenge [8] and KonIQ-10k [13] consists of 1169 and 10073 images containing differ-

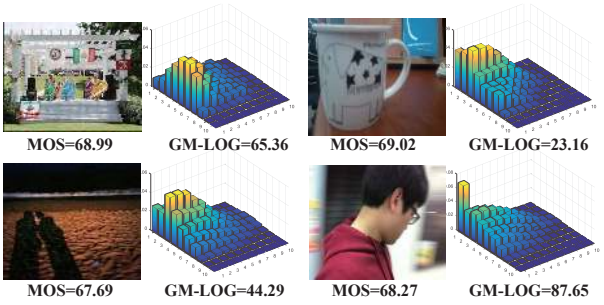

Figure 1. Images captured in the wild contain complex distortions and various contents, resulting in that extracted features differ from each other, though images showed above share similar quality scores. Top left: a synthetically distorted image taken from the LIVE database and GM-Log [38] features extracted, the rest images are taken from the authentic IQA database LIVE Challenge. MOS scores from two databases are aligned to the same scale.

*† The first two authors contributed equally to this work. This work was partially supported by NSFC (61871328, 61901384), ARC (DP160100703) and National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology. S. Su was supported by a scholarship from HUAWEI. ‡ Corresponding author: Yu Zhu.

ent contents respectively. This great content variation has raised a big challenge to the generalization ability of existing IQA methods.

Due to distortion diversity and content variation, IQA for authentically distorted images still has not been well solved. As shown in Figure 1, the extracted features vary when images vary, leading to inconsistent quality predictions with mean opinion score (MOS). In previous work, neither handcrafted feature based approaches nor networks with shallow architectures, which both solve synthetic IQA tasks well, are able in handling realistic distortions. This indicates low level features are not powerful enough in representing complicated distortion in real world. As a result, attempts have been made to use deep semantic features as quality descriptors: deep models which are pretrained on classification tasks are adopted to predict real world distortions. The hypothesis lying behind is that authentic distortions actually exist in photographically generated classification databases such as ImageNet [7], and these pre-trained features are already, to some extent, quality aware.

Although these attempts achieved promising improvements, further efforts are still lacked. Specifically, there are two drawbacks exist by simply adopting network architectures, which are initially designed for learning how to recognize objects, to the task of IQA. First, current deep models only learn global features for classification. For authentic IQA, however, distortions diverse in many ways, most of which exist in local areas. Ignoring local patterns may lead to an inconsistency between predicted quality and human visual perception, since human visual system (HVS) is sensitive to local distortions when the rest part of the image exhibits fairly good quality [21]. Secondly, as image content varies, the way human percepts quality of different objects varies. As illustrated in [22], an image of clear blue sky will be considered of high quality by human inspectors while mistaken by most of IQA methods to be a blurry image due to large flattened area the image contains. Therefore, directly predicting image quality before recognizing image content does not conform to the rule how humans perceive the world. In HVS, the top-down perception model indicates that human tries to understand the image before paying attention to other relevant sub-tasks such as quality assessment. However, in current models, fusing IQA task into semantic recognition network forces the network to learn image content and quality simultaneously, while it is more properly to let the network learn how to judge image quality after it has recognized the image content.

In this paper, we aim at developing an authentic IQA approach by considering the above two challenges which often appear at real world images: distortion diversity and content variation. We propose a local distortion aware module to extract local features from multi-scale to handle distortion diversity, and we introduce a hyper network architecture which dynamically generates weights for a quality prediction network to cover wide content variation. In our method, the proposed hyper network can adaptively learn the rule for perceiving quality according to its recognized content, and the target network follows this manner to give a final quality prediction. By judging quality based upon image content, the network is supposed to give predictions which are more consistent with human perception. In general, the main contributions of the proposed method can be summarized into three-folds:

- To enhance the ability of assessing image in the wild, we propose a novel IQA model based on hyper network which adaptively adjusts the quality prediction parameters. The proposed network predicts image quality in a content-aware manner, and the perception after recognition procedure is more consistent with the way how human realizes the world.
- Since local features are beneficial to handle non-uniform distortions in the image, we introduce a local distortion aware module to further capture image quality. We aggregate both local distortion features and global semantic features for gathering fine-grained details and holistic information, image quality is then predicted upon this multi-scale representation.
- Experimental results demonstrate that our approach not only outperforms the other competitors on authentic IQA databases, but also achieves competing results on synthetic IQA databases, despite we did not specifically design our model to extract synthetic features. This indicates the powerfulness and generalizability of our proposed model.

2. Related Work

2.1. IQA for Synthetically Distorted Images

In the past decades, great efforts have been put into the field of synthetic IQA, the approaches follow either of the two categories: hand-crafted feature based IQA and learning feature based IQA. Hand-crafted feature based approaches generally utilize NSS models to capture distortion. By modeling scene statistics which is sensitive to the appearance of distortion, degradation level of quality can be detected and quantified. These quality aware natural scene parameters include discrete wavelet coefficients [30], the correlation coefficients across subbands [1], DCT coefficients [33], locally normalized luminance coefficients with their pairwise products [29], image gradient, log-Gabor responses and color statistics [3]. Distribution models used to capture the statistics from synthetically distorted image include generalized Gaussian distribution (GGD) [29, 30], asymmetric generalized Gaussian distribution (AGGD) [3, 29], Weibull distribution [3], third order

polynomial tting [33] and histogram counting [38]. These hand-crafted features, however, require expertly design and are time-consuming. In addition, scene statistic features represent image quality from a global view, thus are not able to measure local distortions which commonly appear in authentically distorted images.

Inspired by the successes of machine learning in many computer vision tasks [9, 10, 39, 40], some learning based approaches are also proposed. In the early stage, codebook based learning approaches are introduced [37, 42, 43, 45]. Due to their strong learning power, CNN based methods are then proposed and achieved significant progress in synthetic IQA. In [14], a simple CNN with pooling strategy inherited from [43] is used for quality prediction. Ma *et al.* [27] proposed a deeper network to learn distortion type and image quality simultaneously. In [16, 23, 31], error map of the distorted image is learned to guide quality prediction, the approaches learning error map include training with residual error [16], with quality map calculated from FR-IQA methods [31] and with GAN generated image references [23]. Noticing the limited size of training data from existing IQA databases, [24] and [26] proposed to generate vast training samples by labeling their quality rank instead of quality score. Siamese network [5] and RankNet [4] architectures are used respectively to learn the rank of images.

Although these IQA methods have achieved great performance improvement on synthetic databases, challenge exists when facing large scale data [25, 28], indicating the problem of content variation still not well managed. It has also been shown that IQA models perform well on synthetic databases give inaccurate predictions on authentic IQA databases, suggesting the features of diverse distortion types exist in the wild can not be easily captured by architectures designed for extracting synthetic distortions.

2.2. IQA for Authentically Distorted Images

While most of the IQA models concentrate on synthetically distorted images, there are relatively few works focusing on the more challenging problem of authentic IQA. With the assistant of deep learning, deep semantic features are shown effective in representing image quality. In [17], Kim *et al.* showed that deep features from AlexNet [20] and ResNet [12] pretrained on classification databases such as ImageNet exhibit strong relationship with perceived quality and achieved standout accuracies. In [13], more pretrained baseline networks are tested, results confirmed the power of semantic features in solving IQA problem in the wild. In [46], a two stream network architecture is introduced to both predict synthetic and authentic image distortions. In their work, the authentic quality prediction stream adopted VGG-16 [35] for feature extraction. In [22], Li *et al.* proposed to use statistics from ResNet50 features of multi-patches for quality prediction. Recently, Zhang *et al.* [47] proposed to use image pairs both in synthetic and authentic databases for training IQA model, and the backbone used for feature extraction is ResNet-34. As can be seen, current models directly use output features from semantic learning networks for quality prediction, there are, however, mainly two drawbacks lying behind: first, mixing semantic learning and quality prediction in one network ignores how image semantics influence the way of quality perception, yet in HVS, image quality is judged after image content is recognized. Second, as deep semantic features are extracted from global scale, local distortions, which commonly exists in graphically obtained images, are ignored. As a result, networks are not able to capture detailed quality in an image, leading to inaccurate predictions.

In this work, we propose a novel multi-scale feature fused hyper network architecture to predict image quality in the wild. While previous models mix semantic understanding and quality prediction in one task, we divide the quality prediction procedure into two steps: image semantic features are learned first and quality is predicted based upon what content the image delivers. This procedure follows the top-down perception flow of humans, and we design a hyper network connection to mimic this mapping from image content to the manner of perceiving quality. In addition, instead of simply using global semantic features for content understanding, we also propose to fuse local distortion features from multi-scale to better represent image quality. In this way, our quality prediction procedure becomes self-adaptive, content-aware and capable of capture both detail and holistic information from the image.

3. Proposed Method

In this study, we aim at developing a quality assessment network which adaptively predicts image quality according to image content. The architecture of our network is shown in Figure 2. The proposed network consists of three parts: a backbone network which extracts semantic features, a target network which predicts image quality and a hyper network which generates a series of self-adaptive parameters for the target network. We will introduce our self-adaptive IQA model first and then present details of the three sub-networks in the following.

3.1. Self-Adaptive IQA Model

Traditional deep learning based quality prediction models receive an input image and directly map it to a quality score, the procedure can be described as follows:

$$\varphi(\boldsymbol{x}, \theta) = q, \tag{1}$$

where φ denotes the network model, $\boldsymbol{x}$ is the input image, θ represents the weight parameters. Note that once the training stage completes, weight parameters are fixed for all test

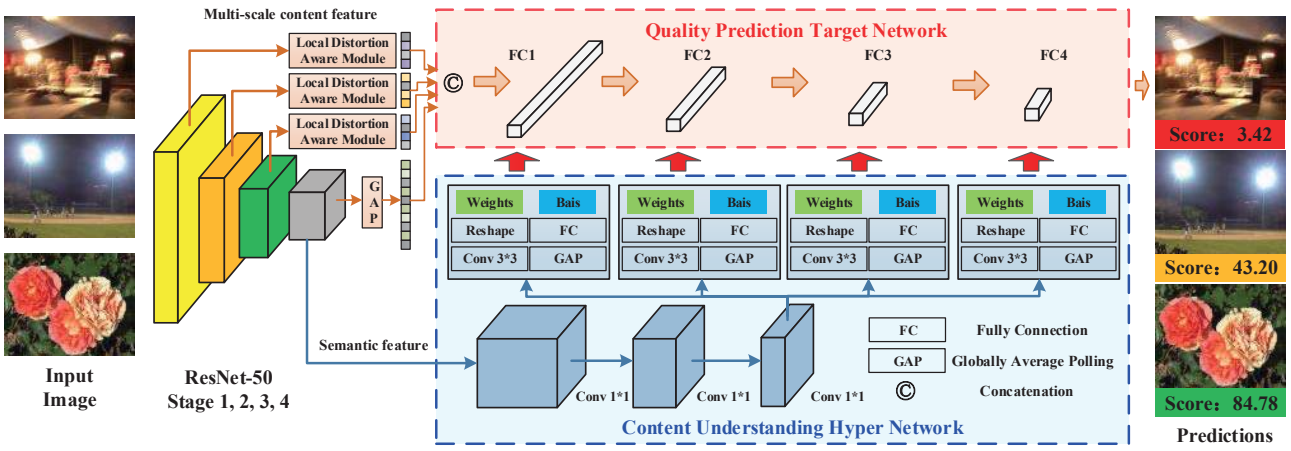

Figure 2. The pipeline of the proposed network. Given an image, we first extract semantic features from the basic model ResNet50, and import them to a hyper network which generates weights for a quality prediction target network. The input of the quality prediction target network is from aggregating multi-scale content features of the image, capturing both local and global distortions. In our module, the hyper network plays the role of formulating quality perception rule according to image content, and the target network makes quality prediction based on what an image specifically exhibits.

images. This prediction model implies that the same kind of quality features are extracted for predicting diverse images. In practical, however, as image contents vary, using the same rule for predicting varies images' quality is not thorough to cover their differently exhibited structures. As illustrated in [22], humans will take an image of clear blue sky as high quality, while for quality prediction models, this picture is most likely to be regarded as a blur contaminated one due to large flatten area it contains. The reason for this mistaken prediction is the ignorance of image semantic. For humans, under the condition of understanding image content, corresponding rules are then used to judge image quality. Therefore, to mimic the perception procedure of humans, we model the task of IQA as follows:

$$\varphi(\boldsymbol{x}, \theta_{\boldsymbol{x}}) = q, \tag{2}$$

where network parameters $\theta_{\boldsymbol{x}}$ are dependent on the image itself instead of being fixed for all inputs. For easier understanding, parameters $\theta_{\boldsymbol{x}}$ can be regarded as quality perceiving rules. As image content varies, the way of perceiving image quality varies. In this way, our IQA model becomes self-adaptive as it extracts different quality indicators with respect to different contents. Ideally, one can train images of the same content with an individual network for quality prediction with more flexibility, however, training a set of networks covering such widely spread contents is computation inefficient and not practical. Therefore, we introduce hyper network to simplify this problem:

$$\theta_{\boldsymbol{x}} = H(S(\boldsymbol{x}), \gamma), \tag{3}$$

where H stands for a hyper network mapping function and γ represents hyper network parameters. We define the input of hyper network as $S(\boldsymbol{x})$, meaning semantic features extracted from the input image $\boldsymbol{x}$. Thus the function of hyper network is to learn the mapping from image content to the rule of how to judge image quality. The learned perception rule will further guide our target network to extract self-adaptive quality features for prediction.

By introducing the intermediate variable $\theta_{\boldsymbol{x}}$ and hyper network, we actually divide the task of IQA into three steps: **semantic feature extraction**, **perception rule establishment** and **quality prediction**. We use a backbone network to extract image semantic features $S(\boldsymbol{x})$, a hyper network to learn the quality perception rule $\theta_{\boldsymbol{x}}$ and a quality prediction target network to obtain the final quality score q. Unlike the quality prediction model in Equation (1), where image quality is directly estimated without semantic understanding or content recognition, our proposed model follows the top-down perception mechanism as it tries to understand image in the first place, until when it executes the task of quality judgement. This designation makes our network more flexible in extracting quality influential factors when facing content varying images. In addition, the proposed quality prediction procedure is also more consistent with the way how human perceives image quality.

In order to reduce the amount of target network parameters $\theta_{\boldsymbol{x}}$ and also for easier training, we simplify the input of the target network to a content aware vector $\boldsymbol{v}_{\boldsymbol{x}} = S_{ms}(\boldsymbol{x})$, where S_{ms} stands for the meaning that the content aware vector is also extracted by the backbone semantic extraction network, but fuses multi-scale features to capture local distortions in the image. Under this alteration, the whole

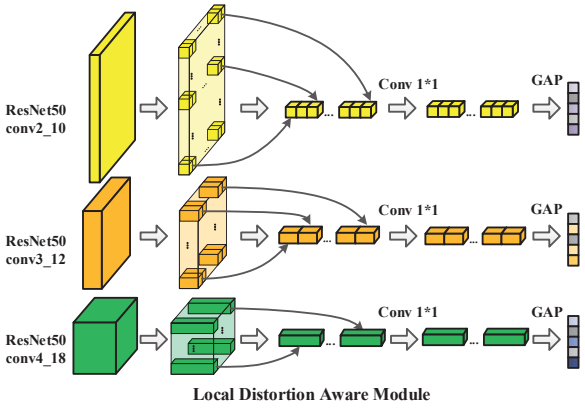

Figure 3. The architecture of the proposed local distortion aware module.

hyper network based IQA model can be described as:

$$\varphi(\boldsymbol{v_x}, H(S(\boldsymbol{x}), \gamma)) = q. \tag{4}$$

Based on the quality prediction model, we then present the architecture of the three sub-networks in the following.

3.2. Semantic Feature Extraction Network

As shown in Figure 2, the front part of our network architecture is a common semantic feature extraction network. The semantic extraction network focuses on understanding image content, and outputs two streams of features for quality prediction. The semantic feature $S(\boldsymbol{x})$ is directly fed to hyper network for weight generation, and the multi-scale content feature stream $S_{ms}(\boldsymbol{x})$ is treated as the input of the target network. The reason why we extract multi-scale content features is that semantic features extracted from the last layer merely represent holistic image content. In order to capture local distortions in real world, we propose to extract multi-scale features through a local distortion aware module. As illustrated in Figure 3, our designed local distortion aware module consists of a series of operations including dividing multi-scale feature maps into non-overlapping patches, stacking the patches along the channel dimension, conducting 1×1 convolution and globally average pooling them into vectors. The proposed module can be regarded to serve as an attention based patch extractor, which is aware of feature patches corresponding to local distortions for better capturing its quality.

Specifically, we use ResNet50 [12] as the backbone model for semantic feature extraction. The pretrained model on ImageNet [7] is used for network initialization. In our network, the last two layers of the origin ResNet50, *i.e.* an average pooling layer and a fully connected layer are removed to output feature stream. We extract multi-scale features from conv2_10, conv3_12, conv4_18 layers as the input to the local distortion aware module, which outputs multi-scale content vector $\boldsymbol{v_x}$.

3.3. Hyper Network for Learning Perception Rule

Inspired by [19], our hyper network consists of three 1×1 convolution layers and several weight generating branches. Since in the proposed network, fully connected layers are used as basic target network component (see Section 3.4), two types of network parameters, *i.e.* fully connected layer weights and biases, should be generated. For different types of parameters, we use different weight generating approaches. Fully connected layer weights are generated from convolution followed by reshape operation of extracted features, while fully connected layer biases are generated by simply average pooling and fully connection, as bias weights have much less amount of parameters. The output channels of convolution and fully connected layers are decided based upon dimensions of corresponding layers in target network for size match. The generated weights are regarded as the rule of perceiving image quality and will further instruct target network for predicting image quality.

3.4. Target Network for Quality Prediction

As multi-scale features extracted by the semantic extraction network are content-aware, the function of the target network is simply mapping learned image contents to a quality score. Therefore, we use a small and simple network for quality prediction. As shown in Figure 2, our target network consists of four fully connected layers, it receives multi-scale content feature vector as input, and propagates through weight determined layers to get the final quality score. In the target network, we choose sigmoid function as the activation function.

3.5. Implementation Details

We implemented our model by PyTorch and conducted training and testing on the NVIDIA 1080Ti GPUs. Following the training strategy from [17], we randomly sample and horizontally flipping 25 patches with size 224×224 pixels from each training image for augmentation. Training patches inherited quality scores from the source image, and we minimize L_1 loss over the training set:

$$\ell = \frac{1}{N}\sum_{i}^{N} ||\varphi(\boldsymbol{v_{p_i}}, H(S(\boldsymbol{p_i}), \gamma)) - Q_i||_1, \tag{5}$$

where $\boldsymbol{p_i}$ and Q_i refers to the i-th training patch and the ground truth score, respectively. We used Adam [18] optimizer with weight decay 5×10^{-4} to train our model for 15 epochs, with mini-batch size of 96. Learning rate is first set to 2×10^{-5}, and reduced by 10 after every 5 epochs. For faster convergence, un-pretrained layers of our model, which are Xavier initialized, applied learning rate 10 times larger. During testing stage, 25 patches with 224×224 pixels from test image are randomly sampled and their cor-

responding prediction scores are average pooled to get the final quality score.

4. Experiments

4.1. Datasets

We used three authentically distorted image databases including LIVE Challenge (LIVEC) [8], KonIQ-10k [13] and BID [6] for evaluation. LIVEC contains 1162 images taken from different photographers with varies camera devices in the real world, hence these images contain complex and composite distortions. KonIQ-10k consists of 10073 images which are selected from the large public multimedia database YFCC100m [36], the sampled images try to cover a wide and uniform quality distribution in the sense of brightness, colorfulness, contrast and sharpness. BID is a blur image database containing 586 images with realistic blur distortions such as motion blur and out of focus, *etc*.

Except for authentic image databases, we also tested our model on synthetic image databases LIVE [34] and CSIQ [21]. There are 779 and 866 synthetically distorted images included in each database.

4.2. Evaluation Metrics

Two commonly used criteria, Spearman's rank order correlation coefficient (SRCC) and Pearson's linear correlation coefficient (PLCC) are adopted to measure prediction monotonicity and prediction accuracy. The two criteria both range from 0 to 1 and a higher value indicates better performance. Before calculating PLCC, logistic regression is first applied to remove nonlinear rating caused by human visual observation, as suggested in the report from Video Quality Expert Group (VQEG) [11].

For each database, 80% images are used for training and the rest 20% are used for testing. For synthetic image databases LIVE and CSIQ, the split is implemented according to reference images to avoid content overlapping. We run 10 times of this random train-test splitting operation and the median SRCC and PLCC values are reported.

4.3. Comparison with the State-of-the-art Methods

Eight state-of-the-art BIQA methods are selected for performance comparison. The comparison methods including hand-crafted based approaches [3, 29, 37], deep learning based synthetic IQA approaches [2, 15] and deep learning based authentic IQA approaches [22, 44, 46].

Single database evaluations. We first analyze experimental results on single databases. As shown in Table 1, our approach outperforms all the state-of-the-art methods on all three authentic image databases (LIVEC, BID and KonIQ-10k) for both SRCC and PLCC evaluations. This suggests that learning image content firstly assists in perceiving image quality, when image data covers a wide range of variety.

Table 1. Overall performance evaluation on five image databases.

SRCC	LIVEC	BID	KonIQ	LIVE	CSIQ
BRISQUE [29]	0.608	0.562	0.665	0.939	0.746
ILNIQE [3]	0.432	0.516	0.507	0.902	0.806
HOSA [37]	0.640	0.721	0.671	0.946	0.741
BIECON [15]	0.595	0.539	0.618	0.961	0.815
WaDIQaM [2]	0.671	0.725	0.797	0.954	**0.955**
SFA [22]	0.812	0.826	0.856	0.883	0.796
PQR [44]	0.857	0.775	0.880	0.965	0.873
DBCNN [46]	0.851	0.845	0.875	**0.968**	0.946
Ours	**0.859**	**0.869**	**0.906**	0.962	0.923
PLCC	LIVEC	BID	KonIQ	LIVE	CSIQ
BRISQUE [29]	0.629	0.593	0.681	0.935	0.829
ILNIQE [3]	0.508	0.554	0.523	0.865	0.808
HOSA [37]	0.678	0.736	0.694	0.947	0.823
BIECON [15]	0.613	0.576	0.651	0.962	0.823
WaDIQaM [2]	0.680	0.742	0.805	0.963	**0.973**
SFA [22]	0.833	0.840	0.872	0.895	0.818
PQR [44]	0.882	0.794	0.884	0.971	0.901
DBCNN [46]	0.869	0.859	0.884	**0.971**	0.959
Ours	**0.882**	**0.878**	**0.917**	0.966	0.942

Though we did not especially add modules for synthetic image feature extraction, our approach achieved competing performance with the state-of-the-art methods on two synthetic image databases LIVE and CSIQ. Note that compared with PQR and SFA, which also utilize backbone classification networks to extract deep semantic features, our approach significantly outperforms PQR on CSIQ database and outperforms SFA on both LIVE and CSIQ dataset.

We further present performance comparison of our approach on individual distortion types. Since distortion types are of high diversity on authentic image databases, we only evaluate the performance on synthetic image databases LIVE and CSIQ, as shown in Table 2. Compared with other methods which introduce specific module to handle synthetic IQA task, our proposed method uses a simply network to obtain competing performances on individual distortion types. This proved that the effectiveness of image content understanding based IQA method.

Generalization ability test. We first run cross database tests for performance comparison, the tests are conducted on intra databases belonging to either authentic or synthetic distortions. We select the most competing two approaches PQR and DBCNN for comparison, and the results are shown in Table 3. Among six authentic cross database tests, our approach achieves four times of top performance. For synthetic cross database tests, our approach still performs competitively to other algorithms, indicating the strong generalization power of our approach.

To further evaluate the generalization ability of our approach, we train competing models on the whole LIVE

Table 2. SRCC comparisons on individual distortion types on the LIVE and CSIQ databases.

Database	LIVE					CSIQ					
Type	JP2K	JPEG	WN	GB	FF	WN	JPEG	JP2K	FN	GB	CC
BRISQUE [29]	0.929	0.965	**0.982**	**0.964**	0.828	0.723	0.806	0.840	0.378	0.820	0.804
ILNIQE [3]	0.894	0.941	0.981	0.915	0.833	0.850	0.899	0.906	0.874	0.858	0.501
HOSA [37]	0.935	0.954	0.975	0.954	**0.954**	0.604	0.733	0.818	0.500	0.841	0.716
BIECON [15]	0.952	**0.974**	0.980	0.956	0.923	0.902	**0.942**	0.954	0.884	0.946	0.523
WaDIQaM [2]	0.942	0.953	**0.982**	0.938	0.923	**0.974**	0.853	0.947	0.882	**0.979**	**0.923**
PQR [44]	0.953	0.965	0.981	0.944	0.921	0.915	0.934	0.955	0.926	0.921	0.837
DBCNN [46]	**0.955**	0.972	0.980	0.935	0.930	0.948	0.940	0.953	**0.940**	0.947	0.870
Ours	0.949	0.961	**0.982**	0.926	0.934	0.927	0.934	**0.960**	0.931	0.915	0.874

Table 3. SRCC evaluations on cross database tests.

Training	Testing	PQR	DBCNN	Ours
LIVEC	BID	0.714	**0.762**	0.756
	KonIQ	0.757	0.754	**0.772**
BID	LIVEC	0.680	0.725	**0.770**
	KonIQ	0.636	**0.724**	0.688
KonIQ	LIVEC	0.770	0.755	**0.785**
	BID	0.755	0.816	**0.819**
LIVE	CSIQ	0.719	**0.758**	0.744
CSIQ	LIVE	0.922	0.877	**0.926**

Table 4. D-Test, L-Test and P-Test results on the Waterloo Exploration Database.

Model	D-Test	L-Test	P-Test
BRISQUE [29]	0.9204	0.9772	0.9930
GM-Log [38]	0.9203	0.9106	0.9748
CORNIA [43]	0.9290	0.9764	0.9947
HOSA [37]	0.9175	0.9647	0.9983
dipIQ [26]	0.9346	**0.9846**	**0.9999**
deepIQA [2]	0.9074	0.9467	0.9628
MEON [27]	0.9384	0.9669	0.9984
Two Stream CNN [41]	0.9301	0.9765	0.9952
DB-CNN [46]	**0.9387**	0.9527	0.9984
Ours	0.9006	0.9747	0.9971

Dataset and test them on the large scale synthetic database Waterloo Exploration Database [25]. Firstly, three testing criteria, D-Test, L-Test and P-Test are calculated, which respectively measure pristine-distortion discriminability, consistency with distortion levels and pairwise quality discriminability. As shown in Table 4, our approach achieves competing performance, though it is not specifically designed for synthetically distorted IQA.

Then, we conducted gMAD competition [28] on the Waterloo Database for a direct visualization. gMAD efficiently selects image pairs with maximum quality difference predicted by an attacking IQA model to challenge an other defending model which considers them are of the same level of quality. The selected pairs are shown to the observer to determine whether the attacker or the defender is robust. In Figure 4, we fix our model as a defender in the first two columns, image pairs selected from a bad quality level and a good quality level are presented respectively. In the last two columns, our model attacks other competing methods where each column represents images selected from a bad and a good quality level predicted by the defender.

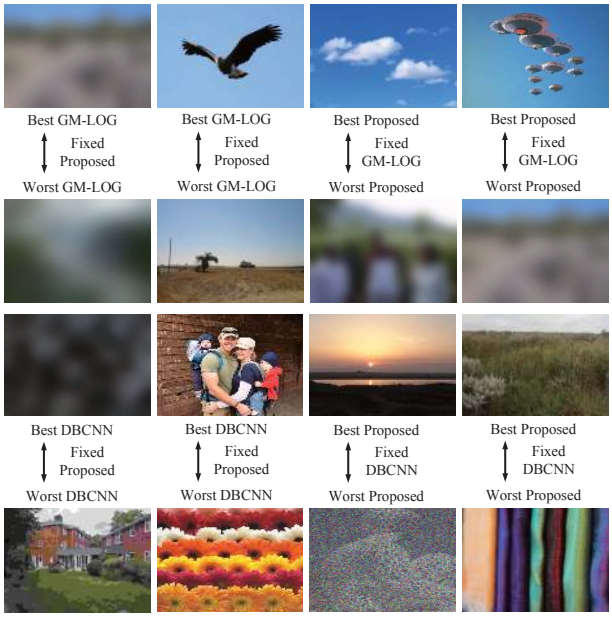

Figure 4. gMAD competition results on the Waterloo Database against GM-LOG [38] and DBCNN [46].

As can be seen from Figure 4, when our model plays as defender, image pairs selected by the attacker do not vary much in perceived quality, while our model successively selects image pairs with huge quality difference when acts as attacker. This indicates our model is both powerful in defending and attacking. In addition, it is worth mentioning that our model successfully recognizes high quality im-

ages with flatten contents, despite they deceive the defending models to have low quality (the image "sky" and the image "sunset" on the third column). These results further demonstrate that our proposed model has a strong generalization ability regarding the challenge of content variation. gMAD results against more IQA models can be found in our supplementary material.

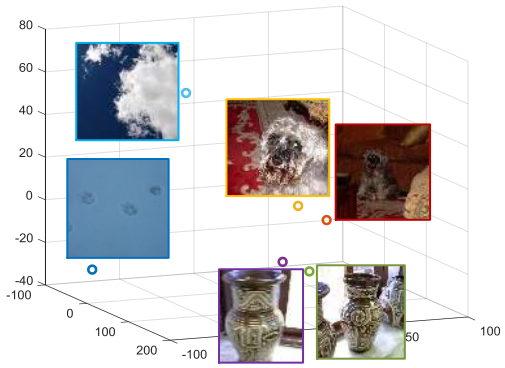

Figure 5. Generated weights of different images are plotted in the 3D space after PCA transformation. This figure shows the weights extracted from the first layer of the target network, weights from other layers also exhibits similar distribution.

4.4. Visualization of Self-Adaptive Weights

In order to verify the effectiveness of the weight generating procedure, we extract generated weights of the quality prediction network from several images of varies contents. We then do PCA transform to the weights and plot them in a 3D space for visualization. In Figure 5, we plot transformed weights from the first layer of the target network, weights from other layers also show similar characteristics. From Figure 5, several interesting findings can be discovered:

First, for images of different contents, the generated weights vary. This indicates our network adopts distinct weights for evaluating image quality in a self-adaptive manner. Whereas for traditional IQA models, weights of the model are fixed for all input images, which will coincide in the same location in the weight space if we plot them.

Second, images of the same object generate similar weights despite they exhibit distinct levels of quality. As can be seen from Figure 5, though their quality varies, images of the same class "dog" or "vase" generate similar weights for quality prediction. This verifies that our model successfully learns image contents to instruct quality prediction. We believe this predicting after understanding scheme makes our model self-adaptive, thus is able to evaluate image quality more flexibly and more precisely when facing the challenge from a large diversity of images.

Third, for flatten images "snow footprint" and "sky", the corresponding weights distinguish from each other. This suggests our network indeed learns to understand high-level image content though they exhibit similar low-level quality indicators such as smoothness. Therefore, our model is prevented from mistaking image quality due to content variation, such as confusing a flatten image with blurriness or mistaking an image abundant of textures to a noisy one.

Table 5. Ablation results on LIVE Challenge and LIVE databases.

Components	LIVE Challenge		LIVE	
	SRCC	PLCC	SRCC	PLCC
Res50	0.827	0.852	0.923	0.947
Res50+MS	0.836	0.859	0.954	0.963
Res50+Hyp	0.854	0.879	0.944	0.959
Res50+MS+Hyp	**0.859**	**0.882**	**0.962**	**0.966**

4.5. Ablation Study

To evaluate the efficiency of our proposed components, we conduct several ablation experiments on the LIVEC and LIVE database. We first use a pretrained ResNet50 with fine-tuning as our backbone model and analyze the effect of each individual component by comparing both SRCCs and PLCCs. The results are shown in Table 5.

We first examine the effectiveness of our proposed local distortion aware module by concatenating them with ResNet50 output features (Res50+MS). The SRCC slightly improved on the LIVE Challenge Database and obviously improved on the LIVE Database with around 1.6% and 3%.

Then, we add hyper network and target network module to the backbone network. The input and weights of the target network are both from the last feature layer of ResNet50. By modifying the hyper network to our proposed architecture, we can see major SRCC and PLCC improvements on both LIVE Challenge and LIVE databases. On LIVE Challenge, SRCC and PLCC increased both 2.7% and on LIVE, they increase 2.1% and 1.2% respectively.

At last, we add multi-scale features to the target network's input, and SRCC and PLCC further improved to the highest value of 85.9%, 88.2% on the LIVE Challenge Database and 96.2% and 96.6% on the LIVE Database.

5. Conclusion

In this paper, we propose a novel network to overcome two challenging problems that appear in the task of authentic IQA: distortion diversity and content variation. The proposed network separates quality prediction from content understanding to mimic how humans perceive image quality. We employ hyper network architecture to accomplish this perception flow, and further introduce a multi-scale local distortion aware module to capture complex distortions. Experimental results showed that our proposed approach possesses strong generalization ability which offers the prospect of more extensive applications of IQA task.

References

[1] Moorthy Anush Krishna and Bovik Alan Conrad. Blind image quality assessment: from natural scene statistics to perceptual quality. *IEEE Transactions on Image Processing*, 20(12):3350–3364, 2011.

[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. *IEEE Transactions on Image Processing*, 27(1):206–219, 2017.

[3] Alan C Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015.

[4] Christopher Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Gregory N Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine learning (ICML-05)*, pages 89–96, 2005.

[5] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005.

[6] Alexandre Ciancio, André Luiz N Targino da Costa, Eduardo AB da Silva, Amir Said, Ramin Samadani, and Pere Obrador. No-reference blur assessment of digital pictures based on multifeature classifiers. *IEEE Transactions on image processing*, 20(1):64–75, 2010.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[8] Deepti Ghadiyaram and Alan C Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, 2015.

[9] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *IEEE International Conference on Computer Vision*, pages 1705–1714, 2019.

[10] Dong Gong, Jie Yang, Lingqiao Liu, Yanning Zhang, Ian Reid, Chunhua Shen, Anton Van Den Hengel, and Qinfeng Shi. From motion blur to motion flow: A deep learning solution for removing heterogeneous motion blur. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3806–3815, 2016.

[11] Video Quality Experts Group et al. Final report from the video quality experts group on the validation of objective models of video quality assessment. In *VQEG meeting, Ottawa, Canada, March, 2000*, 2000.

[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.

[13] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *arXiv preprint arXiv:1910.06180*, 2019.

[14] Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1733–1740, 2014.

[15] Jongyoo Kim and Sanghoon Lee. Fully deep blind image quality predictor. *IEEE Journal of selected topics in signal processing*, 11(1):206–220, 2016.

[16] Jongyoo Kim, Anh-Duc Nguyen, and Sanghoon Lee. Deep cnn-based blind image quality predictor. *IEEE transactions on neural networks and learning systems*, 30(1):11–24, 2018.

[17] Jongyoo Kim, Hui Zeng, Deepti Ghadiyaram, Sanghoon Lee, Lei Zhang, and Alan C Bovik. Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment. *IEEE Signal Processing Magazine*, 34(6):130–141, 2017.

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[19] Sylwester Klocek, Łukasz Maziarka, Maciej Wołczyk, Jacek Tabor, Jakub Nowak, and Marek Śmieja. Hypernetwork functional image representation. In *International Conference on Artificial Neural Networks*, pages 496–510. Springer, 2019.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[21] 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.

[22] Dingquan Li, Tingting Jiang, Weisi Lin, and Ming Jiang. Which has better visual quality: The clear blue sky or a blurry animal? *IEEE Transactions on Multimedia*, 21(5):1221–1234, 2018.

[23] Kwan-Yee Lin and Guanxiang Wang. Hallucinated-iqa: No-reference image quality assessment via adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 732–741, 2018.

[24] Xialei Liu, Joost van de Weijer, and Andrew D Bagdanov. Rankiqa: Learning from rankings for no-reference image quality assessment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1040–1049, 2017.

[25] Kede Ma, Zhengfang Duanmu, Qingbo Wu, Zhou Wang, Hongwei Yong, Hongliang Li, and Lei Zhang. Waterloo exploration database: New challenges for image quality assessment models. *IEEE Transactions on Image Processing*, 26(2):1004–1016, 2016.

[26] Kede Ma, Wentao Liu, Tongliang Liu, Zhou Wang, and Dacheng Tao. dipiq: Blind image quality assessment by learning-to-rank discriminable image pairs. *IEEE Transactions on Image Processing*, 26(8):3951–3964, 2017.

[27] Kede Ma, Wentao Liu, Kai Zhang, Zhengfang Duanmu, Zhou Wang, and Wangmeng Zuo. End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing*, 27(3):1202–1213, 2017.

[28] Kede Ma, Qingbo Wu, Zhou Wang, Zhengfang Duanmu, Hongwei Yong, Hongliang Li, and Lei Zhang. Group mad competition-a new methodology to compare objective image quality models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1664–1673, 2016.

[29] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012.

[30] Anush Krishna Moorthy and Alan Conrad Bovik. A two-step framework for constructing blind image quality indices. *IEEE Signal processing letters*, 17(5):513–516, 2010.

[31] Da Pan, Ping Shi, Ming Hou, Zefeng Ying, Sizhe Fu, and Yuan Zhang. Blind predicting similar quality map for image quality assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6373–6382, 2018.

[32] Nikolay Ponomarenko, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Lina Jin, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, and Federica Battisti. Color image database tid2013: Peculiarities and preliminary results. In *European Workshop on Visual Information Processing*, 2013.

[33] Michele A Saad, Alan C Bovik, and Charrier Christophe. Blind image quality assessment: a natural scene statistics approach in the dct domain. *IEEE Transactions on Image Processing*, 21(8):3339, 2012.

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

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[36] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *arXiv preprint arXiv:1503.01817*, 2015.

[37] Jingtao Xu, Peng Ye, Qiaohong Li, Haiqing Du, Yong Liu, and David Doermann. Blind image quality assessment based on high order statistics aggregation. *IEEE Transactions on Image Processing*, 25(9):4444–4457, 2016.

[38] Wufeng Xue, Xuanqin Mou, Lei Zhang, Alan C Bovik, and Xiangchu Feng. Blind image quality assessment using joint statistics of gradient magnitude and laplacian features. *IEEE Transactions on Image Processing*, 23(11):4850–4862, 2014.

[39] Qingsen Yan, Dong Gong, Qinfeng Shi, Anton van den Hengel, Chunhua Shen, Ian Reid, and Yanning Zhang. Attention-guided network for ghost-free high dynamic range imaging. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1751–1760, 2019.

[40] Qingsen Yan, Dong Gong, Pingping Zhang, Qinfeng Shi, Jinqiu Sun, Ian Reid, and Yanning Zhang. Multi-scale dense networks for deep high dynamic range imaging. In *IEEE Winter Conference on Applications of Computer Vision*, pages 41–50, Jan 2019.

[41] Qingsen Yan, Dong Gong, and Yanning Zhang. Two-stream convolutional networks for blind image quality assessment. *IEEE Transactions on Image Processing*, 28(5):2200–2211, 2018.

[42] Peng Ye and David Doermann. No-reference image quality assessment using visual codebooks. *IEEE Transactions on Image Processing*, 21(7):3129–3138, 2012.

[43] Peng Ye, Jayant Kumar, Le Kang, and David Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1098–1105. IEEE, 2012.

[44] Hui Zeng, Lei Zhang, and Alan C Bovik. A probabilistic quality representation approach to deep blind image quality prediction. *arXiv preprint arXiv:1708.08190*, 2017.

[45] Lin Zhang, Zhongyi Gu, Xiaoxu Liu, Hongyu Li, and Jianwei Lu. Training quality-aware filters for no-reference image quality assessment. *IEEE MultiMedia*, 21(4):67–75, 2014.

[46] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[47] Weixia Zhang, Kede Ma, and Xiaokang Yang. Learning to blindly assess image quality in the laboratory and wild. *arXiv preprint arXiv:1907.00516*, 2019.