


## Research Article

# Blind Image Quality Assessment via Multiperspective Consistency

Ning Guo,<sup>1</sup> Letu Qingge,<sup>2</sup> YuanChen Huang,<sup>1</sup> Kaushik Roy,<sup>2</sup> YangGui Li,<sup>3</sup> and Pei Yang <sup>1</sup>

<sup>1</sup>Department of Computer Technology and Application, Qinghai University, Xining, China

<sup>2</sup>Department of Computer Science, North Carolina A & T State University, Greensboro, USA

<sup>3</sup>State Key Laboratory of Plateau Ecology and Agriculture, Qinghai University, Xining, China

Correspondence should be addressed to Pei Yang; yangpeinmgdx@sina.com

Received 30 November 2022; Revised 19 May 2023; Accepted 3 June 2023; Published 26 July 2023

Academic Editor: Alexander Hošovský

Copyright © 2023 Ning Guo et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Blind image quality assessment (BIQA) has made significant progress, but it remains a challenging problem due to the wide variation in image content and the diverse nature of distortions. To address these challenges and improve the adaptability of BIQA algorithms to different image contents and distortions, we propose a novel model that incorporates multiperspective consistency. Our approach introduces a multiperspective strategy to extract features from various viewpoints, enabling us to capture more beneficial cues from the image content. To map the extracted features to a scalar score, we employ a content-aware hypernetwork architecture. Additionally, we integrate all perspectives by introducing a consistency supervision strategy, which leverages cues from each perspective and enforces a learning consistency constraint between them. To evaluate the effectiveness of our proposed approach, we conducted extensive experiments on five representative datasets. The results demonstrate that our method outperforms state-of-the-art techniques on both authentic and synthetic distortion image databases. Furthermore, our approach exhibits excellent generalization ability. The source code is publicly available at <https://github.com/gn-share/multi-perspective>.

## 1. Introduction

Nowadays, the digital images have become a crucial media format in people's daily life, thanks to the widespread of intelligent devices. However, various distortions can occur during the image capture, processing, and transmission processes, making image quality assessment (IQA) an urgent need. IQA methods can generally be categorized into subjective image quality assessment and objective image quality assessment [1]. Subjective image quality assessment is reliable and accurate as it relies on human participation. However, it is also time-consuming and laborious. Therefore, considerable effort has been dedicated to objective image quality assessment in past decades [2–6]. The goal of objective image quality assessment is to explore image quality perception models that conform to the human vision system (HVS). Based on the availability of reference images, objective image quality assessment methods can be further divided into three categories, namely, full-reference IQA (FR-IQA) [7, 8], reduced-reference IQA (RR-IQA) [9], and no-reference IQA (NR-IQA) [10]. FR-IQA and RR-IQA

models utilize either the entire or part of the pristine images to predict the quality scores, which usually perform well [11]. However, their application scenarios are very limited as reference images are not available in most cases. On the other hand, the NR-IQA predicts image quality without any pristine image information. Despite being the most challenging problem in IQA, blind image quality assessment (BIQA) continues to attract significant attention due to its wide range of applications [12].

In addition to the absence of reference images, the existing datasets for blind image quality assessment (BIQA) exhibit diverse image contents and distortions. Figure 1 shows several sample images from LIVE Challenge (LIVEC) and LIVE datasets. It is evident that the LIVEC dataset, comprising authentic distortion images, encompasses a wide range of content, including indoor and outdoor scenes, day and night scenarios, as well as natural and artificial landscapes. Similarly, the synthetic distortion images in the LIVE dataset demonstrate significant differences compared to authentic distortions and cover various categories. The diversity in both distortion and image content



FIGURE 1: Images from LIVE Challenge (LIVEC) and LIVE datasets. Images on the left are authentic distorted samples, which contain various content. Images on the right are synthetic distorted samples, including JPEG/JPEG2K compression distortion, white noise, fast fading, and Gaussian blur.

variation further amplifies the challenge associated with the BIQA problem. Firstly, it necessitates a more robust representation capability to effectively capture the nuances of images with diverse content. Secondly, adapting the model to encompass a broad spectrum of authentic and synthetic distortions poses significant difficulties. Over the past few decades, extensive research has focused on identifying effective quality-aware features that accurately represent image content and distortion. Early studies predominantly employed handcrafted features such as natural scene statistics (NSS) [13] and the generalized Gaussian distribution (GGD) [13]. In recent years, learning-based approaches, particularly convolutional neural network (CNN) methods [4, 14–16], have gained significant attention in BIQA research. While these studies have achieved promising results, further efforts are required to bridge the gap between BIQA methods and the human visual system (HVS) for enhanced performance.

An exemplary method that demonstrates the advantages of utilizing powerful feature learning and content-aware hyperparameter generation is HyperIQA [4]. HyperIQA leverages the ResNet-50 architecture [17], known for its robust feature learning capabilities, and incorporates a content-aware hyperparameter generation mechanism based on hypernetworks [18]. This approach surpasses the performance of state-of-the-art methods when evaluated on databases containing authentic distorted images. However, it is worth noting that HyperIQA’s performance on synthetic distorted image databases is comparatively weaker. This observation further highlights the challenge of adapting the model to handle a wide range of distortion types and characteristics. The difficulty in achieving consistent performance across various distorted images underscores the need for further advancements in BIQA research.

As the ancient Chinese poem described, “*It’s a range viewed in face and peaks viewed from the side,*” the concept of perceiving different aspects through various perspectives serves as inspiration for our proposed approach. We aim to

enhance the adaptability of our algorithms to accommodate the content variation and diverse distortions present in images. Interestingly, similar ideas can be observed in contrastive self-supervised learning algorithms [19, 20], where two augmented views of an input image are processed by two encoders to generate similarity features in an embedding space. In our approach, we deviate from contrastive self-supervised learning by employing distinct architectures to simulate different perspectives specifically tailored for the blind image quality assessment (BIQA) task. By leveraging multiple perspectives, we aim to capture a more comprehensive understanding of image quality, effectively addressing the challenges posed by content variation and diverse distortions. More specifically, we apply two different ResNet architectures to extract information from two different perspectives. Figure 2 illustrates the visualization results of partial feature maps using different networks. For the same image, different perspectives learn different cues. When using multiple perspectives, we must solve the problem of how to integrate these perspectives into the model. To effectively incorporate both the multiperspective cues and network complexity, we propose a consistency supervision strategy to integrate multiple perspectives. This strategy allows us to merge and harmonize the information from multiple perspectives. The proposed training strategy is similar to knowledge distillation [21], which is utilized in a recently proposed dual-branch semisupervised framework named SSLIQA [22]. The main difference between our model and the knowledge distillation based method is that subnetworks in our proposed model promote each other in the training process instead of using a single direction supervision. Moreover, to take the advantage of content-aware ability of hypernetworks, we employ HyperIQA as a backbone for two different perspectives.

In this paper, we present a novel approach to address the BIQA problem using a multiperspective way. The main contributions of our paper are outlined as follows:

- (1) We introduce a multiperspective approach for BIQA that enhances the adaptability of the algorithm to account for content variation and diverse distortions. By incorporating multiple perspectives, we capture a more comprehensive understanding of image quality. To simulate these perspectives, we employ different ResNet architectures, each representing a distinct viewpoint.
- (2) We devise a training strategy based on multiperspective consistency to effectively integrate the perspectives. This strategy leverages the specificity of individual perspectives and the generality achieved by considering multiple perspectives. The integration of these perspectives leads to improved assessment accuracy. Extensive experiments conducted on five representative IQA datasets validate the effectiveness and generalization ability of our proposed method. The results demonstrate significant improvements in blind image quality assessment, highlighting the advantages of our multiperspective approach.

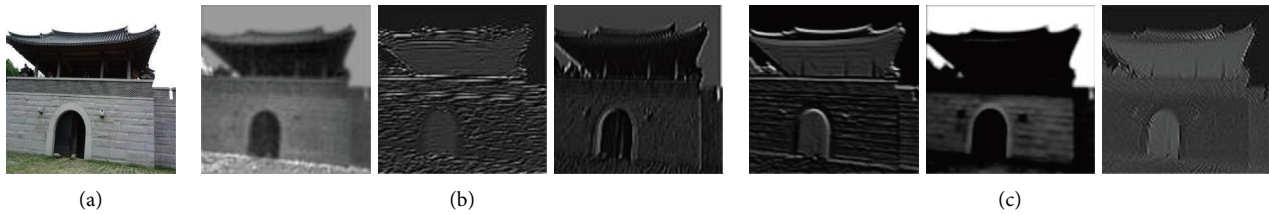


FIGURE 2: c different perspectives. (a) Original image; (b) feature maps from perspective 1 (ResNet-50); (c) feature maps from perspective 2 (ResNet-18).

## 2. Related Work

In the past two decades, many algorithms have been proposed to address the BIQA problem. These approaches can be broadly categorized into two main groups: the handcrafted feature and learning feature-based methods.

Early studies in BIQA focused on representing distorted images by designing artificial features. One common used approach was to leverage natural scene statistics (NSS) as a basis for handcrafted feature design. For example, the blind/referenceless image spatial quality evaluator (BRISQUE) [13] used NSS of locally normalized luminance coefficients to measure the unnaturalness of an image. These extracted features were then fed into a support vector regression (SVR) model to predict the image score. NIQE [23] constructed “quality-aware” collection of statistical features based on the NSS model. Zhang et al. [24] integrated the local natural image quality evaluator (ILNIQE) by incorporating more local quality-aware information into NIQE, which measures the distance between statistical features of the NSS model learned from pristine images and statistical features of the distorted image. Additionally, Jain et al. proposed a model that combined NSS with CNN and achieved promising results [25]. Multiple distributions such as generalized Gaussian distribution (GGD) [13], asymmetric generalized Gaussian distribution (AGGD) [13, 24], and histogram counting [26] were also used to capture the statistics from distorted images. Yue et al. combined statistical property, NSS-based features, and structure and texture features to predict the quality of transparently encrypted images [27]. Moreover, some works used corner descriptors (e.g., SIFT [28] and Harris [29]) to predict image quality.

Learning feature-based methods seek to automatically learn quality-aware features from images. For example, Xu et al. [30] proposed an efficient and robust BIQA model based on a high order statistics aggregation (HOSA). It was a codebook-based approach, which utilized local normalized image patches as local features and constructed codebook using K-means. Very recently, a convolutional neural network (CNN) has been adopted for BIQA and made a great progress. Additionally, Kang et al. [14] addressed the BIQA problem using a simple end-to-end CNN model consisting of one convolutional layer, one maxpooling, one min pooling layer, and three fully connected layers, which is considered to be the earliest CNN-based approach for BIQA. Kim and Lee [15] proposed a blind image evaluator based on a convolutional neural network (BIECON), which imitated

FR-IQA behavior by generating a local quality map using a deep convolutional neural network. To simultaneously handle both synthetic and authentic distortions, Zhang et al. [16] proposed a deep bilinear CNN (DB-CNN) model for BIQA. They adopted a specific CNN architecture inspired by VGGnet [31] to extract features for synthetic distortion and a tailored VGGnet for authentic distortion. VGGnet was also used to construct feature extractors in [12], in which the authors proposed weighted average deep image quality measure (WaDIQaM) for both FR-IQA and BIQA. In addition to VGGnet, AlexNet [32] and ResNet [3, 32, 33] were also typical learning feature-based approaches for BIQA.

Recent studies in blind image quality assessment (BIQA) have aimed to construct more powerful architectures to tackle the challenges in this task. Similar to DB-CNN, Yue et al. proposed a dual-branch network for screen content images’ quality assessment [34]. The original image was first decomposed into predicted and unpredicted portions, which were then fed into two branches for feature extraction. Su et al. [4] proposed a model that learns multiscale features for distorted images and estimated the quality score in a self-adaptive manner through a hyper-network. Compared with the previous supervised learning methods, Madhusudana et al. [35] considered the image quality prediction problem in a self-supervised manner. They used an unlabelled image dataset containing both synthetic and authentic distortions to train a CNN model. Furthermore, Zhang et al. [36] proposed a *unified* BIQA model optimized by a pairwise learning-to-rank training strategy to overcome the challenge of cross-distortion-scenario. Moreover, Golestaneh et al. [37] extracted both local and nonlocal features for BIQA by using a hybrid approach that benefits from CNN and self-attention mechanism in transformers. Zhang et al. [38] proposed a continual learning approach that incorporates the concept of distillation learning to address the devastating forgetfulness brought by the growth of IQA new databases. Similarly, Liu et al. [39] proposed a lifelong blind image quality assessment (LIQA) approach to effectively mitigate the catastrophic forgetting in cases of continuous distortion types and even dataset shifts.

## 3. The Proposed Method

In this paper, we present a novel approach for image quality assessment that leverages multiperspectives to better represent image content and distortion. The proposed method utilizes information from different aspects of an image to

better capture its quality characteristics. Figure 3 shows the overall architecture of our proposed method. Our model includes two content-aware subnetworks, namely, Master Network and Assistant Network. We will use Master Network and Assistant Network to learn quality prediction from two different perspectives. To construct Master Network and Assistant Network, we adopt the hypernetwork architecture of HyperIQA [4], which demonstrates powerful feature learning capabilities and content awareness.

The reason we use two subnetworks is that we design two networks to collaborate on image quality prediction, in which each network is associated with a different perspective. We name these two subnetworks, Master Network and Assistant Network, based on their roles in the test phase. More specifically, during the training stage, the Master Network and Assistant Network interact and provide each other with valuable cues from different perspectives. This collaboration allows them to assist each other to learn more cues effectively. However, in the test phase, only the Master Network is used for image quality prediction. To achieve this goal, we introduce a perspective consistency training strategy to integrate two perspectives learned from two networks for the BIQA problem. We will discuss more details about the proposed model in the following subsections.

**3.1. Multiperspective BIQA Model.** Multiperspective strategy is applied for BIQA to take more beneficial cues into consideration for the prediction task. To capture different perspectives for image quality assessment, we employ different feature extraction architectures that simulate distinct viewpoints. Specifically, we use two different ResNet modules (ResNet-50 and ResNet-18) as a feature extractor to extract features from two different perspectives to construct the Master Network and Assistant Network. The architecture details of Master Network and Assistant Network are shown in Figure 3. It can be seen that the architecture difference between the two networks only exists in the backbone network for feature extraction (ResNet-50 for Master Network and ResNet-18 for Assistant Network). It needs to be clarified again that the only role difference between Master Network and Assistant Network is that Master Network is used for quality prediction in the test phase. We denote the proposed network by  $\mathcal{H} = (\mathcal{H}^M, \mathcal{H}^A)$ , where  $\mathcal{H}^M$  and  $\mathcal{H}^A$  represent the Master Network and Assistant Network, and the superscript  $M$  and  $A$  of  $\mathcal{H}^M$  and  $\mathcal{H}^A$  stand for “Master” and “Assistant,” respectively. As shown in Figure 3, the Master Network  $\mathcal{H}^M$  and Assistant Network  $\mathcal{H}^A$  are structurally independent of each other. Both networks are integrated through perspective consistency constraints. Since  $\mathcal{H}^M$  and  $\mathcal{H}^A$  share the similar structures, we apply the unified notation  $\mathcal{H}^{\mathcal{C}}$ , where  $\mathcal{C} \in \{M, A\}$  represents  $\mathcal{H}^M$  and  $\mathcal{H}^A$ . Given an input image  $X \in \mathbb{R}^{W \times H \times C}$ , we learn the two subnetworks  $\mathcal{H}^M$  and  $\mathcal{H}^A$  so as to map the input image  $X$  to a scalar score as

$$q^{\mathcal{C}}(X) = \mathcal{H}^{\mathcal{C}}(X; \Theta^{\mathcal{C}}), \mathcal{C} \in \{M, A\}, \quad (1)$$

where  $\Theta^{\mathcal{C}}$  is the parameters of  $\mathcal{H}^{\mathcal{C}}$  (Master Network when  $\mathcal{C}$  is  $M$  and Assistant Network when  $\mathcal{C}$  is  $A$ ) and  $q^{\mathcal{C}}(X) \in \mathbb{R}$  represents the scalar quality score generated by  $\mathcal{H}^{\mathcal{C}}$ .

Next, we will demonstrate more details about the proposed two subnetworks  $\mathcal{H}^{\mathcal{C}}$ ,  $\mathcal{C} \in \{M, A\}$ . From Figure 3, we can see that both the Master Network  $\mathcal{H}^{\mathcal{C}}$  ( $\mathcal{C} = M$ ) and Assistant Network  $\mathcal{H}^{\mathcal{C}}$  ( $\mathcal{C} = A$ ) are composed of a *feature extractor*  $\varphi^{\mathcal{C}}(\cdot; \Theta_1^{\mathcal{C}})$ , a *hypernetwork*  $\psi^{\mathcal{C}}(\cdot; \Theta_2^{\mathcal{C}})$  and a *target network*  $\phi^{\mathcal{C}}(\cdot; \Theta_3^{\mathcal{C}})$  with  $\Theta_1^{\mathcal{C}}$ ,  $\Theta_2^{\mathcal{C}}$ , and  $\Theta_3^{\mathcal{C}}$  as of their parameters, where  $\Theta^{\mathcal{C}} = \Theta_1^{\mathcal{C}} \cup \Theta_2^{\mathcal{C}} \cup \Theta_3^{\mathcal{C}}$ . Therefore, we rewrite  $\mathcal{H}^{\mathcal{C}}$  as  $\mathcal{H}^{\mathcal{C}} = (\varphi^{\mathcal{C}}, \psi^{\mathcal{C}}, \phi^{\mathcal{C}})$ . To extract representative features from different perspectives, ResNet-50 and ResNet-18 are adopted to construct the feature extractors for Master Network and Assistant Network, respectively. Suppose for an input image  $X$ , the output of the four stages of ResNet (conv2\_10, conv3\_12, conv4\_18, and conv5\_9 in ResNet-50 and conv2\_5, conv3\_4, conv4\_4, and conv5\_4 in ResNet-18) are denoted as  $s_2^{\mathcal{C}}(X)$ ,  $s_3^{\mathcal{C}}(X)$ ,  $s_4^{\mathcal{C}}(X)$  and  $s_5^{\mathcal{C}}(X)$ , where  $\mathcal{C} = M$  for ResNet-50 is used in Master Network and  $\mathcal{C} = A$  for ResNet-18 is used in Assistant Network, and then we can extract multiscale features for the input image using the feature extractor  $\varphi^{\mathcal{C}}$  as

$$\varphi^{\mathcal{C}}(X; \Theta_1^{\mathcal{C}}) = \text{LDA}(s_2^{\mathcal{C}}(X)) \oplus \text{LDA}(s_3^{\mathcal{C}}(X)) \oplus \text{LDA}(s_4^{\mathcal{C}}(X)) \oplus \text{GAP}(s_5^{\mathcal{C}}(X)), \quad (2)$$

where  $\oplus$  represents Concatenate operation and LDA and GAP are the *local distortion aware module* and *global average pooling*, respectively.

To cover wide image content variation, the dynamically generated parameter strategy is adopted to adaptively learn the quality perception rule according to perceived contents. The dynamic parameters in this type of network are known as hypernetworks [18]. Moreover, the hypernetwork and target network are used to form a hypernetwork for quality regression. The hypernetwork  $\psi^{\mathcal{C}}$  takes the output of the last stage of ResNet  $s_5^{\mathcal{C}}(X)$  as input. For an input image  $X$ , the output of the hypernetwork is

$$\Theta_{ho}^{\mathcal{C}} = \psi^{\mathcal{C}}(s_5^{\mathcal{C}}(X); \Theta_2^{\mathcal{C}}). \quad (3)$$

As the hypernetwork consists of three  $1 \times 1$  convolution layers and four hyperparameter modules (HPM), the computing procedure of  $\psi^{\mathcal{C}}$  is as follows:

$$\psi^{\mathcal{C}}(\cdot) = (\text{HPM}(\text{Conv} \times 3(\cdot)), \text{HPM}(\text{Conv} \times 3(\cdot))) \text{HPM}(\text{Conv} \times 3(\cdot)), \text{HPM}(\text{Conv} \times 3(\cdot)). \quad (4)$$

The target network  $\phi^{\mathcal{C}}$  takes the multiscale feature extracted by  $\varphi^{\mathcal{C}}$  as input and consists of four fully connected layers. The target network maps the multiscale feature extracted from image  $X$  to a scalar quality score as

$$q^{\mathcal{C}}(X) = \phi^{\mathcal{C}}(\varphi^{\mathcal{C}}(X; \Theta_1^{\mathcal{C}}); \Theta_3^{\mathcal{C}}). \quad (5)$$

As introduced previously, the target network and hypernetwork together form a hypernetwork. We replace parameters of  $\phi^{\mathcal{C}}$  with the output of  $\psi^{\mathcal{C}}$ . It means that we use



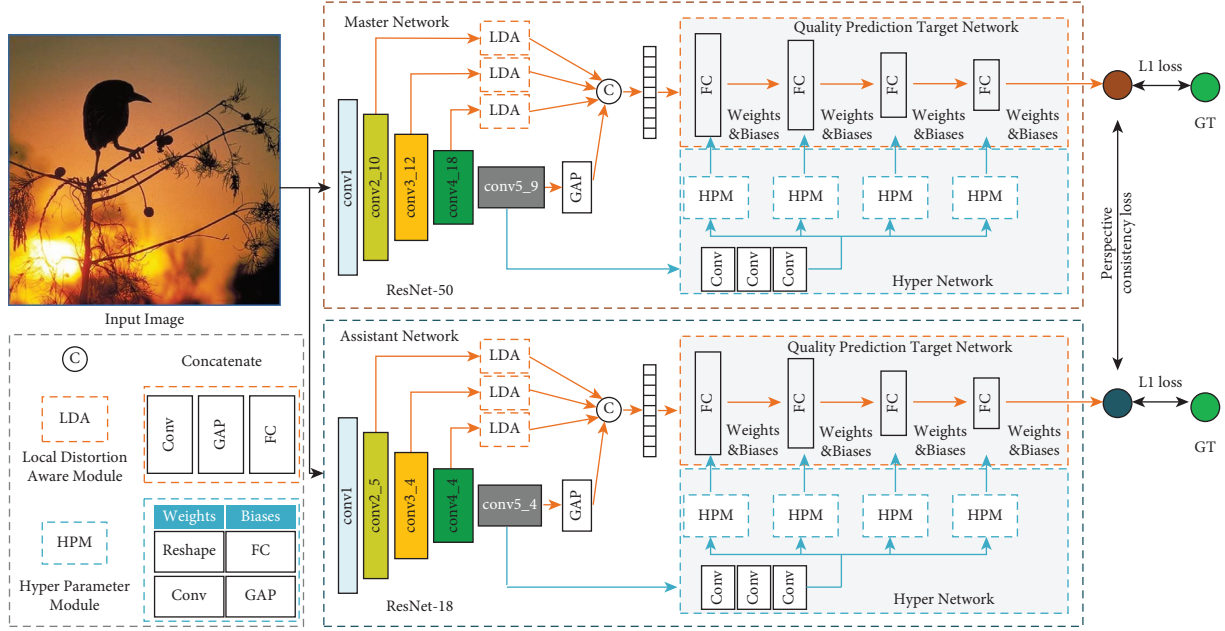


FIGURE 3: Framework of the proposed method for BIQA. The proposed model contains two subnetworks, which are master network and assistant network. Both master and assistant networks are HyperIQA network [4], and each corresponds to a different perspective. The two perspectives are integrated by using the mutual learning training strategy.

$\Theta_{ho}^{\mathcal{E}}$  instead of  $\Theta_3^{\mathcal{E}}$  in equation (5). Therefore, equation (5) can be rewritten as

$$\begin{aligned} q^{\mathcal{E}}(X) &= \phi^{\mathcal{E}}(\varphi^{\mathcal{E}}(X; \Theta_1^{\mathcal{E}}); \Theta_{ho}^{\mathcal{E}}) \\ &= \phi^{\mathcal{E}}(\varphi^{\mathcal{E}}(X; \Theta_1^{\mathcal{E}}); \psi^{\mathcal{E}}(s_s^{\mathcal{E}}(X); \Theta_2^{\mathcal{E}})). \end{aligned} \quad (6)$$

Thus, given an input image  $X$ , equation (6) is to compute the scalar quality scores  $q^M(X)$  and  $q^A(X)$  for the Master Network  $\mathcal{H}^M$  and Assistant Network  $\mathcal{H}^A$ , respectively.

**3.2. Multiperspective Consistency-Based Model Training.** The main idea behind our proposed method is to use cues learned from different perspectives for image quality assessment. To meet this goal, we need to not only consider the specific features of the individual perspective but also effectively use the generality cues of different perspectives. Based on such requirements, we will design a training objective function for the proposed network  $\mathcal{H}$ .

Let  $\mathcal{D} = \{X_i, q_i | i = 1, \dots, N\}$  be a training set, where  $X_i$  is the  $i$ -th training image and  $q_i$  represents the ground true (mean opinion scores (MOS) or different mean opinion scores (DMOS)) for  $X_i$ . Then, to use the specificity features of each perspective, we train both Master Network  $\mathcal{H}^M$  and Assistant Network  $\mathcal{H}^A$  to predict that scalar scores are as close to the ground true scores as possible. We use 1-norm to evaluate the distance between the predicted score and the ground truth and obtain the perspective specificity loss term  $\mathcal{L}_S^{\mathcal{E}}$  for individual subnetworks as

$$\begin{aligned} \mathcal{L}_S^{\mathcal{E}} &= \frac{1}{N} \sum_{i=1}^N \|q^{\mathcal{E}}(X_i) - q_i\|_1 \\ &= \frac{1}{N} \sum_{i=1}^N |q^{\mathcal{E}}(X_i) - q_i|. \end{aligned} \quad (7)$$

The subscript  $S$  of  $\mathcal{L}_S^{\mathcal{E}}$  indicates *specificity*. Note that when  $\mathcal{E} = M$ , the specificity loss term  $\mathcal{L}_S^M$  is for the Master Network  $\mathcal{H}^M$ , and when  $\mathcal{E} = A$ , the specificity loss term  $\mathcal{L}_S^A$  is for the Assistant Network  $\mathcal{H}^A$ .

The generality of the two perspectives makes two subnetworks to learn unified representation for image content and distortion from different aspects. This integration strategy is a consistency constraint between perspectives. We propose a multiperspective consistency loss term  $\mathcal{L}_{PC}$ , where the subscript  $PC$  refers to *perspective consistency*, to constrain each subnetwork to learn under the supervision of each other.  $L_1$  loss is used to measure the difference between the outputs of two perspectives. Thus, the multiperspective consistency loss term  $\mathcal{L}_{PC}$  can be defined as

$$\begin{aligned} \mathcal{L}_{PC} &= \frac{1}{N} \sum_{i=1}^N \|q^M(X_i) - q^A(X_i)\|_1 \\ &= \frac{1}{N} \sum_{i=1}^N |q^M(X_i) - q^A(X_i)|. \end{aligned} \quad (8)$$

Note that both the Master Network and Assistant Network have the same consistency loss term  $\mathcal{L}_{PC}$  as follows:

$$\mathcal{L}_{PC}^M = \frac{1}{N} \sum_{i=1}^N |q^M(X_i) - q^A(X_i)| = \mathcal{L}_{PC}, \quad (9)$$

$$\mathcal{L}_{PC}^A = \frac{1}{N} \sum_{i=1}^N |q^A(X_i) - q^M(X_i)| = \mathcal{L}_{PC}.$$

After defining the specificity loss term  $\mathcal{L}_S^{\mathcal{C}}$  and the perspective consistency loss term  $\mathcal{L}_{PC}$ , we finally obtain the optimization loss function for  $\mathcal{H}^M$  and  $\mathcal{H}^A$  as equations (10) and (11), respectively.

$$\mathcal{L}^M = (1 - \lambda_1)\mathcal{L}_S^M + \lambda_1\mathcal{L}_{PC}, \quad (10)$$

$$\mathcal{L}^A = (1 - \lambda_2)\mathcal{L}_S^A + \lambda_2\mathcal{L}_{PC}. \quad (11)$$

With the previously defined loss functions for Master Network  $\mathcal{H}^M$  and Assistant Network  $\mathcal{H}^A$ , the Adam algorithm is used as an optimizer to optimize the parameters of the the proposed network  $\mathcal{H}$ . Furthermore, the training procedure is described in Algorithm 1. Adam( $\cdot$ ) in Algorithm 1 adjusts the original gradient using adaptive momentum estimation, and we update  $\Theta^M$  and  $\Theta^A$  using  $\Delta\Theta^M$  and  $\Delta\Theta^A$ . The outputs of Adam( $\cdot$ ) are illustrated in equations in Algorithm, respectively. In the procedure of Algorithm 1, the Master Network and Assistant Network play the same role during training. However, once we have obtained the trained network  $\mathcal{H}$ , only the subnetwork  $\mathcal{H}^M$  (Master Network) is used for image quality score prediction in the test phase. This is the reason we name  $\mathcal{H}^M$  as “Master Network.”

## 4. Experimental Results and Discussion

**4.1. Datasets.** To test our proposed model on both authentically and synthetically distorted images, three authentic distortion image databases including LIVE Challenge (LIVEC) [40], KonIQ-10k [41], and BID [42] and two synthetic distortion databases including LIVE [43] and CSIQ [44] are used for evaluation. The score type for the three authentic distortion image databases LIVEC, KonIQ-10k, and BID is MOS. The score type for LIVE and CSIQ is DMOS. The authentic distortion dataset LIVE contains five different types of distortion including JP2K (JPEG2000) compression, JPEG compression, White Gaussian Noise (WN), Gaussian Blurring (GB), and Fast Fading (FF). Similarly, CSIQ contains six types of distortions including JP2K compression, JPEG compression, additive White Gaussian Noise (WN), additive Pink Gaussian Noise (PN), global Contrast Decrements (CD), and Gaussian Blurring (GB). More details about image number, score range, *etc.*, of each dataset are shown in Table 1.

**4.2. Comparison Methods and Evaluation Metrics.** To evaluate the performance of our proposed model, thirteen state-of-the-art BIQA methods are selected for comparison. Among the comparison methods, ILNIQE [24] and BRISQUE [13] are handcrafted feature-based approaches. The other approaches including HOSA [30], BIECON [15], WaDIQaM [12], SFA [33], PQR [32], DB-CNN [16], HyperIQA [4], CONTRIQUE [35], UNIQUE [36], GraphIQA [45], and TReS [37] are learning feature or deep learning-based methods.

We employ two commonly used criteria, namely, Spearman’s rank-order correlation coefficient (SRCC) and Pearson’s linear correlation coefficient (PLCC) to evaluate the performance of the proposed method and compared methods. Before computing PLCC, the predicted quality scores are first processed by a four-parameter logistic regression to remove nonlinear rating, which is caused by human visual observation according to the report from the Video Quality Expert Group [46]. The better method produces a higher SRCC and PLCC ranging between  $-1$  and  $1$ . The definitions of SRCC and PLCC are as follows:

$$\text{SRCC} = 1 - \frac{6\sum_i d_i^2}{n(n^2 - 1)}, \quad (12)$$

$$\text{PLCC} = \frac{\sum_i (q_i - q_m)(\hat{q}_i - \hat{q}_m)}{\sqrt{\sum_i (q_i - q_m)^2 \sum_i (\hat{q}_i - \hat{q}_m)^2}},$$

where  $d_i$  is the rank difference between MOS and the predicted score of the  $i$ -th image and  $n$  represents the number of images.  $q_i$  and  $\hat{q}_i$  refer to MOS and the predicted score of the  $i$ -th image, respectively, and  $q_m$  and  $\hat{q}_m$  are corresponding mean values for all images.

**4.3. Implementation Details.** We train and test our model using an NVIDIA Tesla K40 Graphics Card with video memory 12 GB. The Adam optimizer with a learning rate  $2e-5$  and weight decay  $5e-4$  is employed to train the network for 15 epochs with a batch size of 48. In addition, we set parameters  $\lambda_1 = 0.3$  and  $\lambda_2 = 0.3$  throughout the experiment. We employ the same experimental protocol as HyperIQA [4]. Specifically, we split each dataset into the training set and test set by 4:1. Note that synthetic distortion image datasets LIVE and CSIQ are split into train and test sets according to reference images to avoid content overlapping. In the test phase, we randomly select  $N$  cropped subimages from each test image  $X$ . The final quality score  $q(X)$  of the test image is defined as the mean of the scores of all subimages predicted by the Master Network as follows:

$$q(X) = \frac{1}{N} \sum_{i=1}^N q^M(X_i), \quad (13)$$

```

Input:  $\mathcal{D} = \{X_i, q_i | i = 1, \dots, N\}$ , learning rate  $\eta_t^M$  and  $\eta_t^A$ , Epochs,  $\lambda_1, \lambda_2$ 
Output:  $\mathcal{H}^M, \mathcal{H}^A$ 
Initialize parameters of  $\mathcal{H}^M$  and  $\mathcal{H}^A$ : initialize  $\Theta_1^M, \Theta_1^A$  with pretrained parameters on ImageNet, initialize  $\Theta_2^M, \Theta_3^M, \Theta_2^A, \Theta_3^A$  randomly;  $t \leftarrow 0$ ;
while  $t < \text{Epochs}$  do
  while Fetch minibatch from  $\mathcal{D}$  do
    Compute parameters  $\Theta_{ho}^M, \Theta_{ho}^A$  using equation (3);
    Compute  $q^M, q^A$  for images in minibatch using equation (6);
    Compute  $\mathcal{L}^M, \mathcal{L}^A$  using equations (10) and (11), respectively;
    Compute the gradient and update  $\Theta^M$  and  $\Theta^A$  using Adam optimizer as equations:
       $\Delta\Theta^M = \text{Ada m}(\partial\mathcal{L}^M/\partial\Theta^M)$ 
       $\Theta^M \leftarrow \Theta^M + \eta_t^M \Delta\Theta^M$ 
       $\Delta\Theta^A = \text{Ada m}(\partial\mathcal{L}^A/\partial\Theta^A)$ 
       $\Theta^A \leftarrow \Theta^A + \eta_t^A \Delta\Theta^A$ 
  end
   $t \leftarrow t + 1$ ;
end
return  $\mathcal{H}^M, \mathcal{H}^A$ ;

```

ALGORITHM 1: Multiperspective Consistency-Based Model Training.

TABLE 1: Details of each image dataset.

Dataset	Distortion type	No. of images	No. of distortions	Score range
LIVEC	Authentic	1162	—	[0, 100]
KonIQ-10k	Authentic	10073	—	[0, 100]
BID	Authentic	586	—	[0, 5]
LIVE	Synthetic	29 + 779	5	[0, 100]
CSIQ	Synthetic	30 + 866	6	[0, 1]

where  $X_i$  is the cropped subimage of  $X$ , and  $N$  is set to 25 for all the test datasets. We repeat the experiment 10 times and implement the random train-test splitting operation at each time. The median SRCC and PLCC values are used as the final results. For more details, refer to our released source code at <https://github.com/gn-share/multi-perspective>.

#### 4.4. Performance Evaluation

**4.4.1. Quantitative Evaluation.** First, we conduct experiments on a single dataset and summarize the results in Table 2. The colors red, blue, and green refer to the highest, second, and third score for all comparison methods. For the three authentic distortion datasets, the results indicate that our proposed method outperforms others on LIVEC and BID. Both SRCC and PLCC of our method are only less than those of TReS. For the two synthetic distortion datasets, our proposed model achieves the best results on LIVE for both SRCC and PLCC evaluation, and the SRCC and PLCC are the second and third largest on CSIQ. We highlight that (1) our method outperforms HyperIQA on all the five test datasets; (2) our proposed model achieves the best results on all authentic distortion datasets except for KonIQ compared to the state-of-art HyperIQA [4], UNIQUE [36], CONTRIQUE [35], GraphIQA [45], and TReS [37]. Furthermore, our method performs only weaker than TReS on KonIQ; (3) our proposed method also achieves competitive results on the two synthetic distortion datasets (i.e., LIVE and CSIQ).

In particular, our model shows a significant performance improvement over HyperIQA; (4) the average SRCC and PLCC of our method are larger than those of all the compared approaches, which indicate that the overall performance of our proposed model is better than compared methods.

Then, we further conduct experiments to evaluate the performance of our approach on different distortion types of images. As not all comparison methods reported SRCC values for individual distortion, we only show the results of methods reported from [4] in Table 3. Table 3 presents the SRCC values on individual distortion of each method on LIVE and CSIQ datasets. Based on the experimental results, our approach outperforms the compared methods on four of the five distortion types on the LIVE dataset. For Gaussian blurring (GB) distortion on the LIVE dataset, our approach obtains the second largest SRCC (0.956), which is only lower than the result of BRISQUE (0.964). For the CSIQ dataset, our method achieves the best results on three of the six distortion types, while the WaDIQaM obtains the best results on two out of the six distortion types as shown in Table 3. Note that WaDIQaM has the best performance for the CSIQ dataset in Table 2. Overall, our method is more efficient compared with the other methods on the individual distortion test.

In order to validate the generalization ability of our proposed method, we run cross database tests for the performance evaluation. Due to the lack of source code and reported results, three competitive methods PQR, DB-CNN, and HyperIQA are

TABLE 2: The SRCC and PLCC values of various methods on LIVEC, BID, KonIQ, LIVE, and CSIQ datasets. Red, blue, and green refer to the best, second, and third score among all comparison methods, respectively. The values listed in the *Average* columns are the average of datasets except for BID for fair comparison.

Dataset		LIVEC		BID		KonIQ		LIVE		CSIQ		Average	
Methods	PublicationYear	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
BRISQUE [13]	2012	0.608	0.629	0.562	0.593	0.665	0.681	0.939	0.935	0.746	0.829	0.740	0.769
ILNIQE [24]	2015	0.432	0.508	0.516	0.554	0.507	0.523	0.902	0.865	0.806	0.808	0.662	0.676
HOSA [30]	2016	0.640	0.678	0.721	0.736	0.671	0.694	0.946	0.947	0.741	0.823	0.750	0.786
BIECON [15]	2016	0.595	0.613	0.539	0.576	0.618	0.651	0.961	0.962	0.815	0.823	0.747	0.753
WaDIQaM [12]	2017	0.671	0.680	0.725	0.742	0.797	0.805	0.954	0.963	0.955	0.973	0.844	0.855
PQR [32]	2017	0.857	0.882	0.775	0.794	0.880	0.884	0.965	0.971	0.873	0.901	0.894	0.910
SFA [33]	2018	0.812	0.833	0.826	0.840	0.856	0.872	0.883	0.895	0.796	0.818	0.837	0.855
DB-CNN [16]	2018	0.851	0.869	0.845	0.859	0.875	0.884	0.968	0.971	0.946	0.959	0.910	0.921
HyperIQA [4]	2020	0.859	0.882	0.869	0.878	0.906	0.917	0.962	0.966	0.923	0.942	0.913	0.927
UNIQUE [36]	2021	0.854	0.890	0.858	0.873	0.896	0.901	0.969	0.968	0.902	0.927	0.905	0.922
CONTRIQUE [35]	2022	0.845	0.857	-	-	0.894	0.906	0.960	0.961	0.942	0.955	0.910	0.920
GraphIQA [45]	2022	0.845	0.862	-	-	0.911	0.915	0.979	0.980	0.947	0.959	0.921	0.929
TReS [37]	2022	0.846	0.877	-	-	0.915	0.928	0.969	0.968	0.922	0.942	0.913	0.929
Ours		0.869	0.887	0.879	0.883	0.910	0.922	0.975	0.976	0.950	0.956	0.925	0.935

TABLE 3: SRCC values for each distortion type on LIVE and CSIQ datasets. Red, blue, and green refer to the highest, second, and third score among all comparison methods, respectively.

Database	LIVE					CSIQ					
Type	JP2K	JPEG	WN	GB	FF	WN	JPEG	JP2K	PN	GB	CD
BRISQUE [13]	0.929	0.965	0.982	0.964	0.828	0.723	0.806	0.840	0.378	0.820	0.804
ILNIQE [24]	0.894	0.941	0.981	0.915	0.833	0.850	0.899	0.906	0.874	0.858	0.501
HOSA [30]	0.935	0.954	0.975	0.954	0.954	0.604	0.733	0.818	0.500	0.841	0.716
BIECON [15]	0.952	0.974	0.980	0.956	0.923	0.902	0.942	0.954	0.884	0.946	0.523
WaDIQaM [12]	0.942	0.953	0.982	0.938	0.923	0.974	0.853	0.947	0.882	0.979	0.923
PQR [32]	0.953	0.965	0.981	0.944	0.921	0.915	0.934	0.955	0.926	0.921	0.837
DB-CNN [16]	0.955	0.972	0.980	0.935	0.930	0.948	0.940	0.953	0.940	0.947	0.870
HyperIQA [4]	0.949	0.961	0.982	0.926	0.934	0.927	0.934	0.960	0.931	0.915	0.874
Ours	0.966	0.982	0.989	0.956	0.959	0.967	0.958	0.954	0.956	0.934	0.935

selected for comparison. We use four test protocols, which are (1) *train on LIVEC and test on BID*, (2) *train on BID and test on LIVEC*, (3) *train on LIVE and test on CSIQ*, and (4) *train on CSIQ and test on LIVE*. The first two test protocols are for the authentic distortion, and the last two are for the synthetic distortion. The SRCC values of each comparison method for each test protocol are summarized in Table 4. The results show that our approach significantly outperforms the compared methods for all of four cross database test cases. Specifically, when using LIVEC for training and BID for test set, the SRCC value of our approach is 0.882, which is much higher than the second largest SRCC of 0.762 (DB-CNN).

**4.4.2. Qualitative Evaluation.** In addition, to intuitively evaluate the performance of our proposed method, we present the scoring results for the authentic distortion dataset LIVEC and synthetic distortion dataset LIVE in this section. Figure 4 shows scoring results for images of LIVEC, from which we can see that our proposed method produces remarkable results in the 1st to the 4th columns although the content of images in LIVEC varies. Some failure cases are observed, as shown in the last column of Figure 4, which

correspond to two images with serious distortion and very high quality.

Moreover, Figures 5 and 6 are scoring results for distorted images from the LIVE dataset. Figure 5 shows the predicted scores and the corresponding standard deviations of the plane images distorted by five different types of distortion. The results indicate that our method produces prediction scores close to the DMOS for images of different distortions with satisfactory standard deviations (maximum std 7.06 (std of FF distorted image) and minimum std 1.09 (std of WN distorted image)). Figure 6 presents the prediction scores for images with the same distortion (GB) but different distortion intensities. The distortion intensities increase from left to right, and our model generates prediction scores that are consistent with the expectations.

Figure 7 is the scatter plots of DMOS/MOS versus prediction scores on the test sets of LIVE and LIVEC. The blue solid line represents the fitting line, which is the fitting for all scatter points, while the red dash line represents the desired fitting line. From the scatter plots we conclude that, on one hand, these scatter points can be distributed along the desired line. On the other hand, the result of LIVE is better than that of LIVEC, which is consistent with the quantitative results shown in Table 2.



TABLE 4: SRCC comparison on cross database tests. Best results in boldface.

Train	Test	PQR	DB-CNN	HyperIQA	Ours
LIVEC	BID	0.714	0.762	0.756	<b>0.882</b>
BID	LIVEC	0.680	0.725	0.770	<b>0.795</b>
LIVE	CSIQ	0.719	0.758	0.744	<b>0.762</b>
CSIQ	LIVE	0.922	0.877	0.926	<b>0.934</b>

Values in boldface denote the best SRCC (PLCC) result in Tables 4 and 5.



FIGURE 4: Scoring results from LIVEC. Results in the left four columns are successful examples, and the last column shows the failure cases. Values in parentheses behind the predicted score are standard deviations.



FIGURE 5: Scoring results of distorted images from LIVE. Image in the left top corner is the pristine image, and other five ones correspond to distortion types FF, GB, HP2K, JPEG, and WN. Values in parentheses behind the predicted score are standard deviations.



FIGURE 6: Scoring results of distorted images containing different intensities of Gaussian blurring (the distortion intensities increase from left to right).

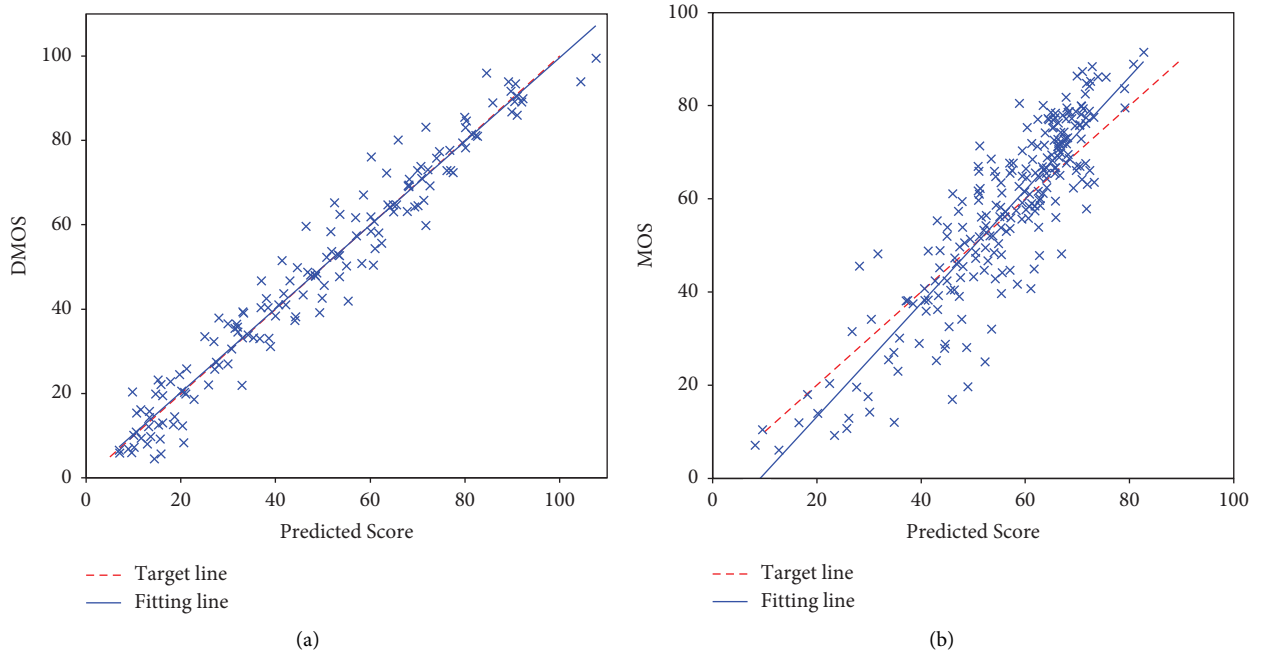


FIGURE 7: Scatter plots of DMOS/MOS versus prediction scores on the test sets of LIVE and LIVEC. The fitting line is the fitting of all scatter points, and the target line represents the desired fitting line. (a) LIVE and (b) LIVEC.

**4.5. Effect of Different Experimental Settings.** In this section, more experiments are conducted to explore the effect of different hyperparameters, backbone architectures, and training set size.

**4.5.1. Selection of Weight Parameter Value for Perspective Consistency Term.** As the loss function for  $\mathcal{H}^M$  and  $\mathcal{H}^A$  are composed of two terms, namely, the perspective specificity loss term  $\mathcal{L}_S^C$  and multiperspective consistency loss term  $\mathcal{L}_{PC}$ , we further conduct experiments to verify the performance of our proposed model using different weight values. During the implementation, we always set  $\lambda_1$  and  $\lambda_2$  as the same value. Table 5 presents the SRCC and PLCC values on LIVE and LIVEC datasets using different  $\lambda_1$  and  $\lambda_2$  ranging from 0.0 to 0.9 with a step length of 0.3. The results demonstrate both SRCC and PLCC obtain the maximum value when  $\lambda_1 = \lambda_2 = 0.3$  on both LIVE and LIVEC datasets. Note that  $\lambda_1 = \lambda_2 = 0.0$  means that the model is not using the perspective consistency constraint during the model training. The results on both LIVE and LIVEC indicate that the perspective consistency term can improve the performance

of the algorithm. However, an improper weight value may lead to degradation of model's performance (e.g., both SRCC and PLCC values corresponding to  $\lambda_1 = \lambda_2 = 0.9$  are less than those of without using perspective consistency).

**4.5.2. Evaluation of Architectures of Master Network and Assistant Network.** The network architecture significantly affects the performance of the algorithm. We also experimentally compare different network structures to test their effectiveness. We test ResNet-18 and ResNet-50 for both Master Network  $\mathcal{H}^M$  and Assistant Network  $\mathcal{H}^A$  and obtain four test cases results as shown in Table 6. The results indicate that, on one hand, the model performs better when we use ResNet-50 as the architecture of  $\mathcal{H}^M$ . On the other hand, using different architectures for  $\mathcal{H}^M$  and  $\mathcal{H}^A$  has more advantages than using the same architectures. As a result, we apply ResNet-50 and ResNet-18 for  $\mathcal{H}^M$  and  $\mathcal{H}^A$ , respectively, which generate the best testing performance both on LIVE and LIVEC as shown in Table 6. We further implement *t*-tests between average SRCC values of the above  $\mathcal{H}^M$  and  $\mathcal{H}^A$  combinations to ascertain whether the results

TABLE 5: The SRCC and PLCC corresponding to different  $\lambda_1$  and  $\lambda_2$  values on LIVE and LIVEC datasets, and we always set  $\lambda_1 = \lambda_2$  throughout all experiments.

Dataset	LIVE		LIVEC	
	SRCC	PLCC	SRCC	PLCC
$\lambda_1, \lambda_2$				
0.0	0.965	0.966	0.857	0.878
0.3	<b>0.975</b>	<b>0.976</b>	<b>0.869</b>	<b>0.887</b>
0.6	0.966	0.968	0.857	0.878
0.9	0.970	0.971	0.842	0.866

The bold values indicate the best SRCC and PLCC values.

are significant or not. The test results indicate that the architecture of  $\mathcal{H}^M = \text{ResNet-50}$  and  $\mathcal{H}^A = \text{ResNet-18}$  is significantly better than the combination of  $\mathcal{H}^M = \text{ResNet-18}$  and  $\mathcal{H}^A = \text{ResNet-50}$  ( $p = 0.0496$  on LIVE and  $p = 0.0109$  on LIVEC) and  $\mathcal{H}^M = \text{ResNet-18}$  and  $\mathcal{H}^A = \text{ResNet-18}$  ( $p = 0.0052$  on LIVE and  $p = 0.0202$  on LIVEC). However, the performance advantage of the architecture  $\mathcal{H}^M = \text{ResNet-50}$  and  $\mathcal{H}^A = \text{ResNet-18}$  over the combination of  $\mathcal{H}^M = \text{ResNet-50}$  and  $\mathcal{H}^A = \text{ResNet-50}$  is not significant, with  $p = 0.2946$  on LIVE and  $p = 0.2148$  on LIVEC.

**4.5.3. Performance of the Proposed Method Using Different Sizes of the Training Set.** In this section, we conduct experiments to verify the performance of our proposed model with different sizes of training sets. The relationship between the performance of the proposed method and the proportion of the training set is shown in Figure 8. It is obvious that both SRCC and PLCC on LIVE and LIVEC gradually decrease when the proportion of the training set in all the datasets decreases. It indicates the importance of training data. The model still performs well even when trained with 20% samples of the whole dataset, which further verifies the effectiveness of our proposed method.

**4.6. More Results for the Multiperspective Strategy on Different Backbone Networks.** To further verify the proposed method, we conduct more experiments using different backbone networks including VGGNet [31], DenseNet [47], ResNet [17], and GoogleNet [48]. Additionally, to better study the performance impact of different backbone networks, we do not use HyperNet throughout the experiments in this section. Specifically, we remove the last softmax layer of each backbone network and combine it with a three-layer MLP to form a baseline. It should be noted that we use the same MLP in the following experiments. Each layer of the MLP contains 512, 32, and 1 neurons, respectively, and ReLU is used as the activation function for the first two layers. The weights for perspective consistency terms are set to  $\lambda_1 = \lambda_2 = 0.3$ . The SRCC and PLCC values for these baseline methods on LIVE and LIVEC are listed in Table 7. The results show that (1) ResNet-50 and DenseNet-169 achieved better performance compared to other baseline algorithms, especially on dataset LIVEC, and (2) for models of the same type, deeper networks (i.e., VGG-16, DenseNet-169, and ResNet-50) perform better.

TABLE 6: The SRCC and PLCC results correspond to different architectures of  $\mathcal{H}^M$  and  $\mathcal{H}^A$  on LIVE and LIVEC datasets.

Dataset	$\mathcal{H}^A$	LIVE		LIVEC	
		SRCC	PLCC	SRCC	PLCC
$\mathcal{H}^M$					
ResNet-18	ResNet-18	0.960	0.963	0.836	0.857
ResNet-18	ResNet-50	0.966	0.968	0.834	0.871
ResNet-50	ResNet-18	<b>0.975</b>	<b>0.976</b>	<b>0.869</b>	<b>0.887</b>
ResNet-50	ResNet-50	0.972	0.973	0.859	0.871

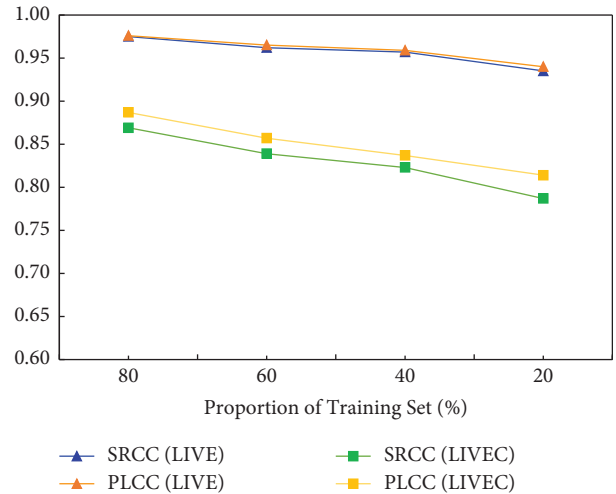


FIGURE 8: Relationship between the performance and the size of the training set.

TABLE 7: The SRCC and PLCC values for baselines using different backbone networks on LIVE and LIVEC datasets.

Dataset	LIVE		LIVEC	
	SRCC	PLCC	SRCC	PLCC
Backbone				
VGG-11	0.952	0.956	0.804	0.819
VGG-16	0.955	0.960	0.825	0.855
DenseNet-121	0.965	0.968	0.828	0.846
DenseNet-169	0.967	0.970	0.833	0.852
ResNet-18	0.961	0.965	0.827	0.850
ResNet-50	0.968	0.971	0.836	0.859
GoogleNet	0.959	0.962	0.824	0.859

To verify the effectiveness of the proposed multiperspective strategy, we conduct experiments to test the performance of different combinations of baseline networks. Table 8 shows the SRCC and PLCC values for different combinations of backbone networks. From the results in Tables 7 and 8, we conclude that the proposed multiperspective strategy has a significant improvement for all baselines (i.e., DenseNet-169, ResNet-50, GoogleNet, and VGG-16). For example, the SRCC value on LIVEC increases from 0.833 (line 6 of Table 7) to 0.857 (line 4 of Table 8) if we use ResNet-50 to assist DenseNet-169, and the SRCC value on LIVE increases from 0.959 (line 9 of Table 7) to 0.974 (line 12 and 13 of Table 8) if we use ResNet-50 or VGG-16 to assist GoogleNet.

TABLE 8: Results for different combinations of  $\mathcal{H}^M$  and  $\mathcal{H}^A$  on LIVE and LIVEC datasets.

$\mathcal{H}^M$	$\mathcal{H}^A$	LIVE		LIVEC	
		SRCC	PLCC	SRCC	PLCC
DenseNet-169	DenseNet-121	0.976	0.977	0.845	0.862
	ResNet-50	0.973	0.976	0.857	0.876
	GoogleNet	0.971	0.975	0.843	0.869
	VGG-16	0.968	0.967	0.854	0.873
ResNet-50	ResNet-18	0.975	0.976	0.853	0.861
	DenseNet-169	0.972	0.975	0.847	0.868
	GoogleNet	0.975	0.978	0.846	0.876
	VGG-16	0.972	0.972	0.841	0.855
GoogleNet	DenseNet-169	0.963	0.966	0.835	0.860
	VGG-16	0.974	0.975	0.848	0.877
VGG-16	VGG-11	0.963	0.966	0.839	0.845
	DenseNet-169	0.963	0.964	0.836	0.847
	ResNet-50	0.967	0.969	0.827	0.854
	GoogleNet	0.973	0.972	0.833	0.851

**4.7. Discussion.** The main idea of this study is to improve the performance of BIQA with valuable cues learned from different perspectives. To achieve this, we utilize different architectures to construct different perspectives and propose a multiperspective consistency-based training strategy. Specifically, architecture of HyperIQA is used to construct different perspectives. As a BIQA model designed for real-world images, HyperIQA achieves competitive performance on authentic distortion datasets. However, its performance on synthetic distorted images is relatively limited. This discrepancy can potentially be attributed to the limited content diversity present in synthetic distortion datasets, which poses challenges for effective model learning. The proposed multiperspective strategy can alleviate this challenge as it can take advantage of both specific features and generality cues of different perspectives. By considering information from different perspectives, the proposed strategy enables mutual exchange of valuable insights while also constraining each perspective to reduce the risk of overfitting.

Moreover, the selection of Master Network has a great impact on performance of our proposed model. Based on experimental results, we adopt the ResNet-50-based HyperIQA as the Master Network. The deeper networks generally have stronger representation abilities. Intuitively, selecting a deeper network as the Master Network from multiple perspectives seems reasonable. In fact, deeper networks usually have better performance with proper training configurations. The experimental results in Tables 7 and 8 support this view to some extent. The results in Table 8 show that using a deeper network as Master Network always performs better.

Lastly, while our proposed model achieves better overall performance compared to other state-of-the-art methods, it is important to acknowledge the notable advancements made by some recent models. For instance, some latest models (e.g., transformer-based TReS [37] and graphical convolutional network (GCN)-based GraphIQA [45]) show great promise, especially for large scale of data. In addition,

some early deep models still show competitive performance. For example, the overall performance of the dual-stream network DB-CNN is very close to some recent models, which is still instructive for BIQA model design.

## 5. Conclusions

In this paper, we propose a novel model for BIQA problem to increase the adaptability of the BIQA model for image content variation and a diverse range of distortions. To represent the image from different aspects, we employ a multiperspective strategy that incorporates more cues. Specifically, we present a perspective consistency constrained training strategy to integrate different perspectives effectively, considering both multiperspective cues and the complexity of the network. Extensive experimental results show that our proposed approach has a promising performance both of authentic and synthetic distorted image databases compared to the state-of-the-art methods. Moreover, generalization ability of the proposed method is also remarkable, further enhancing its practical applicability.

Despite the competitive performance of the proposed multiperspective method and recent models, there is still considerable room for improvement in effectively handling authentic images with diverse content and distortion. Moving forward, our future research will focus on exploring architectures with enhanced representational capabilities, leveraging transformer, GCN, and other advanced techniques. Additionally, we aim to conduct further investigations into the multiperspective integration strategy to enhance the model's adaptability to a wide range of content and distortion variations.

## Data Availability

All datasets supporting this study are publicly available.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China under grant no. 61866031.

## References

- [1] G. T. Zhai and X. K. Min, "Perceptual image quality assessment: a survey," *Science China Information Sciences*, vol. 63, no. 11, Article ID 211301, 2020.
- [2] 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.
- [3] L. Zhang, L. Zhang, X. Q. 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.



- [4] S. L. Su, Q. S. Yan, Y. Zhu et al., "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3667–3676, Seattle, WA, USA, June 2020.
- [5] D. Muthusamy and S. Sathyamoorthy, "Deep belief network for solving the image quality assessment in full reference and no reference model," *Neural Computing & Applications*, vol. 34, no. 24, pp. 21809–21833, 2022.
- [6] D. Muthusamy and S. Sathyamoorthy, "Feature sampling based on multilayer perceptive neural network for image quality assessment," *Engineering Applications of Artificial Intelligence*, vol. 121, Article ID 106015, 2023.
- [7] J. Lin, M. Wang, and W. Xie, "A lightweight quality assessment of screen content images using directional derivative filters," in *Proceedings of the 2018 IEEE 3rd International Conference on Signal and Image Processing (ICSIP)*, pp. 292–296, Shenzhen, China, July 2018.
- [8] Q. Jiang, W. Zhou, X. Chai, G. Yue, F. Shao, and Z. Chen, "A full-reference stereoscopic image quality measurement via hierarchical deep feature degradation fusion," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 12, pp. 9784–9796, 2020.
- [9] W. H. Zhu, G. T. Zhai, X. K. Min et al., "Multi-channel decomposition in tandem with free-energy principle for reduced-reference image quality assessment," *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2334–2346, 2019.
- [10] Y. M. Fang, H. W. Zhu, Y. Zeng, K. D. Ma, and X. Wang, "Perceptual quality assessment of smartphone photography," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3677–3686, Seattle, WA, USA, June 2020.
- [11] Z. Zhang, W. Sun, X. Min et al., "A no-reference evaluation metric for low-light image enhancement," in *Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, Shenzhen, China, July 2021.
- [12] S. Bosse, D. Maniry, K. R. Müller, 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, 2018.
- [13] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.
- [14] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1733–1740, Columbus, OH, USA, June 2014.
- [15] J. Kim and S. Lee, "Fully deep blind image quality predictor," *IEEE Journal of selected topics in signal processing*, vol. 11, no. 1, pp. 206–220, 2017.
- [16] W. X. Zhang, K. D. Ma, J. Yan, D. X. 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] K. M. He, X. Y. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, June 2016.
- [18] D. Ha, A. M. Dai, and Q. V. Le, "Hypernetworks," in *Proceedings of the International Conference on Learning Representations*, New York, NY, USA, December 2017.
- [19] J.-B. Grill, F. Strub, F. Althé et al., "Bootstrap your own latent—a new approach to self-supervised learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21271–21284, 2020.
- [20] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, New Orleans, LA, USA, August 2021.
- [21] G. Hinton, Oriol Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *Computer Science*, vol. 14, no. 7, pp. 38–39, 2015.
- [22] G. Yue, D. Cheng, L. Li, T. Zhou, H. Liu, and T. Wang, "Semi-supervised authentically distorted image quality assessment with consistency-preserving dual-branch convolutional neural network," *IEEE Transactions on Multimedia*, vol. 15, pp. 1–13, 2022.
- [23] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.
- [24] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2579–2591, 2015.
- [25] P. Jain, Gitam Shikkenawis, and S. K. Mitra, "Natural scene statistics and cnn based parallel network for image quality assessment," in *Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP)*, pp. 1394–1398, Anchorage, AK, USA, September 2021.
- [26] W. F. Xue, X. Q. Mou, L. Zhang, A. C. Bovik, and X. C. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and laplacian features," *IEEE Transactions on Image Processing*, vol. 23, no. 11, pp. 4850–4862, 2014.
- [27] G. Yue, C. Hou, K. Gu, T. Zhou, and H. Liu, "No-reference quality evaluator of transparently encrypted images," *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2184–2194, 2019.
- [28] L. Ma, L. Xu, Y. Zhang, Y. Yan, and K. N. Ngan, "No-reference retargeted image quality assessment based on pairwise rank learning," *IEEE Transactions on Multimedia*, vol. 18, no. 11, pp. 2228–2237, 2016.
- [29] I. F. Nizami, M. Majid, M. Rehman, S. M. Anwar, A. Nasim, and K. Khurshid, "No-reference image quality assessment using bag-of-features with feature selection," *Multimedia Tools and Applications*, vol. 79, no. 11–12, pp. 7811–7836, 2020.
- [30] J. T. Xu, P. Ye, Q. H. Li, H. Q. Du, Y. Liu, and D. Doermann, "Blind image quality assessment based on high order statistics aggregation," *IEEE Transactions on Image Processing*, vol. 25, no. 9, pp. 4444–4457, 2016.
- [31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the International Conference on Learning Representations*, New York, NY, USA, May 2015.
- [32] H. Zeng, L. Zhang, C. Alan, and Bovik, "Blind image quality assessment with a probabilistic quality representation," in *Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 609–613, Athens, Greece, October 2018.
- [33] D. Q. Li, T. T. Jiang, W. S. Lin, and M. Jiang, "Which has better visual quality: the clear blue sky or a blurry animal?" *IEEE Transactions on Multimedia*, vol. 21, no. 5, pp. 1221–1234, 2019.
- [34] G. Yue, C. Hou, W. Yan, L. K. Choi, T. Zhou, and Y. Hou, "Blind quality assessment for screen content images via



- convolutional neural network,” *Digital Signal Processing*, vol. 91, pp. 21–30, 2019.
- [35] P. C. Madhusudana, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, “Image quality assessment using contrastive learning,” *IEEE Transactions on Image Processing*, vol. 31, pp. 4149–4161, 2022.
- [36] W. Zhang, K. Ma, G. Zhai, and X. Yang, “Uncertainty-aware blind image quality assessment in the laboratory and wild,” *IEEE Transactions on Image Processing*, vol. 30, pp. 3474–3486, 2021.
- [37] S. A. Golestaneh, S. Dadsetan, and M. K. Kris, “No-reference image quality assessment via transformers, relative ranking and self-consistency,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1220–1230, Waikoloa, HI, USA, January 2022.
- [38] W. Zhang, D. Li, C. Ma, G. Zhai, X. Yang, and K. Ma, “Continual Learning for Blind Image Quality Assessment,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, 2022.
- [39] J. Liu, W. Zhou, X. Li, J. Xu, and Z. Chen, “Liqua: Lifelong Blind Image Quality Assessment,” *IEEE Transactions on Multimedia*, vol. 2022, 2022.
- [40] D. Ghadiyaram and A. C. Bovik, “Massive online crowd-sourced study of subjective and objective picture quality,” *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372–387, 2016.
- [41] 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.
- [42] A. Ciancio, A. L. N. T. da Costa, E. A. B. da Silva, A. Said, R. Samadani, and P. Obrador, “No-reference blur assessment of digital pictures based on multifeature classifiers,” *IEEE Transactions on Image Processing*, vol. 20, no. 1, pp. 64–75, 2011.
- [43] 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.
- [44] D. M. Chandler 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, Article ID 011006, 2010.
- [45] S. Sun, T. Yu, J. Xu, W. Zhou, and Z. Chen, “Graphiqa: learning distortion graph representations for blind image quality assessment,” *IEEE Transactions on Multimedia*, vol. 1, 2022.
- [46] Video Quality Experts Group, “Final report from the video quality experts group on the validation of objective models of video quality assessment,” in *VQEG meeting*, International Telecommunication Union, Geneva, Switzerland, 2000.
- [47] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, Honolulu, HI, USA, July 2017.
- [48] C. Szegedy, W. Liu, Y. Jia et al., “Going deeper with convolutions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, Boston, MA, USA, June 2015.