Journal of Multimedia Information System VOL. 9, NO. 2, June 2022 (pp. 75-86): ISSN 2383-7632 (Online)
https://doi.org/10.33851/JMIS.2022.9.2.75

No-Reference Image Quality Assessment based on Quality Awareness Feature and Multi-task Training

Lijing Lai[1], Jun Chu[1*], Lu Leng[1]

Abstract

The existing image quality assessment (IQA) datasets have a small number of samples. Some methods based on transfer learning or data augmentation cannot make good use of image quality-related features. A No Reference (NR)-IQA method based on multi-task training and quality awareness is proposed. First, single or multiple distortion types and levels are imposed on the original image, and different strategies are used to augment different types of distortion datasets. With the idea of weak supervision, we use the Full Reference (FR)-IQA methods to obtain the pseudo-score label of the generated image. Then, we combine the classification information of the distortion type, level, and the information of the image quality score. The ResNet50 network is trained in the pre-train stage on the augmented dataset to obtain more quality-aware pre-training weights. Finally, the fine-tuning stage training is performed on the target IQA dataset using the quality-aware weights to predicate the final prediction score. Various experiments designed on the synthetic distortions and authentic distortions datasets (LIVE, CSIQ, TID2013, LIVEC, KonIQ-10K) prove that the proposed method can utilize the image quality-related features better than the method using only single-task training. The extracted quality-aware features improve the accuracy of the model.

Key Words: Deep Learning, No-Reference Image Quality Assessment, Multiple Task Learning, Score Prediction.

I. INTRODUCTION

In the era of social media, digital images are everywhere. Image quality assessment has become more critical and derived many application scenarios. In daily use, the quality of media images and photographic images can be assessed. In computer vision tasks, the quality of the generated images in image enhancement [1], image super-resolution [2], and image restoration tasks [3-4] can be assessed. In industrial applications such as detection and recognition tasks, high-quality images are screened through image quality assessment to improve the stability of the application system.

Due to the great success of deep convolutional neural networks in image classification [5], object detection [6], object tracking [7], and other computer vision tasks, researchers have also begun to introduce them into the field of no-reference image quality assessment [8], becoming the mainstream method design thinking. However, the design of datasets in image quality assessment is time-consuming, labor-intensive, and expensive, which cause the number of images in the image quality assessment dataset to be small. The application of deep convolutional networks in image quality assessment is suffered from the network overfitting.

Many researchers use the patch-based method [9] to solve the problem of overfitting, such as dividing the image into multiple patches of size 32×32, which are used as network input to increase training samples. Many others addressing this problem with pretrained strategies [10].

There are plenty new topics emerged in the IQA task. For example, handling IQA with Transformers [11]. Evaluation of the generated image is also a hot spot in the IQA task. [12] proposed the 2021 IQA challenge on the newly PIPAL dataset [13]. The PIPAL dataset includes various types of GAN-generated images in image restoration (deraining and dehazing) tasks, image enhancement tasks and image restoration tasks.

In this work, we mainly focus on and solve the problems in data augmentation-based methods. These methods directly research the data itself, augment the dataset by applying distortion to high-quality original images.

RankIQA [14] designed a strategy to generate large-scale distorted images without laborious human labeling. According to the law that image quality decreases with increasing distortion levels, they synthetically generate ranked image pairs with different distortion levels from the Waterloo Exploration dataset [15]. A Siamese network is pretrained using pairs of sorted images. Finally, they used a

Manuscript received May 19, 2022; Revised May 27, 2022; Accepted May 30, 2022. (ID No. JMIS-22M-05-018)
Corresponding Author (*): Jun Chu, +86-0797-83953414, chujun99602@163.com
[1]School of Software, Nanchang Hangkong University, Nanchang, China, lai_lj@foxmail.com, chujun99602@163.com, leng@nchu.edu.cn

branch of the Siamese network to predict image scores, aiming to convert image distortion levels into quality scores. The limitation of the rank method is that it can only simulate distorted images in synthetic IQA datasets, and it is not easy to apply this method to authentic IQA datasets.

DB-CNN [16] uses two large datasets: the Waterloo Exploration dataset and PASCAL VOC 2012 [17], to generate distorted images. For the final augmented dataset, its labels contain vectors encoding the distortion type and distortion level. Then, they designed a shallow CNN for synthesizing distorted images. Chose a pretrained VGG-16 network for the classification task on ImageNet as another branch to extract relevant features of authentically distorted images. Because distortion in ImageNet is a natural consequence of photography rather than simulation. They combined a shallow CNN for synthetic distortion and VGG-16 for natural distortion into one model and designed a new pooling strategy to calculate the final quality score.

RankIQA [14] used augmented data pair-wise ranking information, and DB-CNN [16] used image distortion type and level information. Unlike methods that directly use the ImageNet dataset to pre-train weights for transfer learning, these augmentation-based methods considered the difference between the samples in the IQA and the ImageNet datasets. Their pre-training network can better extract features related to image quality assessment tasks (quality-aware features) and achieve good results.

However, due to the lack of Mean Opinion Score (MOS) labels of augmented images, they pre-trained a network in a single-task learning strategy, using rank information or distortion type and level information. Quality score--the main target of the IQA is underutilization. The pre-trained weights exist differences between these single-task learning strategies and image quality score prediction. The model ability of extracted quality-aware features can be improved.

We propose a no-reference image quality assessment method based on quality-aware feature learning and multi-task training. To make better use of image quality-related attributes, the idea of weak supervision learning is applied in the dataset augmentation. Several full-reference methods are used to obtain the quality scores of images in the pre-training set, and we called them pseudo-quality scores (Pseudo-MOS, PMOS). Then we apply the multi-task training strategy, take the score prediction as the main task, the distortion type and level classification as auxiliary tasks. The multi-task training makes the pre-trained network extract quality-aware features better. Finally, use the quality-aware weights to initialize the network and fine-tune on the target IQA dataset. Performance on three synthetic distortion datasets and two authentic distortion datasets proved that the proposed method makes better use of image quality-related attributes than methods that only use single-task training. The extracted quality-aware features improve the model's accuracy beyond the current mainstream methods.

The primary contributions of this study are: 1) The method proposed in this paper comprehensively utilizes three attributes related to the IQA task, the distortion type, distortion level, and quality score. Extract more quality-aware features and predict more accurate predictions. 2) The synthetic and authentic distortion datasets are augmented using different strategies. Combined with the FR-IQA method, a reliable pseudo-score label is calculated for the synthetic images. The problem of lacking quality score labels of the augmented data set is solved. 3) Using a multi-task training strategy, comprehensively utilize the distortion type, distortion level, and quality score information of the augmented image, and perform feature fusion in the head of the network, so that the network can extract quality-aware features.

II. DATA AUGMENTATION AND PSEUDO-LABEL

Data augmentation can alleviate the problem of model overfitting due to fewer dataset samples so that we can train a deeper convolutional neural network. Most of the current augmented-based works lack image MOS labels, more likely to pre-training model with distortion type, level, and ranking of image pairs attributes. There is still a distinct difference in the quality score prediction task. We step further on this fact, the idea of weak supervision is introduced, and the pseudo quality score information of the image is generated.

2.1. Data Augmentation

According to [16], a large-scale synthetic distorted dataset was generated. A total of 21,869 high-quality images without distortion from two large datasets Waterloo Exploration Database [15] (4744) and PASCAL VOC2012 [17] (17125), were mixed to serve as the original image. The diversity and richness of its image content far exceed the current image quality assessment dataset with less than 100 original images. Use nine types of synthetic distortion methods: the original four standard synthetic distortion methods in the Waterloo Exploration Database: JPEG compression, JPEG2000 compression, Gaussian blur, and Gaussian white noise. Pink noise, contrast distortion, color dithering, overexposure, and underexposure were added.

Synthetic distortion images contain only one distortion type and level in each augmented image. The distortion in an authentic distortion image is complicated. The simulation and synthesis of authentic distortion images are correspondingly more complicated. Therefore, the images in the authentic distortion datasets CLIVE and KonIQ-10K are directly synthesized and amplified. For images in the original

dataset, apply a blend-type, blend-level distortion.

According to [18], distortions in authentic distortion can be roughly regarded as a mixture of several distortions, such as overexposure, underexposure, blur caused by motion, out of focus, contrast distortion, vignetting, and compression. These distortions are simulated using seven algorithms, specifically, increasing pixel brightness to simulate overexposure distortion, reducing pixel brightness to simulate underexposure distortion, using a motion filter to simulate motion blur distortion, using a Gaussian low-pass filter to simulate image out-of-focus, Image vignetting is simulated by shifting the pixels of each channel of the image RGB, global contrast reduction simulates contrast distortion, and JEPG compression simulates compression distortion. Overexposure and underexposure contain two levels of distortion, and other distortions contain three levels of distortion. Finally, to control the augmented dataset's scale, about 700,000 authentic distorted images of the augmented dataset are selected in equal proportions among the images generated from each original image as the pre-training set.

2.2. Pseudo-Label

For image classification and image detection tasks, label assignments for attributes and locations of content in images are all objective. In contrast, label assignments in IQA are different. The quality score label of distorted images is highly subjective, and the experiment is time-consuming and laborious.

Techniques of weakly supervised learning has been introduced to other domains to deal with the problem of missing labels. Although subjective MOS scores are difficult to obtain, objective FR-IQA scores are easy to calculate. They generally outperform NR-IQA methods. Although the score is not as reliable as the subjective MOS, it has a reference value as a pseudo-label and can be used for pre-training.

Six classic and SOTA FR-IQA methods are used to obtain the PMOS of augmented images, namely SSIM [19], MS-SSIM [20], MDSI [21], VSI [22], FSIM [23], GMSD [24]. The scores of the generated images in the augmented dataset are predicted, and the average score of the six methods is taken as the PMOS label of the generated image.

Fig. 1 presents some distorted image samples with PMOS in our dataset and several distorted images with subjective MOS in TID2013. (a)-(d) are images of different perceptual quality with subjective MOS in TID2013. (e)-(h) are images of different perceptual quality with PMOS in the constructed dataset. It is observed that the distortions in the four images from left to the right are Gaussian blur, contrast distortion, JPEG distortion, and chromatic aberration distortion. When the degree of distortion is similar, the PMOS in our proposed dataset and the subjective MOS in TID2013 are similar. The reliability of our proposed large-scale quality annotation dataset is verified.

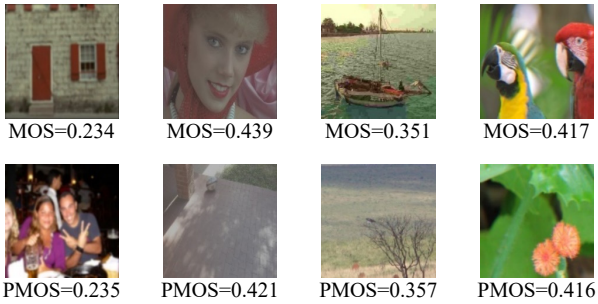


Fig. 1. Quality score of augmented datasets and TID2013 dataset.

III. THE QUALITY PREDICTION FRAMEWORK

The proposed method contains two stages:

1. The pre-training stage. Different from existing methods, which are limited by the lack of image quality score labels and can only do single-task learning. We propose a multi-task learning method that utilizes the quality score information of images in the main task for the score prediction regression task. The auxiliary task uses the image's distortion type and level information to perform the classification task. The combing of image quality-related labels allows our pretrained network to extract quality-aware features better.

2. The fine-tuning stage. We initialize the network with pretrained weights and fine-tune the target dataset. The pretrained weights can extract quality-aware features better than the previous works that use distortion type and level or rank information between image pairs. With better quality-aware features, accuracy score prediction was obtained.

3.1. Multi-Task Learning

The pre-training stage is multi-task learning combining attributes related to image quality. In IQA tasks, compared to using single-task learning to predict image quality scores directly, multi-task learning improves each other's performance by introducing two or more similar tasks into learning and training, correlating the information shared by the tasks, and complementing each other.

ResNet50 is selected as the backbone network. The ResNet series network adopts a residual design. The data output of a specific layer of the first several network layers is skipped multiple layers. It is directly introduced into the input part of the following data layer. This design overcomes the problem of network depth—the problem of low learning efficiency and the inability to improve the accuracy caused by deepening effectively.

In the augmented dataset, the image's distortion type and level information are included automatically when the algorithm applies distortion, and they are closely related to image quality. The image quality assessment model uses the classification information and image quality score. Co-training improves the performance of quality score prediction.

The pre-trained network can extract better quality-aware features from multi-task learning. The network structure of the network is shown in Fig. 2. The ResNet50 backbone network contains four residual blocks, and each residual block contains several residual layers, which finally extract image features and input them into the task-specific head structure. The backbone outputs a 1×2048-dimensionalfeature vector *v*. We send *v* into two fully connected layer branches for multi-task training. Branch 1 is for distortion type and level classification, and branch 2 is for quality score regression.

In the auxiliary task classification branch, the feature *v* from the backbone network is reduced in dimension through the fully connected layers cFC1 and cFC2. A 1×1024-dimensional feature vector *c1* and a 1×*N*-dimensional vector *c2* are outputs, where N denotes the number of classification types. Finally, the classification prediction result is output through the activation function.

For synthetic distortion datasets, according to the number of distortion types and levels of the augmented dataset, *N*=39. Each image in the dataset has only a single type and a single level of distortion. The auxiliary task is a standard classification task, optimized by the SoftMax activation and Cross-Entropy loss functions.

The SoftMax can be formulated as:

$$\hat{\mathrm{C}}_{\mathrm{i}}^{(k)} = \frac{\exp(y_i^{(k)})}{\sum_{j=1}^{39} \exp\left(y_i^{(k)}\right)}, \tag{1}$$

where $\hat{c}^{(k)} = [\hat{c}_1^{(k)}, ..., \hat{c}_{39}^{(k)}]^{\mathrm{T}}$ denotes 39-dimensional classification prediction value of the *k*-th input image, denotes the probability of a specific level of distortion type, denotes the *i*-th activation value of the output of the *k*-th input image in the last fully connected layer cFC2.

The Cross-Entropy can be formulated as:

$$L_{c1} = -\sum_{k=1}^{M}\sum_{i=1}^{39} c_i^{(k)} \log \hat{c}_i^{(k)}. \tag{2}$$

For the authentic distortion dataset, according to the total types of mixed distortion, *N*=26. Each image in the dataset has multiple types of distortion of different levels. The auxiliary task is a multi-label classification task, optimized by the Sigmoid activation function and Binary Cross Entropy loss function. The sigmoid activation function can be formulated as:

$$\sigma(z)=1/(1+e^{(-z)}). \tag{3}$$

The Binary Cross Entropy can be formulated as:

$$Lc_2 = -\frac{1}{M}\sum_{k=1}^{M}\sum_{i=1}^{26}(c_i^{(k)} \log(\hat{c}^{(k)}) + (1 - c_i^{(k)})\log(1 - \log(\hat{c}^{(k)})), \tag{4}$$

where $\hat{c}^{(k)} = [\hat{c}_1^{(k)}, ..., \hat{c}_{26}^{(k)}]^{\mathrm{T}}$ denotes 26-dimensional classification prediction value of the k-th input image, denotes the probability of a specific level of distortion type output by the sigmoid activation function.

In the main task score prediction branch, the feature vector *v* from the backbone network is reduced in dimension through the fully connected layer rFC1, rFC2, andrFC3. A 1×1024-dimensional feature vector *s1* is output from rFC1. We concat *s1* and *c1*(from the classification branch) with ReLU activation function. Finally generates a 1×2048-dimensional mixed feature vector *m*1. The symbol ⊕ denotes the concatenation operation. The mix operation can be formulated as:

$$m1=ReLU(s1 \oplus c1). \tag{5}$$

The mix feature *m1* reduces the dimension and maps through the fully connected layers rFC2 and rFC3, finally

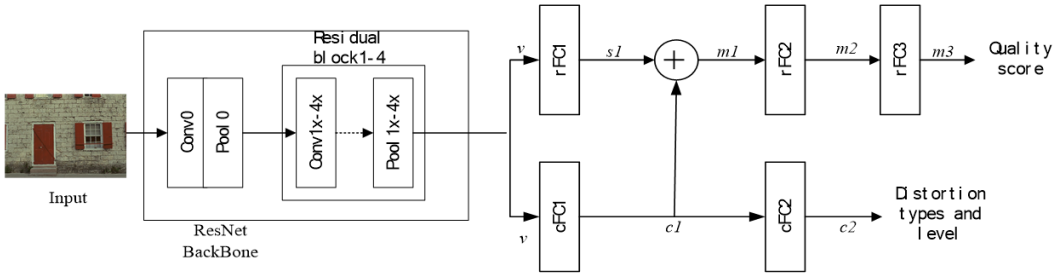


Fig. 2. Multi-task pre-training stage network structure.

outputs the prediction score. The quality score $\hat{s}_i$ is optimized by L_1 loss. The L_1 loss can be formulated as:

$$L_s = \frac{1}{M}\sum_{i=1}^{M}|\hat{s}_i - s_i|, \quad (6)$$

where M denotes the size of the mini batch, $\hat{s}_i$ denotes the quality score predicted by the network, and s_i denotes the pseudo-quality label of the image.

In summary, for the augmented synthetic distortion dataset, the final multi-task training total loss is:

$$L=L_s+L_{c1}. \quad (7)$$

For the augmented authentic distortion dataset, the final multi-task training total loss is:

$$L=L_s+L_{c2}. \quad (8)$$

3.2. Network Fine-Tuning with Quality-Aware Weights

The fine-tuning stage combines the pretrained network weights with better extraction quality perception ability to perform fine-tuning on the target dataset. Early transfer learning used pretrained weights for classification tasks on ImageNet, ignoring the feature differences between classification tasks and IQA tasks. Most of the current work based on augmented datasets uses pre-training for the classification of distortion types and levels or pair-wise rank information. After multi-task training, the pre-trained network can extract features that are more perceptive to extraction quality. Fine-tuning the target dataset can result in more accurate prediction scores.

To alleviate the over-fitting phenomenon, the fully connected layer of the network is modified into two fully connected layers, the output neuron size is 256 and 1 respectively, and finally, the predicted value of the image quality score is output. When initializing the parameters of the ResNet50 network layer of the backbone network, the initialization weights that have undergone multi-task pre-training and are more quality-aware are used. The Fine-tuning of the target dataset is trained with the ground truth scores, and the quality score is optimized by the L_1 loss.

IV. EXPERIMENTS

To verify the effectiveness of the proposed method, we conduct multiple types of experiments. Compared with the current state-of-the-art related mainstream methods on three synthetic distortion datasets, LIVE [25], CSIQ [26], and TID2013 [27], and two authentic distortion datasets, LIVEC [18] and KonIQ-10K [28]; cross-dataset verification is designed to verify the generalization of the method; an ablation experiment is also designed to verify the effectiveness of each module. The Pearson Linear Correlation Coefficient (PLCC) and the performance indicator Spearman's Rank Ordered Correlation Coefficient (SROCC) are used as the evaluation indicators of the method.

4.1. Experimental Setups

Dataset division: In the pre-training stage, for the synthetic distortion data set, the image content does not overlap with the target data set, and the entire augmented dataset is used as the training set. For the authentic distortion dataset, the augmented dataset is divided into a training set (80%) based on the content of the reference images. In the fine-tuning stage, for the target dataset, the dataset is also divided into a training set (80%) and a test set (20%) based on the content of the reference images. Note that for the authentic distortion dataset, both two stage's training sets have the same content, so the image content does not overlap with the target data set.

The data augmentation algorithms were implemented in MATLAB code, and the version of MATLAB is 2018b. Using functions in MATLAB to distorted the image, e.g. "fspecial('gaussian', hsize, hsize/6)" for gaussian blur distortion. All models and loss functions and optimizers in the experiments are implemented in a Linux system with ubuntu18.04. Pytorch is a deep learning package for Python. The version of Python is 3.6.9, and Pytorch is 1.3. Using an NVIDIA RTX 3090 GPU. We use a ResNet-50 pre-trained on ImageNet as the backbone for CNN in the first stage, the FC layer of the head is initialized using the He [29] method and used the ADAM optimizer. The learning rate is set to α, and the optimizer parameters β_1=0.9, β_2 =0.999. In the first stage, the image is scaled to 256×256, and then 224×224 image patches are taken as network input. We set the training iterations to 30, mini-batch=256, and backbone network $\alpha=10^{-4}$. For the synthetic distortion dataset, the fully connected layer cFC$\alpha=10^{-5}$, the fully connected layer sFC$\alpha=10^{-6}$; for the authentic distortion dataset, the fully connected layer cFC$\alpha=10^{-4}$, the fully connected layer sFC$\alpha=10^{-6}$. The combination of learning rates is selected with the best result through experiments. In the second stage, 100 image patches of 224×224 are randomly cropped as input to augment the dataset. We set training iterations to 10, set mini-batch=32, the backbone network, and the two fully connected layers $\alpha=10^{-5}$, where a dropout layer is set before the first fully connected layer, and the dropout rate is 0.5. In the test, 60 224×224 image patches are randomly cropped for each test image.

4.2. Compare with SOTA Methods

For the synthetic distortion datasets, the proposed method is compared with the existing SOTA method on various datasets, and 18 mainstream methods are selected, namely: PSNR, SSIM [19], FSIM [23], BRISQUE [30], CORNIA [31], IL-NIQE [32], CNN [33], HOSA [34], FRIQUEE [35], RANK [14], DMIR-IQA [36], MMMNet [37], AIGQA [38], DB-CNN [15], Deep-FL [39], CaHDC [40] NSSADNN [41] and the Baseline (ResNet50), the performance of ResNet50 on the target dataset is selected as the benchmark for evaluation. The comprehensive performance (i.e., mean value) of SROCC and PLCC on each dataset is in the last column, and the experimental results are shown in Table 1, where the best and second-best performing methods are marked with bold and underlined, respectively.

From the results in Table 1, it can be observed that:

1. The proposed method has the top two performances in almost every dataset, especially on the TID2013 dataset, with more diverse images and distortion types. Meanwhile, it ranked first in the comprehensive performance, proving that the data augmentation applying distortion to many images with different contents is effective. The diversification of image content and image distortion improves the feature extraction ability of the model. Only the second-best results are obtained on the LIVE and CSIQ datasets, but they are not much less than the best results. The reason may be due to the small number of samples in these two datasets and the lack of distortion diversity, which makes it hard to further increase higher indicators.

2. The proposed method outperforms DB-CNN and Deep-FL due to multi-task training using image distortion types, levels, and image quality scores in the pre-training stage. Because the pre-training of DB-CNN only uses the type and level information of image distortion, while Deep-FL only uses the score information of the image. Proves that the multi-task learning taking advantage of more information related to image quality can improve the performance of the model.

Compared with synthetic distorted datasets, the research on authentic distorted images is more challenging. Hence, the existing datasets and related methods are also lacking.

The proposed methods are compared on the authentic distorted image datasets LIVEC and KonIQ-10K. Compared with 9 existing mainstream IQA methods, these 9 methods are: BRISQUE [30], FRIQUEE [35], WaDIQaM-NR [41], MMMNet [37], NSSADNN [42], DB-CNN [15], MetaIQA [43], Deep-FL [39] and the Baseline (ResNet50), the performance of ResNet50 on the target dataset is selected as evaluation. The content of "-" in the table indicates that the corresponding method has no data in the dataset. Bold and underlined are the best and second-best results,

Table 1. Experimental results of synthetic distortion dataset.

Datasets	LIVE [25]		CSIQ [26]		TID2013 [27]		Weight average	
Methods	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
PSNR	0.876	0.872	0.806	0.800	0.636	0.706	0.773	0.793
SSIM [19]	0.913	0.945	0.834	0.861	0.775	0.691	0.841	0.833
FSIMc [23]	0.963	0.960	0.913	0.919	0.802	0.877	0.893	0.919
BRISQUE [30]	0.939	0.942	0.775	0.817	0.572	0.651	0.762	0.803
CORNIA [31]	0.942	0.943	0.714	0.781	0.549	0.613	0.735	0.779
IL-NIQE [32]	0.902	0.908	0.821	0.865	0.521	0.648	0.748	0.807
CNN [33]	0.956	0.954	0.683	0.754	0.558	0.653	0.732	0.787
HOSA [33]	0.948	0.949	0.781	0.841	0.688	0.764	0.806	0.851
FRIQUEE [35]	0.940	0.944	0.835	0.874	0.680	0.753	0.818	0.857
RANK [14]	<u>0.981</u>	<u>0.982</u>	0.861	0.893	0.780	0.793	0.874	0.889
DMIR-IQA [36]	0.967	0.971	0.823	0.881	0.796	0.821	0.862	0.912
MMMNet [37]	0.970	0.970	0.924	0.937	0.832	0.853	0.908	0.920
AIGQA [38]	0.960	0.957	0.927	0.952	<u>0.871</u>	0.893	<u>0.919</u>	0.934
DB-CNN [15]	0.968	0.971	**0.946**	**0.959**	0.816	0.865	0.910	0.931
Deep-FL [39]	0.972	0.978	0.930	0.946	0.858	0.876	0.891	0.907
CaHDC [40]	0.965	0.964	0.903	0.914	0.816	0.865	0.895	0.914
NSSADNN [41]	**0.986**	**0.984**	0.893	0.927	0.844	**0.910**	0.907	<u>0.940</u>
BaseLine	0.950	0.954	0.876	0.905	0.712	0.756	0.846	0.872
Ours	0.976	0.980	<u>0.942</u>	<u>0.954</u>	**0.895**	<u>0.903</u>	**0.940**	**0.945**

respectively.

From the results in Table 2, it can be observed that:

1. The proposed method achieves the top two levels on both authentic distortion datasets. On the LIVEC dataset with only 1162 images, the method performs well, proving that the augmentation method of applying mixed distortion on the images of the LIVEC dataset is effective. Compared with the baseline using ImageNet's pre-trained weights, the augmented data. The ability of the model for feature extraction of authentic distorted images is facilitated. The SROCC metric on LIVEC is second-best, probably because, in DBCNN, the distortion of synthetically distorted images is incorporated, which is missing in our method.

2. On the KonIQ-10K dataset, a large-scale authentic distortion dataset containing 10073 images, it is noted that only the baseline of ImageNet's pre-training weights is used, and its effect on the KonIQ-10K data is already very good, exceeding the current IQA methods. The improvement of the proposed method on the baseline is relatively small. There are two main reasons for the excellent baseline effect: First, the authentic distorted images are more similar in content to the images in the ImageNet dataset, and the ImageNet pre-training weights are aware of the authentic distorted image features to a certain extent. Second, compared with LIVEC, the number of images in the KonIQ-10K dataset is nearly ten times that of LIVEC, and the model overfitting phenomenon is weakened, which leads to achieving good performance. Meanwhile, due to the large number of images of KonIQ-10K, when it is augmented, the distortion types and distortion levels of mixed diversity are less, resulting in the auxiliary tasks of multi-task training cannot well promote the network to extract image quality-related information feature.

In summary, the proposed method achieves SOTA performance on both synthetic and authentic databases.

Table 2. Experimental results of authentic distortion dataset.

Datasets	LIVEC [18]		KonIQ-10K [28]	
Methods	SROCC	PLCC	SROCC	PLCC
BRISQUE [30]	0.608	0.629	-	-
FRIQUEE [35]	0.682	0.705	-	-
WaDIQaM [42]	0.671	0.680	-	-
MMMNet [37]	**0.852**	0.876	-	-
NSSADNN [41]	0.745	0.813	-	-
DB-CNN [15]	0.851	0.869	0.875	0.884
MetaIQA [43]	0.802	0.835	0.877	0.850
Deep-FL [39]	0.734	0.769	0.887	0.877
BaseLine	0.819	0.849	0.904	**0.912**
Ours	0.847	**0.881**	**0.912**	**0.912**

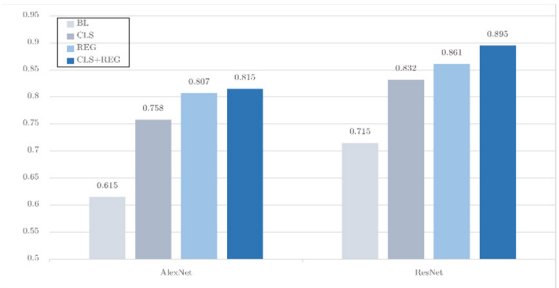
4.3. Ablation Experiment

Several ablation experiments are designed to demonstrate the effectiveness of each module of the proposed method, which is performed on synthetic distortion and authentic distortion datasets, respectively. In Fig. 3, AlexNet and ResNet correspond to different backbone networks, respectively. FT means that in the second stage, the ImageNet pre-training weights are directly tuned, that is, the baseline in the beforementioned; CLS means that only the image distortion type and level label are used in the first stage conducting classification single-task training; REG means that only the pseudo-quality score of the image is used for regression single-task training in the first stage; CLS+REG is the proposed multi-task training strategy.

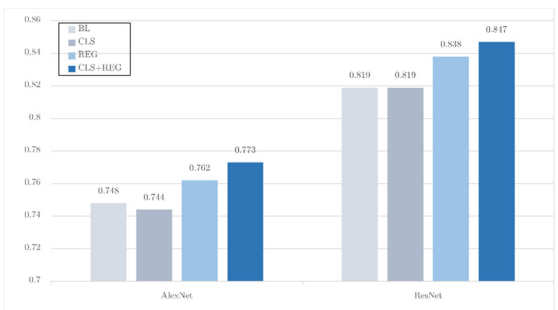
From Fig. 3, it can be observed that:

1. The proposed method is portable in different networks. The performance is improved compared to the baseline on different backbone networks AlexNet and Resnet.

2. Multi-task training combines the advantages of the quality-related labels, and the effect is significantly improved. Different training strategies in the pre-training stage have improved results relative to the baseline, indicating that the network weights obtained in the first stage are more quality-aware than those of ImageNet. Meanwhile, the results of different pretrain strategies for the subsequent tuning stage are CLS, REG, and Multi, respectively, from low to high. Using the image distortion type and level to pretrain the model, performs worse than using the image quality score. Multi-task training performs best.



(a) SROCC of ablation experiments on TID2013



(b) SROCC of ablation experiments on LIVEC

Fig. 3. Experimental results of ablation on synthetic distortion and authentic distortion datasets.

4.4. Cross Database Evaluations

To verify the robustness of the proposed method, cross-dataset experiments are conducted, and the experimental results are compared with several current competitive NR-IQA methods. Cross-dataset experiments refer to training the model on one complete dataset and testing on another complete dataset. Table 3 shows the results of cross-dataset experiments between LIVE, CSIQ, TID2013, and LIVEC datasets. The content of "-" in the table indicates that the corresponding method has no data in the dataset. Bold and underlined are the best and second-best results, respectively.

As can be seen from Table 3, the proposed method shows good generalizability. The method has the top two performances on most datasets. Even when trained on small datasets LIVE and CSIQ, which contain limited distortion types, it can achieve good performance on other datasets during the test. Meanwhile, for training on synthetic datasets and testing on authentic distortion (or vice versa), the results between synthetic distortion and authentic distortion datasets are relatively lower. This is mainly due to the large difference in features between synthetic and authentic distorted images, making such experiments challenging.

V. CONCLUSIONS

We proposed a multi-task learning IQA method in this paper. The method utilizes image distortion type, level, and quality score comprehensively. Various attributes related to image quality make the model can better extract image quality-aware features. It demonstrates state-of-the-art performance on both synthetic distortion datasets and authentic distortion datasets. We believe it is arises from augmenting the datasets with various distortions and levels to reduce the phenomenon of network overfitting and training. In addition, the results of cross-dataset experiments and various ablation experiments also show the reliability of augmented datasets and PMOS, proposed model has good generalization, robustness, and portability.

Meanwhile, our method has many extensibilities and improvements. For the data augmentation, more diverse and refined distortion types and levels would increase the quality of the datasets. In the pretrain stage, we handle the synthetic and authentic datasets separately, deal with the datasets more unified will improve the generalization of our model. In the fine-tuning stage, only the features from the last layer are used for score prediction. Considering the connection between the human visual system and CNN, fusing multi-level features can further improve the model's performance.

ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China (Grant No.62162045,61866028), the Technology Innovation Guidance Project of Jiangxi Province, China (Grant No. 20212BDH81003), and the Graduate Innovation Special Fund Project of Nanchang Hangkong University (Grant No. YC2020127).

Table 3. Cross-dataset validation.

Training datasets	LIVE [25]			CSIQ [26]		
Testing datasets	CSIQ	TID2013	LIVEC	LIVE	TID2013	LIVEC
BRISQUE [30]	0.562	0.358	0.337	0.847	0.454	0.131
FRIQUEE [35]	<u>0.722</u>	0.461	0.411	<u>0.879</u>	0.463	0.264
CORNIA [31]	0.649	0.360	0.433	0.853	0.312	<u>0.393</u>
HOSA [33]	0.594	0.361	<u>0.463</u>	0.773	0.329	0.291
WaDIQaM [42]	0.704	0.462	-	-	-	-
SGDNet [44]	0.719	**0.532**	0.455	0.832	<u>0.521</u>	0.311
Ours	**0.763**	<u>0.517</u>	**0.506**	**0.886**	**0.542**	**0.396**
Training datasets	TID2013 [27]			LIVEC [18]		
Testing datasets	LIVE	CSIQ	LIVEC	LIVE	CSIQ	TID2013
BRISQUE [30]	0.790	0.590	0.254	0.238	0.241	0.280
FRIQUEE [35]	0.755	0.636	0.181	**0.644**	<u>0.592</u>	**0.424**
CORNIA [31]	<u>0.846</u>	0.672	0.293	0.588	0.446	<u>0.403</u>
HOSA [33]	0.846	0.612	**0.319**	0.537	0.336	0.399
WaDIQaM [42]	-	<u>0.733</u>	-	-	-	-
SGDNet [44]	0.759	0.571	<u>0.309</u>	0.491	0.559	0.229
Ours	**0.868**	**0.742**	**0.319**	0.605	**0.596**	0.382

REFERENCES

[1] J. H. Lee, Y. W. Lee, D. Jun, and B. G. Kim, "Efficient color artifact removal algorithm based on high-efficiency video coding (HEVC) for high-dynamic range video sequences," *IEEE Access (IEEE),* vol. 8, pp. 64099-64111, 2020.

[2] H. Chen, X. He, L. Qing, Y. Wu, C. Ren, and R. E. Sheriff, et al.," Real-world single image super-resolution: A brief review*," Information Fusion*, vol. 79, pp. 124-145, 2022.

[3] S. Li, W. Ren, F. Wang, I. B. Araujo, E. K. Tokuda, and R. H. Junior, et al., "A comprehensive benchmark analysis of single image deraining: Current challenges and future perspectives," *International Journal of Computer Vision*, vol. 129, no. 4, pp. 1301-1322, 2021.

[4] D. Singh and V. Kumar, "A comprehensive review of computational dehazing techniques," *Archives of Computational Methods in Engineering*, vol. 26, no. 5, pp. 1395-1413, 2019.

[5] W. Wang, Y. Yang, X. Wang, W, Wang, and J, Li, "Development of convolutional neural network and its application in image classification: A survey, "*Optical Engineering,* vol. 58, no. 4, p. 040901, 2019.

[6] Y. Liu, P. Sun, N. Wergeles, and Y. Shang, "A survey and performance evaluation of deep learning methods for small object detection," *Expert Systems with Applications,* vol. 172, no. 4, p. 114602, 2021.

[7] G. Ciaparrone, Fl. Sánchez, S. Tabik, L. Troianoc, R. Tagliaferria, and F. Herrerab, "Deep learning in video multi-object tracking: A survey," *Neurocomputing*, vol. 381, pp. 61-88, 2019.

[8] G. Zhai and X. Min. "perceptual image quality assessment: A survey," *Science China. Information Sciences*, vol. 63, no. 11, pp.1-52, 2020.

[9] X. Xie, Y. Zhang, Wu J, G. Shi, and W. Dong, "Bag-of-words feature representation for blind image quality assessment with local quantized pattern," *Neurocomputing*, vol. 266, pp. 176-187, 2017.

[10] H. O. Shahreza, A. Amini, and H. Behroozi, "No-reference image quality assessment using transfer learning," in*2018 9th International Symposium on Telecommunications (IST). IEEE*, 2018, pp. 637-640.

[11] M. Cheon, S. J. Yoon, B. Kang, and J. Lee, "Perceptual image quality assessment with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 433-442.

[12] J. Gu, H. Cai, C. Dong, J. S. Ren, R. Timofte, and Y. Gong, et al., "NTIRE 2021 challenge on perceptual image quality assessment," in *Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition,* 2021, pp. 677-690.

[13] J. Gu, H. Cai, H. Chen, X. Ye, J. S. Ren, and C. Dong, "PIPAL: A large-scale image quality assessment dataset for perceptual image restoration," in *European Conference on Computer Vision,* Springer, Cham, 2020, pp. 633-651.

[14] X. Liu, J. an De Weijer, and A. D. Bagdanov, "RankIQA: Learning from rankings for no-reference image quality assessment," *IEEE International Conference on Computer Vision (ICCV),* Venice, Italy, 2017, pp. 1040-1049.

[15] K. Ma, Z. Duanmu, Q. Wu, Z. Wang, H. Yong, and H. Li, et al., "Waterloo exploration database: New challenges for image quality assessment models," *IEEE Trans. Image Process*, vol. 26, no. 2, pp. 1004-1016, 2017.

[16] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 30, no. 1, pp. 36-47, 2020.

[17] M. Everingham, L. Vangool, C. K. I. Williams, J. Winn, and A. Zisserman," The PASCAL visual object classes (VOC) challenge," *International Journal of Computer Vision,* vol. 88, no. 2, pp. 303-338, 2010.

[18] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372-387, 2016.

[19] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600-612, 2004.

[20] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers,* CA, USA*, 2003*, vol. 2, pp. 1398-1402. 2003.

[21] H. Z. Nafchi, A. Shahkolaei, R. Hedjam, and M. Cheriet, "Mean deviation similarity index: Efficient and reliable full-reference image quality evaluator," *IEEE Access,* vol. 4, pp. 5579-5590, 2016.

[22] L. Zhang, Y. Shen, and H. Li, "VSI: A visual saliency-induced index for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4270-4281, 2014.

[23] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378-2386, 2011.

[24] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index*," IEEE Transactions on*

Image Processing, vol. 23, no. 2, pp. 684-695, 2014.

[25] H. R. Sheikh, M. F. Sabir, and A.C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440-3451, 2006.

[26] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, pp. 011006:1-011006:21, 2010.

[27] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarianet, and J. Astola, et al., "Image Database Tid2013: Peculiarities, results and perspectives," *Signal Processing: Image Communication*, vol. 30, pp. 57-77, 2015.

[28] V. Hosu, H. H. Lin, T. Sziranyi, and D. Saupe, "Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 4041-4056, 2020.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on Imagenet Classification," *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1026-1034.

[30] A. Mittal, G. S. Muralidhar, J. Ghosh, and A. C. Bovik, "Blind image quality assessment without human training using latent quality factors," in *IEEE Signal Process Letters,* vol. 19, no. 2, pp. 75-78, 2012.

[31] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Providence, Rhode Island,* pp. 1098-1105, 2012.

[32] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," in *IEEE Transactions on Image Processing,* vol. 24, no. 8, pp. 2579-2591, 2015.

[33] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition,* Columbus, USA, 2014, pp. 1733-1740.

[34] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, "blind image quality assessment based on high order statistics aggregation," *IEEE Trans Image Process,* vol. 25, no. 9, pp. 4444-4457, 2016.

[35] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using A bag of features approach," *Journal of Vision*, vol. 17, no. 1, pp. 32-32, 2017.

[36] D. Liang, X. Gao, W. Lu, and J. Li, "Deep blind image quality assessment based on multiple instance regression," *Neurocomputing*, vol. 431, pp. 78-89, 2021.

[37] F. Li, Y. Zhang, and P. C. Cosman, "MMMNet: An end-to-end multi-task deep convolution neural network with multi-scale and multi-hierarchy fusion for blind image quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 12, pp. 4798-4811, 2021.

[38] J. Ma, J. Wu, L. Li, W. Dong, X. Xie, and G. Shi, et al.,"Blind image quality assessment with active inference," *IEEE Transactions on Image Processing,* vol. 30, pp. 3650-3663, 2021.

[39] H. Lin, V. Hosu, and D. Saupe, DeepFL-IQA: Weak Supervision for Deep IQA Feature Learning, http://arxiv.org/abs/2001.08113, 2020.

[40] J. Wu, J. Ma, F. Liang, W. Dong, G. Shi, and W. Lin, "End-to-end blind image quality prediction with cascaded deep neural network," *IEEE Transaction on Image Process*, vol. 29, pp. 7414-7426, 2020.

[41] B. Yan, B. Bare, and W. Tan, "Naturalness-aware deep no-reference image quality assessment," *IEEE Transactions on Multimedia,* vol. 21, no. 10, pp. 2603-2615, Oct. 2019.

[42] S. Bosse, D. Maniry, K. R. Mller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 206-219, Jan. 2018.

[43] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, "MetaIQA: Deep meta-learning for no-reference image quality assessment," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* Seattle, USA, 2020, pp. 14131-14140.

[44] S. Yang, Q. Jiang, W. Lin, and Y. Wang, "SGDNet: An end-to-end saliency-guided deep neural network for no-reference image quality assessment," *ACM International Conference on Multimedia Association for Computing Machinery*, Nice, France, 2019, pp.1383-1391.

Journal of Multimedia Information System VOL. 9, NO. 2, June 2022 (pp. 75-86): ISSN 2383-7632 (Online)
https://doi.org/10.33851/JMIS.2022.9.2.75

AUTHORS

Lijing Lai Master student at the School of Software, Nanchang Hangkong University. His research interests include image processing and computer vision.

Jun Chu received the Ph.D. degree from Northwestern Polytechnic University, Xi'an, China, in 2005. She was a Postdoctoral Researcher with the Exploration Center of Lunar and Deep Space, National Astronomical Observatory of Chinese Academy of Sciences, from 2005 to 2008. She was a Visiting Scholar with the University of California at Merced, USA. She is currently the Director of the Key Laboratory of Jiangxi Province for Image Processing and Pattern Recognition, and a Full Professor with the School of Software, Nanchang Hangkong University. Her research interests include computer vision and pattern recognition. She was also a member of the Computer Vision Special Committee and China Computer Federation.

Lu Leng received the Ph.D. degree from Southwest Jiao tong University, Chengdu, China, in 2012.He performed his Postdoctoral Research with Yonsei University, Seoul, South Korea, and the Nanjing University of Aeronautics and Astronautics, Nanjing, China. He was a Visiting Scholar with West Virginia University, USA. He is currently an Associate Professor with NanchangHangkong University and a Visiting Scholar with Yonsei University. He has published more than 70 international journal and conference papers. He has been granted several scholarships and funding projects for his academic research. He is a Reviewer of several international journals and conferences. His research interests include image processing, biometric template protection, and biometric recognition. Dr. Leng is a member of the Association for Computing Machinery (ACM), the China Society of Image and Graphics (CSIG), and the China Computer Federation (CCF).