Quality Assessment on Authentically Distorted Images by Expanding Proxy Labels

Xiaodi Guan, Fan Li * and Lijun He

School of Information and Communications Engineering, Xi’an Jiaotong University, Xi’an 710049, China; gxd1997@stu.xjtu.edu.cn (X.G.); lijunhe@mail.xjtu.edu.cn (L.H.)
* Correspondence: lifan@mail.xjtu.edu.cn

Received: 7 January 2020; Accepted: 29 January 2020; Published: 3 February 2020

Abstract: In this paper, we propose a no-reference image quality assessment (NR-IQA) approach towards authentically distorted images, based on expanding proxy labels. In order to distinguish from the human labels, we define the quality score, which is generated by using a traditional NR-IQA algorithm, as “proxy labels”. “Proxy” means that the objective results are obtained by computer after the extraction and assessment of the image features, instead of human judging. To solve the problem of limited image quality assessment (IQA) dataset size, we adopt a cascading transfer-learning method. First, we obtain large numbers of proxy labels which denote the quality score of authentically distorted images by using a traditional no-reference IQA method. Then the deep network is trained by the proxy labels, in order to learn IQA-related knowledge from the amounts of images with their scores. Ultimately, we use fine-tuning to inherit knowledge represented in the trained network. During the procedure, the mapping relationship fits in with human visual perception closer. The experimental results demonstrate that the proposed algorithm shows an outstanding performance as compared with the existing algorithms. On the LIVE In the Wild Image Quality Challenge database and KonIQ-10k database (two standard databases for authentically distorted image quality assessment), the algorithm realized good consistency between human visual perception and the predicted quality score of authentically distorted images.

Keywords: no-reference image quality assessment; authentic distortion; proxy label

1. Introduction

Images have become a main carrier of information dissemination as multimedia technology develops rapidly. Nevertheless, the information usually suffers from various degrees of distortion during acquisition, transmission, storage, and other conditions such as camera motion, which makes it extremely difficult to perform subsequent processing. For the purpose of maintaining, controlling, and improving the image quality, it is important to evaluate and quantify it. Since we are in the big data era, the quality of images has a lot of impact in daily life, as the high-definition camera on mobile phones, aerial equipment on drones, and monitoring equipment for public transportation. If the images obtained are not clear enough, or there are noises, it is bound to affect user experience. In the relevant scientific research field, image quality is also closely associated with various scientific works [1,2]. For example, astronomical observations require high quality images, and for medical imaging, the quality even determines final diagnosis. Therefore, the need to develop an efficient and reliable image quality assessment method is indeed increasing.

Image quality assessment is the process of analyzing the features of an image and then, evaluating the degree of its distortion. There are two categories of image quality assessment (IQA) methods, subjective methods and objective methods. Subjective IQA methods rely on the opinions of a great many viewers, making it costly, time consuming, and inefficient in practical applications. The objective
methods overcome these deficiencies by using a mathematical model to create a mapping between image features and quality scores that strives to truly reflect human visual perception. Computers are used to build models that imitate the human visual system. When the powerful computing ability is utilized, the evaluation efficiency is greatly improved.

The objective IQA methods are usually divided into the following three categories, based on the available information of the given reference image (undistorted image): Full-reference IQA (FR-IQA), reduced-reference IQA (RR-IQA) and no-reference IQA (NR-IQA).

Among these three types, NR-IQA has the widest range of application and the most practical value, and therefore has received more attention and has become a hotspot in recent years. In early NR-IQA studies, researchers were committed to design some hand-crafted features that could discriminate distorted images and, then, train a regression model to predict image quality. Some NR-IQA methods have been based on a specific distortion [3,4], which commonly used the prior knowledge of the distortion type. To assess the image quality with no information available, natural scene statistics (NSS)-based methods have been widely used to extract reliable features, which assume the natural images share certain statistics and the occurrence of distortions can change these statistics [5–9]. Nevertheless, the hand-crafted features have always been designed for a specific type of distortion which lies in the low-level feature and leads to insufficient feature extraction and analysis.

The development of convolutional neural networks (CNN) has greatly accelerated computer vision research and application. In 2012, Krizhevsky et al. used CNN to make significant progress in image recognition [10], which enabled researchers to have a greater prospect of computer vision. With the popularity of CNN, the structure of neural networks is getting deeper and deeper. A deeper structure means the network has a stronger ability to extract features with higher accuracy of results, however, however, at the cost of a larger amount of training data.

The success of CNN in image recognition has attracted researchers to look for the application in IQA. When compared with the traditional methods, the application of neural networks has brought great progress to NR-IQA. Researchers have completed the NR-IQA task by training deep neural networks (DNN) [11–13]. Nevertheless, the problem to be solved in the above research is that we lack a great amount of sample data for the task. Especially when the network structure is deep and wide, the number of parameters increases rapidly, in the meantime, the number of training data should be increased accordingly. The labeling process in an IQA image dataset requires that many human reviewers assess each image which is extremely labor intensive and costly. Therefore, most of the IQA datasets are too limited for effectively training neural networks, as shown in Table 1. Therefore, researchers focus on how to increase the amount of data.

### Table 1. Number of images in different database for image classification and image quality assessment (IQA).

| Database                                           | Number of Images   |
|----------------------------------------------------|--------------------|
| ImageNet [14] (for image classification)            | More than 14 million |
| LIVE In the Wild Image Quality Challenge database [15] (The IQA database for authentic distortion) | 1162               |
| TID2013 [16] (the biggest database for IQA)         | 3000               |

The methods for expanding the amount of data can be roughly divided into two types, image segmentation inside the database and extension out of the dataset. The method of image segmentation in the database is to expand the data volume by dividing the entire image into tiles [12,17–22]. However, it has shortcomings in that the assessment of small image patches can ignore the integrity of a whole image and it is not efficient. With regards to methods of augmenting outside the database, researchers always process the available images, such as reversing, contrast changing, and applying varying degrees of Gaussian blur, to ensure sufficient samples [23–25].
The quality of the image is mainly affected by distortion. Distortion is divided into two types according to the way it is generated, synthetic distortion, and authentic distortion. Synthetic distortion refers to distortion generated during processing and transmission, such as distortion caused by image compression. Authentic distortion refers to distortion occurring in daily, for example, camera jitter, overexposure, and underexposure. In the current research work on NR-IQA, the research objective is mainly focused on synthetic distortion. When algorithms for synthetic distortion are used for authentic distortion, they performed poorly. A few algorithms for authentic distortion have not performed excellently because of the limited amount of data, and the fact that authentic distortion images cannot be augmented with image processing methods such as Gaussian blur.

Conversely, in real life, authentic distortion is the main factor affecting image quality. Meanwhile, with the rapid development of mobile devices and camera equipment, as well as the eagerness of people to know whether their devices work well, the requirement to evaluate the quality of authentic images has been rapidly improving. Unfortunately, the existing IQA algorithms do not perform well in the real life daily scene. The difficulty is that authentic distortion types are much more varied and the degree of distortion is much more complex, and therefore when applied to authentic application scenes, the methods for synthesizing distortion underperform in experiment. Therefore, the development of no-reference image quality assessment methods for authentic distortion is of great significance in practical applications. However, methods for authentic distortion face the difficulty that researchers cannot expand data such as by processing the available images using Gaussian blur. Authentic distortions are unlike synthetic distortions, since authentic distortions are generated in human’s daily life with strong authenticity and complexity. Therefore, the available images are too few. Therefore, DNN models are always trained with an IQA database for direct authentic distortion. In Reference [24], researchers designed an architecture called S-CNN for synthetically distorted images, however, considering that the DNN model was not beneficial for authentic IQA databases, they selected the pretrained VGG-16 network for the classification task on ImageNet as another branch to extract relevant features for authentically distorted images. Finally, in the fine-tuning step, they tailored the pretrained S-CNN and VGG-16 [26] and introduced bilinear pooling module for optimization of the entire model. Unfortunately, the limited size of database restricted the performance on the authentic distortion.

In this paper, an NR-IQA algorithm, based on image proxy labels, is proposed, mainly focusing on the assessment of authentically distorted images. In order to distinguish from the human labels, we define the quality score, which is generated by using a traditional NR-IQA algorithm, as “proxy labels”. “Proxy” means that the objective results are obtained by computer after the extraction and assessment of the image features, instead of human judgement. Above all, a traditional IQA algorithm for authentic distortion is applied to obtain amounts of image proxy labels. Generating the proxy labels expands the data size without processing the natural images, to ensure the authenticity of natural image distortion. Then, we use the proxy labels to train the model followed by a fine-tuning.

The innovation lies in applying cascading transfer learning to IQA. Since VGG16 [26] is a classic model for image classification tasks, the feature it can extract is related to classification of the images in natural scenes. In our work, the classification task can be regarded as the source task and the database for training the original VGG16, i.e., the ImageNet database, can be regarded as the source domain. The target task is the IQA, and the target domain is the image in the IQA database for authentic distortion. The source domain and the target domain are similar to some extent, that is, ImageNet contains the images whose scenes are similar with authentic distortion images. Therefore, we use the images in the source domain to generate proxy labels, then, train the VGG16 model and transfer it to a preliminary IQA model. After the first transfer learning, the model has the ability to assess the image quality. For the sake of adequately meeting the IQA task, the twice transfer learning is necessary. In order to be consistent with image type in the first transfer learning, we use authentic distortion images for training in the twice training. After the twice training, the model is completely transferred to the target task.
We apply cascading transfer learning in our algorithm. The first expands the authentic images data size and learning preliminarily, then, the second inherits the foundation from the previous learning and fine-tunes the model. Therefore, the model is transferred to the target task completely, and can more accurately predict the quality of authentically distorted images. When compared with the algorithms which only fine-tune, our algorithm performs better in terms of expanding data size and adequate training. The experimental results strongly demonstrate that the predictions generated by our algorithm are highly correlated with human perception and achieves state-of-art performance in a LIVE In the Wild Image Quality Challenge database [15].

The remainder of this paper is organized as follows: Related work is introduced in Section 2, details of the methodology we proposed are given in Section 3, experimental results and the analysis are presented in Section 4, and conclusions are given in Section 5.

2. Related Works

2.1. Traditional NR-IQA Methods

Regarding FR-IQA, some methods are commonly used, such as Peak Signal to Noise Ratio (PSNR) and structural similarity index (SSIM) [27]. PSNR is the most widely used method in the field of image and video processing. It has low computational complexity and fast implementation speed. It has been applied in video coding standards H.264 and H.265. Although PSNR has the above characteristics, it has obvious limitations. It is greatly affected by pixels and has low consistency with subjective evaluation. It does not take into account some important physiological, psychological, and physical aspects of the human visual system (HVS) feature. According to the HVS, an evaluation method of error sensitivity analysis and structural similarity analysis (structural similarity index, SSIM) is proposed. Structural similarity assumes that the HVS is highly adapted to extracting structural information from the scene in an attempt to simulate the structural information of the image. However, due to the necessity of pristine images, the commonly used PSNR and SSIM cannot be used for NR-IQA.

Most of the traditional NR-IQA methods can be categorized into natural scene statistics (NSS) methods and learning-based methods. NSS methods assume that images of different qualities have different statistical properties for the response of a particular filter. Discrete cosine transform (DCT) and other domains are typically used to extract features. The purpose of these methods is to assess image quality by estimating the distribution parameters. There are also researchers who extract features directly from the airspace, which greatly improves performance [28]. The DIIVINE [29] deploys summary statistics under an NSS wavelet coefficient model. Another model, BLIINDSS [30] extracts a small number of NSS features in the DCT domain. BRISQUE [28] trains an SVR on a small set of spatial NSS features. CORNIA [31], which is not an NSS-based model, builds distortion-specific code words to compute image quality. NIQE [32] is an unsupervised NR-IQA technique driven by spatial NSS-based features that requires no exposure to distorted images at all. C-DIIVINE [33] is a complex extension of the NSS-based DIIVINE IQA model which uses a complex steerable pyramid. Features computed from it enable changes in local magnitude and phase statistics induced by distortions to be effectively captured.

To tackle the difficult problem of quality assessment of images in the wild, FRIQUEE [34] produces a large and comprehensive collection of “quality-sensitive” statistical image features drawn from among the most successful traditional NR-IQA models.

Furthermore, FRIQUEE deployed these models in a variety of color spaces representative of both chromatic image sensing, bandwidth-efficient, and perceptually motivated color processing. This large collection of features defined in various complementary perceptually relevant color and transform-domain spaces drives the “feature maps” based approach. Therefore, FRIQUEE conducted a discriminant analysis of an initial set of 564 features designed in different color spaces, which, when used to train a regressor, produced an NR-IQA model delivering a high level of quality prediction power.
In the learning-based approach, researchers use support vector machines or neural networks to extract local features of an image and map it to mean opinion score (MOS), or differential mean opinion score (DMOS), to build a predictive model. The codebook method [35] combines different features rather than directly using local features. Since the amount of data in the existing dataset is too small, it is of significance to construct the codebook through unsupervised learning using a dataset without MOS. Saliency maps can also be used to simulate human visual systems and improve the accuracy of these methods [36].

2.2. DNN Methods for NR-IQA

The hand-crafted features could be not sufficient to fully characterize the complicated image structures and distortions. Since deep neural network (DNN) can automatically capture more deep features, a series of DNN methods are proposed to deal with the NR-IQA task [37–40] and provides a very promising performance.

In Reference [41], according to the different role of DNN, authors, first, divided the DNN methods into two categories. One category is the support vector regression (SVR)-based NR-IQA methods, which use DNN to extract deep features and SVR methods to predict image quality. The other category is the DNN-based NR-IQA methods, which take full advantage of the back-propagated capability of DNN to optimize prediction accuracy.

The SVR-based methods are classified into two major schemes as follows: (1) Extracting from low-level features of image [42–44] and (2) extracting from data of image/image patch [45–47]. Instead of using DNN models to extract deep features, DNN-based methods directly use the DNN model to predict image quality. According to different input in the DNN, the patch-input methods and the image-input methods can be summarized. Since the performance of the DNN heavily depends on the number of training data, the patch-input method aims to divide images into multiple patches as DNN input to increase training samples. According to the different labels of training patches, two ways can be adopted. One is to use the image subjective score (SS) as the image patch label [12,17–19]. The other is to use the score which is generated by the FR-IQA method as the image patch label [20–22]. Rather than using image patches as the input, the image-input methods aim to train a prediction model by using the whole image and its associated ground truth, which can effectively estimate the quality of a whole image. However, there has been limited effort towards end-to-end optimized NR-IQA using DNN, primarily due to the lack of sufficient ground truth labels of images. For expanding distorted images, there are generally two expanded, taking the advantage of large databases, such as the ImageNet [14], and the artificial generation images [23–25]. The DNN, then, is trained by the transfer-learning method. This is a common way to overcome the small database task. In Reference [25], Li et al. utilized network in network (NIN), which is pretrained for the classification task on the large-scale ImageNet database, followed by transfer learning to deal with the NR-IQA problem. RankIQA [23] designed a new strategy to generate the large-scale distortion images without laborious human labeling. According to the rule that the image quality decreases with the increase of the distortion levels, they synthetically generated the ranked image pairs with five different distortion levels from a big database. Then, using the pairs of the ranked images, they pretrained a Siamese network to learn image distortion levels. Ultimately, they fine-tuned a branch of Siamese network to predict image score, which aim to transfer image distortion levels to quality scores. To improve performance of different IQA databases, Zhang et al. designed an end-to-end DB-CNN solution for NR-IQA that worked for both synthetically and authentically distorted images [24]. They designed the architecture of the S-CNN for synthetically distorted images, which aimed to classify the probability of each distortion type at the specific degradation level. Considering that this DNN model is not beneficial for authentic IQA databases, they selected the pretrained VGG-16 network for the classification task on ImageNet as another branch to extract relevant features for authentically distorted images. Finally, in the fine-tuning step, they tailored the pretrained S-CNN and VGG-16 and introduced bilinear pooling module for optimization of the entire model.
Nevertheless, all the methods underperform on the authentic distortion database. To address the mentioned problems, we combine the traditional NR-IQA method to generate somewhat dependable labels without human labeling and apply the VGG-16, which performs well for NR-IQA, to acquire more credible results for authentic NR-IQA problem.

3. Proxy Labels Based No-Reference Image Quality Assessment

In this section, a NR-IQA algorithm based on obtaining proxy labels for authentically distorted images is detailed. First, the general framework and process of the algorithm are described and, then, the key points and innovations of the algorithm proposed.

3.1. Overview of Our Approach

Most of the IQA research is focused on synthetic distortion. In real-life scenarios, due to the complexity and variety of real scenes, the complexity of image distortion is not a simple analogy of synthetic distortion. Therefore, the method for synthesizing distortion cannot meet the actual needs when applied to real-life scenes. At the same time, we noticed that the application of deep learning to blind IQA has been the focus of many researches. The IQA method based on deep learning establishes the mapping from the underlying image to the high-level perception through the deep neural network, so that the image quality is more closely related to human perception. However, deep networks have obvious shortcomings that cannot be ignored such as a large number of samples are needed for training. However, manual labeling of image quality is time-consuming and labor intensive, and it is extremely unrealistic to create a dataset with a variety of distortions, such as ImageNet, and therefore researchers are also looking to expand the number of training available.

Notably, there are a lot of authentically distorted images in life, but these images are not manually labeled. If we could effectively use this large number of unlabeled distortion images, we could significantly expand the amount of data, thus, making network training more accurate. However, manual labeling of these images is time-consuming and labor intensive. If the image is labeled using a traditional IQA algorithm, the computing power of the computer is fully utilized, and an objective quality evaluation of a large number of images is performed in a short time. In this way, the efficiency is greatly improved as compared with manual labeling. Although the objective assessment based on the machine is not close to the human visual perception, it can predict by extracting and analyzing the features of a large number of images, and the network can learn the IQA-related features during the training, and therefore the network has the function of the assessment of image quality.

The ultimate goal of the IQA is to simulate the perception of the person to evaluate the image. If the model is trained using only a large number of machine assessment labels, although the problem of insufficient data volume is solved, the model has the function of evaluating image quality. However, the mapping established by the model stops at objectively assessing the quality, and it is necessary to continue to adjust in order to imitate the perception of the person. In order to avoid this problem, we apply the manually labeled images to continue to learn. By fine-tuning the model through the data in the existing IQA dataset, the network can further imitate human perception based on the function of assessing image quality, adjust the model mapping with human perception, and complete the IQA task more perfectly.

Our approach is based on the observation that a huge number of authentically distorted images is available. The first step in this work is to acquire a lot of real distortion images. We select ImageNet [14] for augmenting data size. This set contains a large number of real scene images, which can meet both the authentic distortion and the large amount of data required in this work. Moreover, the images are mostly taken from reality, and therefore model trained by these images can play a more robust role when applied to life practices. After obtaining a large number of authentic distortion images, the low-level features of the images are extracted by FRIQUEE [34], a traditional NR-IQA method, to perform objective quality scores. The scores obtained are saved as proxy labels of the images.
This step aims at labeling without humans, which greatly expands the amount of training data as compared with other methods.

In the second step, the authentic distortion images from ImageNet and their corresponding proxy labels are fed into the VGG16 network for training, that is, the first transfer learning. Through the training by a large number of data, the task of VGG16 is gradually transformed into IQA from classification, meanwhile the deep-level features related to image quality are learned and other irrelevant features are suppressed. In a word, the model has the function of image quality assessment.

Although the network has been preliminary competent for the IQA task after an adequate training, the ability of predicting remains in the objective evaluation stage and is a certain distance from human perception. Therefore, in the third step, the IQA database and labels are fed for fine-tuning, which is regarded as the second transfer learning. The LIVE In the Wild Image Quality Challenge database which contains the various type of authentically distorted image is selected to adjust the network. Therefore, the model can match the relationship between the artificial perception and the image distortion.

Flow chart is shown as Figure 1.

**Figure 1.** A flowchart of the expanding the proxy labels algorithm (EPL). (1) Generate large number of proxy labels attached to the image in ImageNet [14], using the FRIQUEE algorithm; (2) train VGG16 using the proxy label; (3) fine-tune the model using IQA data, in our method is LIVE In the Wild Image Quality Challenge database [15]. Ultimately, the model can assess the image quality and score images. In the first stage training, the values next to the images are the proxy labels generated by the computer. As for the fine-tuning stage, the values are the quality score given by observers in LIVE In the Wild Image Quality Challenge database [15].

3.2. Generating Proxy Labels

Researchers have established some IQA standard databases, in which the image quality scores are given by the amounts of observers. This kind of labeling is regarded as human labeling. However,
the data size is too limited when fed into a deep neural network, and always leads to an unsatisfied performance. In order to expand the amount of available data for training, this work takes the objective methods by computer instead of manual labeling. In order to distinguish from human labels, we define the score, which a computer gives using a traditional NR-IQA algorithm, as “proxy labels”. “Proxy” means that the objective results are obtained by machine after the extraction and assessment of the image features, instead of human judging. By expanding the labels, the amount of data can be effectively increased, so that the network can fully learn the relevant features of IQA during the training process and avoid the over-fitting problem caused by insufficient data volume.

Since this work is to perform image quality assessment on authentic distorted images, a conventional IQA method specific for authentic distortion is considered when selecting an algorithm for augmenting the proxy labels. When automatically scoring a large number of original images, selecting the algorithm with the best performance to maximize the training accuracy and success rate. At present, in many traditional algorithms, FRIQUEE [34] is one of the best algorithms for authentic distortion image, as shown in Table 2. Therefore, in this work, the FRIQUEE algorithm is chosen as a tool for generating proxy labels.

Table 2. Performance of some algorithms on the LIVE CHALLENGE DATABASE.

| Algorithms | SROCC | PLCC |
|------------|-------|------|
| FRIQUEE [34] | 0.70  | 0.66 |
| BRISQUE [28] | 0.61  | 0.58 |
| DIIVINE [29] | 0.59  | 0.56 |
| BLIINDSS [30] | 0.45  | 0.40 |
| NIQE [32] | 0.48  | 0.42 |
| C-DIIVINE [33] | 0.66  | 0.63 |

Although the proxy labels do not completely match human perception, they have relevant low-level features of image quality. After training large numbers of proxy labels, the model does not directly meet the purpose of imitating people’s perception but does learn the characteristics of distorted images and lays a good foundation for subsequent model adjustment.

3.3. Training Network Using Proxy Labels

In this work, the VGG16 [26] network is selected for training and testing. The VGG network is a model proposed by Oxford University, in 2014. Because its structure is intuitive and easy to understand, it has been applied in various fields and it has become a widely used convolutional neural network model. The VGG16 is a model for image classification tasks, and it shows excellent performance in classification and other image processing areas. Meanwhile, lots of IQA researches based on the VGG16 perform quite well, which indicates that VGG16 is also suitable for IQA. When comparing with other DNN models such as ResNet [48] and AlexNet [10], VGG16 has a moderate structure depth, quite fast operation speed, and good performance. While ensuring the training and computing efficiency, it also guarantees competitive performance.

In the original architecture, the last layer of the network is 1000 channels, corresponding to 1000 categories in image recognition. For the single score output of IQA, the number of channels in the last layer is modified to one to meet the needs of the single-class output in this work.

During the initial training process, the model learns the relevant features of IQA through a large amount of data training. Since the VGG16 network requires that size of input image is $224 \times 224$, first, a large number of original images are subjected to size filtering before input to qualify images with sizes greater than $224 \times 224$ are retained. Random sampling is adopted at the beginning of VGG16, which ensures any image bigger than $224 \times 224$ is randomly cropped to $224 \times 224$. Therefore, despite the different aspect ratios and size, all images fed into VGG16 are $224 \times 224$. The images and their proxy labels are fed into the VGG16 network, and the first stage of training is performed. During the
process, due to the existence of many proxy labels relevant to IQA, the original classification features in
the network are suppressed, and the features related to IQA are learned and enhanced. After this
training, the network already has the function of quality assessment. Nevertheless, since the data used
for training is objective data obtained by computer labeling, it is different from the actual perception of
human beings, therefore, it is necessary to continually train. According to the fundament, the IQA
database is used to fine-tune the network.

3.4. Fine-Tuning the Network

After the preliminary training, the network has the ability to evaluate the quality of images,
but the ability is based on objective methods, with some differences existing as compared with people’s
true feelings. In order to narrow the difference, the second phase of training is carried out, that is,
the network is fine-tuned using the IQA database.

In the existing IQA database, most databases are oriented to synthetic distortion. The LIVE In the
Wild Image Quality Challenge database [15] is an IQA database of authentic distortion images that fit
into this work. Among them are 1162 distorted images, containing various types of distortion which
have been captured by various representative modern mobile devices.

In the second phase of training, the model is fine-tuned using the authentic distortion images and
corresponding human labels. The characteristics of IQA in the network are further enhanced, making
the model more accurate. In addition, more importantly, after fine-tuning, the mapping between
features and scores is closer to a human’s perception of distortion, making the model more practical.

4. Experimental Results

In this section, the performance of the algorithm is demonstrated based on several experiments.
When compared with the basic algorithm and the most advanced algorithm in the NR-IQA field,
the proposed algorithm shows good performance on the authentically distorted images.

4.1. Database

We use two types of databases in the experiment, database for augmenting proxy labels and IQA
database for model tuning and testing.

4.1.1. Database for Augmenting

The database used to expand the data size is the ImageNet database. ImageNet is a large image
dataset set up to facilitate the development of computer image recognition technology [14]. This image
set contains a large number of real scene images, which can meet both the authentic distortion and
the large amount of data required in this work. In this image set, most of the images are taken in life.
The samples of ImageNet are shown in Figure 2.

What we can observe easily is that the image of the actual scene does have a certain degree
of distortion. Therefore, why the VGG model performs well in authentic distortion is explained
because its training set contains a large number of real images which are close to the distorted image in
IQA database.

4.1.2. IQA Database

The subjective IQA database used in this work is to fine-tune the model and verify the performance
of the model, where each image is associated with its subjective score (mean opinion score, MOS).
The experiment is based on a public IQA database, the LIVE In the Wild Image Quality Challenge
database (LIVEC) [15].

This dataset is an authentically distorted images IQA database that matches this work. There are
1162 authentically distorted images and no reference images. The MOS value is obtained from the
estimated 350,000 opinion scores given by more than 8100 observers, with a range of [0,100]. The higher the image quality is, the higher the score is. The samples of LIVE are shown in Figure 3.

Figure 2. Examples of the ImageNet [14]. It contains natural scenes, human landscapes, animals, plants, and portraits, etc. Most of the images have various type and degree of distortion. Dynamic blurring of the subject and underexposure are relatively common. The according proxy labels are under the images. The according proxy labels are under the images, given by FRIQUEE [34].

Figure 3. Examples of the LIVE In the Wild Image Quality Challenge database [15]. It consists of amounts of authentically distorted images. The number based the images are MOS (mean opinion scores), given by 8100 observers.
Another is the KonIQ-10k database [49], which is an IQA dataset to date consisting of 10,073 quality scored images. Through the use of crowdsourcing, researchers obtained 1.2 million reliable quality ratings from 1459 crowd workers, with the score range of [1, 5], and the better the quality is, the higher the value is.

4.2. Experimental Protocol

4.2.1. Evaluation Protocols

In this experiment, four commonly used coefficients are used to evaluate performance: Pearson linear correlation coefficient (PLCC), Spearman rank order correlation coefficient (SROCC), Kendall rank order correlation coefficient (KROCC), and root mean squared error (RMSE).

The PLCC is used to measure the linear correlation between actual scores and predicted quality scores. Given \( N \) distorted images, the actual score of the \( i \)-th image is represented by \( y_i \), and the predicted score of the network is represented by \( \hat{y}_i \). The PLCC is calculated as follows:

\[
LCC = \frac{\sum_{i=1}^{N} (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N} (y_i - \bar{y})^2 \sum_{i=1}^{N} (\hat{y}_i - \bar{\hat{y}})^2}},
\]

(1)

where \( \bar{y} \) and \( \bar{\hat{y}} \) are the average of the actual score and the predicted quality score, respectively. And PLCC is expected to be higher.

The SROCC measures the monotonic relationship between the actual score and the predicted score. Given \( N \) distorted images, the actual score of the \( i \)-th image is represented by \( y_i \), and the predicted score of the network is represented by \( \hat{y}_i \). The SROCC is calculated as follows:

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

(2)

where \( v_i \) is the rank of the true score \( y_i \) and \( p_i \) is the rank of the predicted output score \( \hat{y}_i \), for all images. The higher the SROCC is, better the performance is.

The KROCC also measures the monotonic. Given \( N \) distorted images, the actual score of the \( i \)-th image is represented by \( x_i \), the predicted score of the network is represented by \( y_i \), thus, \( n \) pairs like \( (x_1, y_1) \) and \( (x_2, y_2) \) can be obtained. Then, we randomly select two data pairs \( (x_i, y_i) \) and \( (x_j, y_j) \) to form a pair \( [(x_i, y_i), (x_j, y_j)] \). Thus, we get \( N(N-1)/2 \) pairs. If \( x_i > y_i \) and \( x_j < y_j \), or \( x_i < y_i \) and \( x_j > y_j \), we call such a pair concordant. If \( x_i > y_i \) and \( x_j > y_j \), or \( x_i < y_i \) and \( x_j < y_j \), the pair is defined discordant. \( N_c \) is the number of concordant pairs, and \( N_d \) represents the discordant. KROCC is calculated as follows:

\[
KROCC = \frac{2(N_c - N_d)}{N(N-1)},
\]

(3)

therefore, KROCC can evaluate the rank correlation and is expected to be higher.

The RMSE compares the absolute error between the algorithm’s predicted value and the ground truth to measure the accuracy of the algorithm and it is used in many works [50]. Given \( N \) distorted images, the actual score of the \( i \)-th image is represented by \( y_i \), and the predicted score of the network is represented by \( \hat{y}_i \). The RMSE is calculated as follows:

\[
RMSE = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N}},
\]

(4)

and RMSE is expected to be lower.
4.2.2. Evaluation Process

After the first transfer learning, we evaluate the performance of the model. The experiment can verify whether the model is oriented in the correct direction after the expansion of the proxy labels and training, and whether the image quality can be assessed reasonably.

After the completion of the entire training process, the network performance is verified again and compared with other algorithms to prove the correctness of the algorithm. The result verifies the validity of the extended proxy labels and the two trainings.

Since this method is based on a deep network, the LIVEC and KonIQ-10k database is randomly divided into two groups, a training set and a test set. In the experiment, 80% of the images in the database were used as training sets and the remaining 20% were used as test sets. The contents of all images are different; thus, the training and test sets are randomly selected. This random process is repeated ten times to eliminate the bias. For each iteration, the training and test sets were randomly selected as described above.

To find the proper amount of the initial training data, we perform our algorithm based on 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, and 10,000 images with their proxy labels. The average results of the obtained SROCC, PLCC, KENDALL, and RMSE values are reported for comparison. To ensure the stability of the model, we also compute the variance of the four metrics on different trained models.

In comparison, the performance of the EPL method based on the most proper amount of initial training data is compared with the most advanced NR-IQA methods, including: classical NR-IQA methods (BLIINDSS [30], BRISQUE [28], BWS [5], CORNIA [31], GMLOG [51], IL-NIQE [6], and FRIQUEE [34]), and DNN-based NR-IQA methods (CNN [12], RankIQA [23], BIECON [20], DIQaM [17], DIQA [22], CaHFI [52], NRVPD [53], ESD [54], VS-DDON [55], NQS-GAN [56], and ILGNet [57]). This method was also compared with the well-known DNN models, AlexNet [10], ResNet50 [48], and VGG-16 [26], which were modeled using the LIVEC database.

4.3. Experiment Results

For the sake of convenience, the algorithm proposed is abbreviated as EPL (expanding the proxy labels).

4.3.1. Proper Amount of Proxy Labels

One thousand, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000 and 10,000 images with their proxy labels are using to train 10 VGG16 models. Here, we show the mean and variance of SROCC, PLCC, KENDALL, and RMSE, as shown in Table 3. In Table 3, we divide the model into two types: Base and Base + FT (FT means fine-tune). Base means that the model is only trained by the proxy labels, whereas Base + FT means the model is trained completely.

In Figure 4, the line charts are more intuitively represented. The x-axes indicate the numbers of proxy labels we use in the first stage of training, and the y-axes indicate to the value of the different coefficients. For the SROCC, PLCC, and KENDALL, the values are expected to be higher, whereas the RMSE is expected to be lower. From the table and the figure, we can observe, when the training data is more than 5000 images with their proxy labels, the performance of the base model stays at a level, and there is no significant improvement. However, for the completely trained model, when the amount is 7000, the performance reached the best point overall. Therefore, the proper number of proxy labels is 7000, and we choose the model trained by 7000 proxy labels to compare with other algorithms.
According to the comparison of the performance on Base and Base + FT models, as shown in Figure 5, the performance has indeed greatly improved after fine-tuning no matter the number of proxy labels which demonstrates the effectiveness of the two times transfer learning. Since the final performance is based on expanding the proxy labels, thus, it also demonstrates the effectiveness of expanding the proxy labels.

Meanwhile, the variance is low on both the Base and Base + FT models which demonstrate the stability of the model we trained.

---

**Table 3.** Performance comparison between different amount of training data in the first transfer learning on LIVEC [15].

| Amount | Model   | SROCC Mean | SROCC Variance | PLCC Mean | PLCC Variance | KENDALL Mean | KENDALL Variance | RMSE Mean | RMSE Variance |
|--------|---------|------------|----------------|-----------|---------------|---------------|-----------------|-----------|--------------|
| 1000   | Base    | 0.683      | 0.006          | 0.664     | 0.005         | 0.479         | 0.005           | 18.177    | 0.064        |
|        | Base + FT | 0.799      | 0.004          | 0.824     | 0.003         | 0.605         | 0.004           | 13.095    | 0.087        |
| 2000   | Base    | 0.686      | 0.006          | 0.694     | 0.005         | 0.492         | 0.005           | 17.893    | 0.066        |
|        | Base + FT | 0.836      | 0.004          | 0.841     | 0.005         | 0.65          | 0.005           | 12.764    | 0.129        |
| 3000   | Base    | 0.714      | 0.005          | 0.707     | 0.005         | 0.514         | 0.005           | 17.939    | 0.063        |
|        | Base + FT | 0.778      | 0.004          | 0.800     | 0.004         | 0.581         | 0.005           | 14.089    | 0.085        |
| 4000   | Base    | 0.719      | 0.005          | 0.708     | 0.003         | 0.514         | 0.004           | 17.867    | 0.041        |
|        | Base + FT | 0.796      | 0.004          | 0.811     | 0.004         | 0.603         | 0.004           | 13.366    | 0.096        |
| 5000   | Base    | 0.755      | 0.004          | 0.744     | 0.005         | 0.542         | 0.005           | 17.184    | 0.051        |
|        | Base + FT | 0.827      | 0.005          | 0.839     | 0.004         | 0.632         | 0.006           | 12.690    | 0.111        |
| 6000   | Base    | 0.752      | 0.003          | 0.735     | 0.002         | 0.547         | 0.003           | 16.390    | 0.034        |
|        | Base + FT | 0.872      | 0.004          | 0.882     | 0.002         | 0.665         | 0.005           | 10.836    | 0.084        |
| 7000   | Base    | 0.748      | 0.004          | 0.729     | 0.005         | 0.545         | 0.005           | 16.148    | 0.062        |
|        | Base + FT | 0.870      | 0.004          | 0.890     | 0.002         | 0.688         | 0.004           | 10.616    | 0.077        |
| 8000   | Base    | 0.742      | 0.003          | 0.721     | 0.004         | 0.534         | 0.005           | 16.694    | 0.059        |
|        | Base + FT | 0.845      | 0.003          | 0.842     | 0.003         | 0.657         | 0.004           | 12.380    | 0.078        |
| 9000   | Base    | 0.757      | 0.003          | 0.747     | 0.004         | 0.553         | 0.004           | 16.177    | 0.072        |
|        | Base + FT | 0.852      | 0.003          | 0.865     | 0.003         | 0.658         | 0.004           | 11.725    | 0.078        |
| 10,000 | Base    | 0.736      | 0.003          | 0.735     | 0.003         | 0.532         | 0.003           | 16.148    | 0.059        |
|        | Base + FT | 0.843      | 0.005          | 0.849     | 0.003         | 0.649         | 0.005           | 12.201    | 0.104        |

**Figure 4.** Trend of various coefficients based on different amounts of training data.
4.3.2. Performance Comparison

Generating amounts of proxy labels is conducted to learn the low-level features applied in traditional algorithms. Therefore, we evaluated the efficiency of the model trained by proper numbers of proxy labels, that is, 7000 proxy labels. The data for ten tests, on LIVEC and KonIQ-10k, are listed in Table 4. The performance and the comparison with others are shown in Table 5.

| Time | SROCC LIVEC [15] | PLCC | KENDALL | RMSE | SROCC | PLCC | KENDALL | RMSE |
|------|------------------|------|---------|------|-------|------|---------|------|
| 1    | 0.757            | 0.736| 0.554   | 16.060| 0.656 | 0.658| 0.468   | 13.104|
| 2    | 0.751            | 0.728| 0.546   | 16.196| 0.662 | 0.664| 0.473   | 12.981|
| 3    | 0.742            | 0.728| 0.537   | 16.171| 0.657 | 0.661| 0.470   | 13.046|
| 4    | 0.744            | 0.719| 0.543   | 16.259| 0.655 | 0.657| 0.468   | 13.094|
| 5    | 0.751            | 0.730| 0.549   | 16.116| 0.657 | 0.658| 0.469   | 13.065|
| 6    | 0.750            | 0.735| 0.546   | 16.044| 0.658 | 0.659| 0.470   | 13.052|
| 7    | 0.750            | 0.732| 0.545   | 16.127| 0.656 | 0.659| 0.469   | 13.040|
| 8    | 0.748            | 0.729| 0.544   | 16.140| 0.658 | 0.660| 0.470   | 13.026|
| 9    | 0.742            | 0.725| 0.539   | 16.207| 0.657 | 0.660| 0.469   | 13.044|
| 10   | 0.749            | 0.728| 0.546   | 16.156| 0.567 | 0.660| 0.470   | 13.052|

| Mean | 0.748 | 0.729| 0.545 | 16.148| 0.657 | 0.660| 0.470   | 13.050|
| Variance | 0.004 | 0.005| 0.005 | 0.062 | 0.002 | 0.002| 0.001   | 0.034|

It can be seen from Table 5 that the method proposed in this study is superior to all conventional NR-IQA methods after the expansion of the proxy labels and initial training, that is, EPL can automatically extract not only the hand-crafted features, but also some other features relevant to image quality which are conductive to the assessment. When compared with using only hand-crafted features, the combined strategy outperforms.
Table 5. Performance comparison between the proposed algorithm (after first stage training) and other algorithms.

| Algorithm          | LIVEC [15] | KonIQ-10k [49] |
|--------------------|------------|----------------|
|                    | SROCC      | PLCC           | SROCC      | PLCC           |
| BLINDDSS [30]      | 0.463      | 0.507          | -          | -              |
| BRISQUE [28]       | 0.607      | 0.645          | -          | -              |
| CORNIA [31]        | 0.618      | 0.662          | -          | -              |
| GMLOG [51]         | 0.543      | 0.571          | -          | -              |
| IL-NIQE [6]        | 0.594      | 0.589          | -          | -              |
| BWS [5]            | 0.482      | 0.526          | -          | -              |
| FRIQUEE [34]       | 0.629      | 0.630          | -          | -              |
| AlexNet [10]       | 0.765      | 0.788          | -          | -              |
| VGG16 [26]         | 0.753      | 0.794          | -          | -              |
| ResNet50 [48]      | 0.809      | 0.826          | -          | -              |
| CNN [12]           | 0.516      | 0.536          | -          | -              |
| RankIQA [23]       | 0.641      | 0.675          | -          | -              |
| BIECON [20]        | 0.663      | 0.705          | -          | -              |
| DIQA [17]          | 0.606      | 0.601          | -          | -              |
| DIQA [22]          | 0.703      | 0.704          | -          | -              |
| CaHFI [52]         | 0.738      | 0.744          | -          | -              |
| NRVPD [53]         | 0.759      | 0.775          | -          | -              |
| ESD [54]           | -          | -              | 0.912      | 0.920          |
| VS-DDON [55]       | 0.630      | 0.705          | -          | -              |
| NQS-GAN [56]       | 0.869      | 0.893          | -          | -              |
| ILGNet [57]        | 0.363      | 0.348          | -          | -              |
| EPL (after first transfer training) | 0.748 | 0.729 |           |               |

In addition, compared with the end-to-end deep learning methods (CNN [12], RankIQA [23], BIECON [20], DIQA [17], DIQA [22], CaHFI [52], NRVPD [53], ESD [54], VS-DDON [55], NQS-GAN [56], and ILGNet [57]), since the above algorithms are mostly directed to synthetic distortion, the learning of the authentic distortion features is insufficient. Consequently, although it has not been adjusted by the IQA database, the proposed method is still superior to all the methods.

Nevertheless, when compared with the classification networks (AlexNet [10], VGG16 [26], and ResNet50 [48]) which directly use IQA data for fine-tuning, the model is inferior only after initial training. Because the above models are trained by considerable images in ImageNet and fine-tuned by IQA data, more features can be learned and mappings are closer to the human perception.

In conclusion, the model performance after first stage training is significantly better than most methods in the LIVEC database which strongly proves the effectiveness of using a large quantity of authentic distortion images for initial training.

After fine-tuning the model using LIVEC, performance has been further improved. LCC increased by 22.09% as compared with the first stage of training, and SROCC increased by 16.31%. The data of the detailed ten tests are listed in Table 6. The evaluation results are shown in Table 7.

From Table 7, we can see that the performance of this algorithm has been greatly improved by the fine-tuning of the IQA database and it outperforms the existing algorithms.

After the first stage training, although the model does not perform as well as AlexNet [10], VGG16 [26], and ResNet50 [48], the whole training leads to a competitive result. Because this model is a foundation for the first stage of training, it shows excellent performance after adjustment by the LIVEC database. The above models have a series of features relevant to object classification due to sufficient data to train; however, not all of the features are conducive to IQA. Oppositely, EPL overcomes the necklace by learning from a set of proxy labels attached to a number of images, which prompts the model to enhance the learning of the feature which is strongly relevant to image quality and suppresses the effect of the weakly relevant features. After the fine-tuning, the mapping is closer
to human perception due to the relatively proper features learned. All the processing results in the outstanding performance.

Table 6. Detailed data for 10 tests based on the Base + FT model.

| Time | LIVEC [15] | KonIQ-10k [49] |
|------|-----------|----------------|
|      | SROCC     | PLCC   | KENDALL | RMSE | SROCC | PLCC | KENDALL | RMSE |
| 1    | 0.866     | 0.887  | 0.682   | 10.741| 0.859  | 0.858 | 0.669   | 8.330 |
| 2    | 0.862     | 0.887  | 0.680   | 10.754| 0.859  | 0.859 | 0.669   | 8.307 |
| 3    | 0.871     | 0.891  | 0.689   | 10.522| 0.860  | 0.860 | 0.669   | 8.297 |
| 4    | 0.868     | 0.890  | 0.687   | 10.639| 0.860  | 0.860 | 0.669   | 8.293 |
| 5    | 0.871     | 0.891  | 0.690   | 10.590| 0.858  | 0.858 | 0.667   | 8.330 |
| 6    | 0.874     | 0.891  | 0.692   | 10.590| 0.861  | 0.857 | 0.671   | 8.326 |
| 7    | 0.875     | 0.893  | 0.693   | 10.505| 0.862  | 0.860 | 0.671   | 8.274 |
| 8    | 0.871     | 0.889  | 0.688   | 10.602| 0.860  | 0.859 | 0.669   | 8.305 |
| 9    | 0.873     | 0.890  | 0.692   | 10.587| 0.862  | 0.861 | 0.671   | 8.266 |
| 10   | 0.868     | 0.889  | 0.684   | 10.634| 0.862  | 0.862 | 0.672   | 8.251 |
| Mean | 0.870     | 0.890  | 0.688   | 10.616| 0.860  | 0.859 | 0.670   | 8.298 |
| Variance | 0.004  | 0.002  | 0.004   | 0.077 | 0.001  | 0.001 | 0.001   | 0.026 |

Table 7. Performance comparison between EPL and other algorithms.

| Algorithm  | LIVEC [15]   | KonIQ-10k [49] |
|------------|--------------|---------------|
|            | SROCC       | PLCC          |
|            | SROCC       | PLCC          |
| BLIINDSS [30] | 0.463      | 0.507  | -    | -    |
| BRISQUE [28]   | 0.607      | 0.645  | -    | -    |
| CORNIA [31]    | 0.618      | 0.662  | -    | -    |
| GMLOG [51]     | 0.543      | 0.571  | -    | -    |
| IL-NIQE [6]    | 0.594      | 0.589  | -    | -    |
| BWS [5]        | 0.482      | 0.526  | -    | -    |
| FRIQUEE [34]   | 0.629      | 0.630  | -    | -    |
| AlexNet [10]   | 0.765      | 0.788  | -    | -    |
| VGG16 [26]     | 0.753      | 0.794  | -    | -    |
| ResNet50 [48]  | 0.809      | 0.826  | -    | -    |
| CNN [12]       | 0.516      | 0.536  | -    | -    |
| RankIQA [23]   | 0.641      | 0.675  | -    | -    |
| BIECON [20]    | 0.663      | 0.705  | -    | -    |
| DIoM [17]      | 0.606      | 0.601  | -    | -    |
| DIQA [22]      | 0.703      | 0.704  | -    | -    |
| CaHFI [52]     | 0.738      | 0.744  | -    | -    |
| NRVPD [53]     | 0.759      | 0.775  | -    | -    |
| ESD [54]       | -          | -      | 0.912| 0.920|
| VS-DDON [55]   | 0.630      | 0.705  | -    | -    |
| NQS-GAN [56]   | 0.869      | 0.893  | -    | -    |
| ILGNet [57]    | 0.363      | 0.348  | -    | -    |
| **EPL**        | **0.870**  | **0.890**| 0.859| 0.860|

This proves the importance of acquiring a large number of samples when training the network. In this work, it shows the effectiveness of expanding proxy labels for deep IQA model training. Moreover, the IQA database plays an indispensable role in the training process. After fine-tuning, the network learns a more human-like mapping, and the evaluation result of the image is closer to the real experience of the human.
4.3.3. Supplementary Experiment

Since we propose the EPL method based on the requirement of IQA for authentic distortion, we perform a supplementary experiment on the LIVEC database. However, it is also meaningful to use the idea of synthetic distortion. Meantime, using other method to generate the proxy labels and evaluating the trained model on synthetic IQA database can also prove the effectiveness of proxy labels.

In order to demonstrate that the selected label-generating method has an impact on subsequent results, we select the IL-NIQE [6], a traditional NR-IQA method for synthetic distortion, and FSIMc [58], a traditional FR-IQA method for synthetic distortion, to generate proxy labels on the Waterloo exploration database [59]. FSIMc [58] is an FR-IQA method and outperforms most of NR-IQA methods, including IL-NIQE [6]. A comparison of the results between the two can prove the selected label-generating method truly has an impact. The Waterloo exploration database contains 4744 pristine natural images and 94,880 synthetic distortion images created from them. We randomly choose 20,000 images from Waterloo exploration database to generate proxy labels using IL-NIQE [6] and FSIMc [58].

After getting the proxy labels, similar to the process above, we trained a series of models, and evaluated their performance on SROCC, PLCC, KENDALL, and RMSE based on the Live [60], Tid2013 [16], CSIQ [61], and LIVEMD [62] databases. All of the databases are designed for synthetic distortions. Consistently, we evaluated the models after two times of transfer learning, found the proper number of proxy labels, and adopted the results generated by the most proper model.

We compared the results with other algorithms, as shown in Table 8.

| Algorithm     | LIVE [60] | TID2013 [16] | CSIQ [61] | LIVEMD [62] |
|---------------|-----------|--------------|-----------|--------------|
|               | SROCC   | PLCC   | SROCC   | PLCC   | SROCC   | PLCC   | SROCC   | PLCC   |
| FSIMc [58]    | 0.964   | -      | 0.931   | -      | -      | -      | -      | -      |
| BLIINDSS [30] | 0.912   | 0.916  | 0.536   | 0.628  | 0.780   | 0.832  | 0.887   | 0.902  |
| BRISQUE [28]  | 0.939   | 0.942  | 0.572   | 0.651  | 0.775   | 0.871  | 0.887   | 0.921  |
| CORNIA [31]   | 0.942   | 0.943  | 0.549   | 0.613  | 0.714   | 0.781  | 0.890   | 0.915  |
| GMLOG [51]    | 0.950   | 0.954  | 0.675   | 0.683  | 0.803   | 0.812  | 0.824   | 0.863  |
| IL-NIQE [6]   | 0.902   | 0.908  | 0.521   | 0.648  | 0.821   | 0.865  | 0.902   | 0.914  |
| BFS [5]       | 0.934   | 0.943  | 0.597   | 0.622  | 0.786   | 0.820  | 0.901   | 0.922  |
| AlexNet [10]  | 0.942   | 0.933  | 0.615   | 0.668  | 0.647   | 0.681  | 0.881   | 0.899  |
| VGG16 [26]    | 0.952   | 0.949  | 0.612   | 0.671  | 0.762   | 0.814  | 0.884   | 0.900  |
| ResNet50 [48] | 0.950   | 0.954  | 0.712   | 0.756  | 0.876   | 0.905  | 0.909   | 0.920  |
| CNN [12]      | 0.956   | 0.933  | 0.658   | 0.653  | 0.683   | 0.754  | 0.933   | 0.927  |
| RankIQA [23]  | 0.981   | 0.982  | 0.780   | 0.793  | 0.861   | 0.893  | 0.908   | 0.929  |
| BIECON [20]   | 0.961   | 0.960  | 0.717   | 0.762  | 0.815   | 0.823  | 0.909   | 0.933  |
| DIQaM [17]    | 0.960   | 0.972  | 0.835   | 0.855  | 0.869   | 0.894  | 0.906   | 0.931  |
| DIQA [22]     | 0.970   | 0.972  | 0.843   | 0.868  | 0.844   | 0.880  | 0.920   | 0.933  |
| CaiHFI [52]   | 0.965   | 0.964  | 0.862   | 0.878  | 0.903   | 0.914  | 0.927   | 0.950  |
| NRVPD [53]    | 0.943   | 0.947  | 0.683   | 0.768  | 0.840   | 0.889  | -       | -      |
| ESD [54]      | -       | -      | -       | -      | -       | -      | -       | -      |
| VS-DDON [55]  | 0.982   | 0.980  | 0.852   | 0.861  | 0.860   | 0.875  | -       | -      |
| NQS-GAN [56]  | 0.985   | 0.986  | 0.901   | 0.908  | -       | -      | -       | -      |
| ILGNet [57]   | -       | -      | -       | -      | -       | -      | -       | -      |
| Ours (using IL-NIQE) | 0.955   | 0.952  | 0.715   | 0.750  | 0.828   | 0.830  | 0.920   | 0.928  |
| Ours (using FSIMc) | 0.986   | 0.981  | 0.843   | 0.862  | 0.861   | 0.892  | 0.934   | 0.935  |

The results show the performance is indeed improved by two transfer learning as compared with the traditional method we use to generate proxy label, and comparatively to some extent. Considering that we use IL-NIQE, an NR method, to generate the proxy labels, while the other methods mostly use the FR method (BIECON [20]), or train the model with the no-distorted images, such as the comparisons (DIQA [22], DIQaM [17], RankIQA [23]), the proxy labels we obtained are less accurate than the others. The less accuracy leads to the comparatively ordinary. To improve the performance...
and demonstrate that the accuracy of proxy labels truly affects the performance, we select FSIMc [58], an FR-IQA method to also generate labels. The experimental results show that by using FSIMc to generate labels, the model performs better than most algorithms and also proves that the method of expanding proxy labels is also applicable for synthetically distorted images.

5. Conclusions

To address the scarcity of IQA data and the insufficient research on authentic distortion IQA, we have proposed a method that is based on proxy labels. First, a traditional NR-IQA algorithm is used to generate large numbers of proxy labels to expand the amount of training data. Since the labels can be generated in mass, we can train deeper and wider networks than previous work. Then, the network is trained by the proxy labels, and it learns preliminarily. Ultimately, inheriting the foundation from the previous learning, the network is fine-tuned by IQA database for closing to human perception. Experimental results on the LIVE In the Wild Image Quality Challenge database and the KonlQ-10k database demonstrate that our approach performs superior as compared with existing NR-IQA algorithms on authentic distortion.

Author Contributions: Conceptualization, X.G. and F.L.; methodology, F.L.; software, X.G.; validation, X.G. and L.H.; formal analysis, X.G. and L.H.; investigation, X.G.; resources, F.L.; data curation, X.G.; writing—original draft preparation, X.G.; visualization, X.G.; supervision, F.L.; project administration, F.L.; funding acquisition, F.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research work was supported in part by the National Science Foundation of China (61671365, U1903213), and the Natural Science Basic Research Plan in Shaanxi Province of China (2018Q6022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kwan, C.; Chou, B.; Bell, J.F., III. Comparison of Deep Learning and Conventional Demosaicing Algorithms for Mastcam Images. *Electronics* 2019, 8, 308. [CrossRef]
2. Li, F.; Fu, S.; Liu, Z.; Qian, X. A cost-constrained video quality satisfaction study on mobile devices. *IEEE Trans. Multimed.* 2018, 20, 1154–1168. [CrossRef]
3. Li, L.; Lin, W.; Wang, X.; Yang, G.; Bahrami, K.; Kot, A.C. No-Reference Image Blur Assessment Based on Discrete Orthogonal Moments. *IEEE Trans. Cybern.* 2016, 46, 39–50. [CrossRef] [PubMed]
4. Liu, H.; Klomp, N.; Heynderickx, I. A No-Reference Metric for Perceived Ringing Artifacts in Images. *IEEE Trans. Circuits Syst. Video Technol.* 2010, 20, 529–539. [CrossRef]
5. Yang, X.; Li, F.; Zhang, W.; He, L. Blind Image Quality Assessment of Natural Scenes Based on Entropy Differences in the DCT Domain. *Entropy* 2018, 20, 885. [CrossRef]
6. Zhang, L.; Zhang, L.; Bovik, A.C. A Feature-Enriched Completely Blind Image Quality Evaluator. *IEEE Trans. Image Process.* 2015, 24, 2579–2591. [CrossRef]
7. Ji, W.; Wu, J.; Zhang, M.; Liu, Z.; Shi, G.; Xie, X. Blind Image Quality Assessment with Joint Entropy Degradation. *IEEE Access* 2019, 7, 30925–30936. [CrossRef]
8. Wu, Q.; Li, H.; Ngn, K.N.; Ma, K. Blind Image Quality Assessment Using Local Consistency Aware Retriever and Uncertainty Aware Evaluator. *IEEE Trans. Circuits Syst. Video Technol.* 2018, 28, 2078–2089. [CrossRef]
9. Zhou, Y.; Li, L.; Wang, S.; Wu, J.; Fang, Y.; Gao, X. No-Reference Quality Assessment for View Synthesis Using DoG-Based Edge Statistics and Texture Naturalness. *IEEE Trans. Image Process.* 2019, 28, 4566–4579. [CrossRef]
10. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* 2017, 60, 84–90. [CrossRef]
11. Liang, Y.; Wang, J.; Wan, X. Image Quality Assessment Using Similar Scene as Reference. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; Volume 9909, pp. 3–18.
12. Kang, L.; Ye, P.; Li, Y.; Doermann, D. Convolutional Neural Networks for No-Reference Image Quality Assessment. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1733–1740.
13. Kang, L.; Ye, P.; Li, Y.; Doermann, D. Simultaneous Estimation of Image Quality and Distortion via Multitask Convolutional Neural Networks. In Proceedings of the 2015 IEEE International Conference on Image Processing, Quebec City, QC, Canada, 27–30 September 2015; pp. 2791–2795.

14. Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; Li, F. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; Volume 1–4, pp. 248–255.

15. Gadiyaram, D.; Bovik, A.C. Massive Online Crowdsourced Study of Subjective and Objective Picture Quality. IEEE Trans. Image Process. 2016, 25, 372–387. [CrossRef] [PubMed]

16. Ponomarenko, N.; Jin, L.; Jeremeiev, O.; Lukin, V.; Egiazarlarian, K.; Astola, J.; Vozel, B.; Chehdi, K.; Carli, M.; Battisti, F.; et al. Image database TID2013: Peculiarities, results and perspectives. Signal Process Image 2015, 30, 57–77. [CrossRef]

17. Bosse, S.; Maniry, D.; Mueller, K.-R.; Wiegand, T.; Samek, W. Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment. IEEE Trans. Image Process. 2018, 2018, 27, 206–219. [CrossRef] [PubMed]

18. Yan, Q.; Gong, D.; Zhang, Y. Two-Stream Convolutional Networks for Blind Image Quality Assessment. IEEE Trans. Image Process. 2019, 28, 2200–2211. [CrossRef]

19. Chen, Z.; Takeuchi, M.; Katto, J. A Pre-Saliency Map Based Blind Image Quality Assessment via Convolutional Neural Networks. In Proceedings of the 19th IEEE International Symposium on Multimedia (ISM), Taichung, Taiwan, 11–13 December 2017; pp. 77–82.

20. Kim, J.; Lee, S. Fully Deep Blind Image Quality Predictor. IEEE J. Sel. Top. Signal Process. 2017, 11, 206–220. [CrossRef]

21. Pan, D.; Shi, P.; Hou, M.; Ying, Z.; Fu, S.; Zhang, Y. Blind Predicting Similar Quality Map for Image Quality Assessment. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6373–6382.

22. Kim, J.; Ahn-Duc, N.; Lee, S. Deep CNN-Based Blind Image Quality Predictor. IEEE Trans. Neural Netw. Learn. Syst. 2019, 30, 11–24. [CrossRef]

23. Liu, X.; Van de Weijer, J.; Bagdonav, A.D. RankIQA: Learning from Rankings for No-reference Image Quality Assessment. In Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1040–1049.

24. Zhang, W.; Ma, K.; Yan, J.; Deng, D.; Wang, Z. Blind image quality assessment using a deep bilinear convolutional neural network. IEEE Trans. Circuits Syst. Video Technol. 2018, 28, 36–47. [CrossRef]

25. Li, Y.; Po, L.-M.; Feng, L.; Yuan, F. No-reference Image Quality Assessment with Deep Convolutional Neural Networks. In Proceedings of the IEEE International Conference on Digital Signal Processing, Beijing, China, 16–18 October 2016; pp. 685–689.

26. Simonyan, K.; Zisserman, A.J.C.S. Very Deep Convolutional Networks for Large-Scale Image Recognition. 2014. Available online: https://arxiv.org/abs/1409.1556. (accessed on 4 September 2014).

27. Kahaki, S.M.M.; Nordin, M.J.; Ashtari, A.H.; Zahra, J.S. Invariant Feature Matching for Image Registration Application Based on New Dissimilarity of Spatial Features. PLoS ONE 2016, 11, e0149710.

28. Mittal, A.; Moorthy, A.K.; Bovik, A.C. No-Reference Image Quality Assessment in the Spatial Domain. IEEE Trans. Image Process. 2012, 21, 4695–4708. [CrossRef]

29. Moorthy, A.K.; Bovik, A.C. Blind Image Quality Assessment: From Natural Scene Statistics to Perceptual Quality. IEEE Trans. Image Process. 2011, 20, 3350–3364. [CrossRef]

30. Saad, M.A.; Bovik, A.C.; Charrier, C. Blind Image Quality Assessment: A Natural Scene Statistics Approach in the DCT Domain. IEEE Trans. Image Process. 2012, 21, 3339–3352. [CrossRef] [PubMed]

31. Ye, P.; Kumar, J.; Kang, L.; Doermann, D. Unsupervised Feature Learning Framework for No-reference Image Quality Assessment. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 1098–1105.

32. Mittal, A.; Soundararajan, R.; Bovik, A.C. Making a “Completely Blind” Image Quality Analyzer. IEEE Signal Process. Lett. 2013, 20, 209–212. [CrossRef]

33. Zhang, Y.; Moorthy, A.K.; Chandler, D.M.; Bovik, A.C. C-DIIVINE: No-reference image quality assessment based on local magnitude and phase statistics of natural scenes. Signal Process. Image Commun. 2014, 29, 725–747. [CrossRef]

34. Gadiyaram, D.; Bovik, A.C. Perceptual quality prediction on authentically distorted images using a bag of features approach. J. Vis. 2017, 17, 32. [CrossRef] [PubMed]
35. Ye, P.; Doermann, D. No-Reference Image Quality Assessment Using Visual Codebooks. *IEEE Trans. Image Process.* 2012, 21, 3129–3138.

36. Zhang, P.; Zhou, W.; Wu, L.; Li, H. SOM: Semantic Obviousness Metric for Image Quality Assessment. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2394–2402.

37. Meng, F.; Guo, L.; Wu, Q.; Li, H. A New Deep Segmentation Quality Assessment Network for Refining Bounding Box Based Segmentation. *IEEE Access* 2019, 7, 59514–59523. [CrossRef]

38. Bosse, S.; Maniry, D.; Wiegand, T.; Sameki, W. A Deep Neural Network for Image Quality Assessment. In Proceedings of the 2016 IEEE International Conference on Image Processing, Phoenix, AZ, USA, 25–28 September 2016; pp. 3773–3777.

39. Wu, J.; Zeng, J.; Dong, W.; Shi, G.; Lin, W. Blind image quality assessment with hierarchy: Degradation from local structure to deep semantics. *J. Vis. Commun. Image Represent.* 2019, 58, 353–362. [CrossRef]

40. Migdadi, A.; Bovik, A.C. Blind Image Quality Assessment on Real Distorted Images using Deep Belief Nets. In Proceedings of the IEEE Global Conference on Signal and Information Processing, Atlanta, GA, USA, 3–5 December 2014; pp. 946–950.

41. Lin, H.; Hosu, V.; Saupe, D. KonIQ-10k: Towards an Ecologically Valid and Large-Scale IQA Database. 2018. Available online: http://database.mmsp-kr.de/koniq-10k-database.html (accessed on 22 March 2018).

42. Kahaki, S.M.M.; Arshad, H.; Nordin, M.J.; Ismail, W. Geometric feature descriptor and dissimilarity-based registration of remotely sensed imagery. *PloS ONE* 2018, 13, e0200676. [CrossRef] [PubMed]

43. Xue, W.; Mou, X.; Zhang, L.; Bovik, A.C.; Feng, X. Blind Image Quality Assessment Using Joint Statistics of Gradient Magnitude and Laplacian Features. *IEEE Trans. Image Process.* 2014, 23, 4850–4862. [CrossRef] [PubMed]

44. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

45. Lin, H.; Hosu, V.; Saupe, D. KonIQ-10k: Towards an Ecologically Valid and Large-Scale IQA Database. 2018. Available online: http://database.mmsp-kr.de/koniq-10k-database.html (accessed on 22 March 2018).

46. Kahaki, S.M.M.; Arshad, H.; Nordin, M.J.; Ismail, W. Geometric feature descriptor and dissimilarity-based registration of remotely sensed imagery. *PloS ONE* 2018, 13, e0200676. [CrossRef] [PubMed]
56. Yang, H.; Shi, P.; Zhong, D.; Pan, D.; Ying, Z. Blind Image Quality Assessment of Natural Distorted Image Based on Generative Adversarial Networks. *IEEE Access* 2019, 7, 179290–179303. [CrossRef]

57. Yang, D.; Peltokeito, V.; Kamarainen, J. CNN-Based Cross-Dataset No-Reference Image Quality Assessment. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019.

58. Zhang, L.; Zhang, L.; Mou, X.; Zhang, D. FSIM: A Feature Similarity Index for Image Quality Assessment. *IEEE Trans. Image Process.* 2011, 20, 2378–2386. [CrossRef]

59. Ma, K.; Duanmu, Z.; Wu, Q.; Wang, Z.; Yong, H.; Li, H.; Zhang, L. Waterloo Exploration Database: New Challenges for Image Quality Assessment Models. *IEEE Trans. Image Process.* 2017, 26, 1004–1016. [CrossRef]

60. Sheikh, H.R.; Sabir, M.F.; Bovik, A.C. A Statistical Evaluation of Recent Full Reference Quality Assessment Algorithms. *IEEE Trans. Image Process.* 2006, 15, 3440–3451. [CrossRef]

61. Larson, E.C.; Chandler, D.M. Most apparent distortion: Full-reference image quality assessment and the role of strategy. *J. Electron. Imaging* 2010, 19, 011006.

62. Jayaraman, D.; Mittal, A.; Moorthy, A.K.; Bovik, A.C. Objective Quality Assessment of Multiply Distorted Images. In Proceedings of the Conference Record of the Asilomar Conference on Signals Systems and Computers, Pacific Grove, CA, USA, 4–7 November 2012; pp. 1693–1697.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).