Predictive Auto-Encoding Network for Blind Stereoscopic Image Quality Assessment

Jiahua Xu*, Wei Zhou*, Zhibo Chen, Senior Member, IEEE, Suiyi Ling, Student Member, IEEE, and Patrick Le Callet, Fellow, IEEE

Abstract—Stereoscopic image quality assessment (SIQA) has encountered non-trivial challenges due to the fast proliferation of 3D contents. In the past years, deep learning oriented SIQA methods have emerged and achieved spectacular performance compared to conventional algorithms which are only relied on extracting hand-crafted features. However, most existing deep SIQA evaluators are not specifically built for stereoscopic content and consider little prior domain knowledge of the 3D human visual system (HVS) in network design. In this paper, to better simulate the binocular rivalry phenomenon, we propose a Predictive Auto-encoder Network (PAD-Net) for blind/No-Reference stereoscopic image quality assessment (NR-SIQA). The proposed encoder-decoder architecture is inspired by the predictive coding theory that the cognition system tries to match bottom-up visual information with top-down predictions. Besides, we design the Siamese framework to mimic the binocular rivalry in the 3D HVS based on the likelihood and prior maps generated from the predictive coding process. Extensive experimental results on three publicly available stereoscopic image quality databases demonstrate that the devised approach outperforms state-of-the-art algorithms for predicting the perceptual quality of both symmetrically and asymmetrically distorted stereoscopic images with various distortion types.

Index Terms—Siamese encoder-decoder, stereoscopic image quality, 3D human vision, predictive auto-encoding, quality regression.

I. INTRODUCTION

IMAGE quality assessment (IQA) plays an essential role in various image processing and visual communication systems. In the literature, many traditional 2D IQA metrics have been proposed [1]–[7]. Due to the unprecedented development of 3D/stereoscopic technologies, quality issues for 3D media are becoming more and more attracted in both academy and industry [8]. Through the process of acquisition, compression, transmission, display, etc., original reference stereoscopic images usually suffer from perceptual quality degradation caused by diverse distortion types and degrees. As a result, the demands for effectively assessing the perceptual quality of stereoscopic images are urgent.

Compared with the previous 2D IQA case, stereoscopic image quality assessment (SIQA) is more challenging owing to a wide variety of 2D and 3D influential factors, such as image spatial artifacts, depth perception, visual comfort, and so on [9]. These factors have different effects on evaluating the quality of experience (QoE) for stereoscopic images. Except for [10] that considers image distortion and depth perception quality simultaneously, existing research works mainly focus on modeling each individual factor for 3D QoE [11]–[14]. In this paper, we aim to study the visually perceptual quality assessment of stereoscopic images.

Similar to 2D IQA, according to the availability of original reference stereoscopic images, SIQA models are typically divided into three categories: full-reference (FR) [15]–[20], reduced-reference (RR) [21]–[23], and blind/no-reference (NR) [24]–[30] SIQA metrics.

For FR SIQA algorithms, full information of the reference image is assumed to be exploited. The earliest FR IQA model investigated some off-the-shelf 2D IQA metrics, such as structural similarity index (SSIM) [1], universal quality index (UQI) [2], C4 [3] and so on [4], to assess stereoscopic image quality [15]. Further, disparity information was integrated into the 2D IQA metrics to predict the 3D image quality [16], [17]. Apart from incorporating depth clues, the binocular vision characteristics of the human visual system (HVS) were combined with 2D IQA algorithms. For example, Gorley and Holliman proposed a new stereo band limited contrast (SBLC) metric based on the HVS sensitivity to contrast changes in high frequency regions [18]. Moreover, Chen et al. [19] proposed cyclopean images according to the binocular rivalry in the human eyes. In addition, Lin and Wu developed a binocular integration based computational model for evaluating the perceptual quality of stereoscopic images [20].

As for RR SIQA approaches, only part of original non-distorted image data is available. Qi et al. utilized binocular perceptual information (BPI) to perform RR SIQA [21]. By characterizing the statistical properties of stereoscopic images in the reorganized discrete cosine transform (RDCT) domain, Ma et al. presented the RR SIQA method [22]. Furthermore, a new RR-SIQA metric based on natural scene statistics (NSS) and structural degradation was proposed [23].

However, in most practical applications, the original pristine image cannot be accessible. Therefore, NR SIQA is inevitably required. Some research works have been studied about NR SIQA. Akhter et al. extracted the local features of artifacts and disparity to evaluate the perceptual quality of stereoscopic images [24]. Sazzad et al. also exploited perceptual differences of local features for NR SIQA [25]. However, these methods are distortion-specific NR SIQA approaches, which are only
suitable for JPEG coded stereoscopic image pairs. Thus, several general-purpose NR SIQA metrics have emerged. Chen et al. extracted 2D and 3D NSS features from stereoscopic image pairs. The shape-parameter features were regressed onto subjective quality scores by using the well-known support vector regression (SVR). Su et al. proposed the stereoscopic/3D blind image naturalness quality (S3D-BLINQ) index by constructing a convergent cyclopean image and extracting bivariate and correlation NSS features in spatial and wavelet domains.

With the development of deep learning techniques, deep neural networks (DNN) have achieved remarkable advantages for many image processing and computer vision tasks. It also brings improvement for the study on NR SIQA. Specifically, Shao et al. proposed a blind deep quality evaluator for assessing stereoscopic image quality with monocular and binocular interactions based on deep belief network (DBN). The deep NR stereoscopic/3D image quality evaluator (DNR-S3DIQE) was proposed, which extracted local abstractions and then aggregated them into global features by employing the aggregation layer. In addition, Yang et al. took into account the deep perception map and binocular
Binocular rivalry is an important phenomenon in 3D vision for SIQA [19], in which perception alternates between different views when two eyes see different scenes [10]. In conventional perspective, binocular rivalry is simulated by low-level competition between the input stimulus and its energy related to the stimulus [35]–[37]. Recently, more literatures try to explain binocular rivalry by the predictive coding theory [38], [39]. It is a popular theory about how brain process visual stimuli. According to predictive coding theory, the cognition system tries to match bottom-up visual signal with top-down predictions [40]. Different from the traditional statements that binocular rivalry is low-level inter-ocular competition in early visual cortex, binocular rivalry mechanism based on predictive coding theory (BRM-PC) stress both on low-level and high-level competitions [38]. Moreover, the BRM-PC is the HVS guided and more inline with human cognition system [39]. Therefore, we believe that introducing BRM-PC will be beneficial to SIQA.

In this paper, we propose a generic architecture called Predictive Auto-encoDing Network (PAD-Net), which is an end-to-end network for general-purpose NR SIQA. Our contributions of the proposed method are summarized as follows:

- We propose a novel predictive auto-encoding network inspired by the binocular rivalry mechanism based on predictive coding theory [39] which helps to explain the binocular rivalry phenomenon in 3D vision. The source code of PAD-Net is available online for public research usage.
- According to the predictive coding theory in human brain, we adopt the encoder-decoder architecture to reconstruct the sensory input and further exploit the Siamese network to generate the corresponding likelihood as well as prior maps for modeling binocular rivalry in the cognition system.
- We demonstrate that the human brain can perceive the differences of fusion information under various distortion types and levels for symmetrically and asymmetrically distorted stereoscopic image pairs.
- Compared with state-of-the-art SIQA metrics, the proposed PAD-Net provides more precise quality estimation, especially for those stereopairs under asymmetric distortion thanks to the well consideration of binocular rivalry based on predictive coding theory.

The remainder sections of this paper are organized as follows. Section II explains the proposed Predictive Auto-encoDing Network (PAD-Net) in details. Section III presents the experimental results and analysis. In Section IV, we conclude the paper with an outlook on the future work.

II. PROPOSED METHOD

Inspired by the binocular rivalry mechanism based on predictive coding theory (BRM-PC) [39], we design the no-reference PAD-Net to automatically predict the perceptual quality of stereoscopic images. The proposed PAD-Net is a Siamese based end-to-end network, including auto predictive coding and quality regression modules. Given paired distorted 3D images, they are divided into sub-images with size 256x256 first, then the quality scores of these sub-images are estimated through PAD-Net and aggregated as a final one as done in [41]. We pre-train the proposed auto-encoder and quality regression module on Waterloo Exploration Database [42] and LIVE 2D Database [43], respectively. After that, the entire network is jointly optimized using 3D IQA databases [19], [44] to generate more accurate predictions.

In this section, the architecture of our proposed PAD-Net is described first as shown in Fig. 1. Then, we introduce the Siamese encoder-decoder module and quality regression module in details. Finally, the training and testing methods of our PAD-Net are presented.

A. Overview

The architecture of the PAD-Net is depicted in Fig. 1 which contains a Siamese encoder-decoder module and a quality regression module. The Siamese encoder-decoder module represents the processing procedure of left and right view images. It is inspired by the BRM-PC that the human brain tries match bottom-up visual stimuli with top-down predictions [40] and the view with the larger posterior probability will be dominant during the rivalry and inhibit the other one. In [39], an empirical Bayesian framework based on BRM-PC was proposed and the posterior probability is calculated from the likelihood and prior probability [45]. In our proposed PAD-Net, the likelihood and prior are inferred through the Siamese encoder-decoder, which corresponds to the likelihood and prior in the BRM-PC.

For each branch of the Siamese encoder-decoder, four times down convolution and up convolution with a stride of 2 pixels are performed to reconstruct the distorted images. Non-linear activation function layers follow the first three down convolution and up convolution layers to enhance the representation ability of our neural network. We square the difference between the input and reconstructed image as the residual error to obtain likelihood probability map and the high-level representation after four times down convolution to generate prior probability map. The reasons will be given in Section II B.

Once we get the left and right view images as well as their likelihood and prior probability maps, we can put them into
the quality regression module and compute the final quality score. It can be seen from the bottom of Fig. 1 that the quality regression module is composed of 1) fusion for input images and probability maps, 2) ResNet-18 without last two layers as feature extractor and 3) the final max pooling and fully connected layer. Note that fusion in Fig. 1 includes a one stride convolution layer and activation function layer to match the input size with ResNet-18. The output of the proposed PAD-Net is the predicted quality score of the input stereoscopic image.

B. Siamese Encoder-Decoder Module

The Siamese encoder-decoder module is inspired by the BRM-PC [39]. According to [40], the core task of brain is to represent the environmental causes of its sensory input. Generative model based hierarchical Bayesian inference is used to provide constraints on the mapping between the input and causes (hypotheses). Different from the conventional understanding that exists in lower-level inter-ocular visual cortex, binocular rivalry mechanism based on predictive coding theory stresses both on lower-level and high-level competition [38]. Therefore, we believe the BRM-PC is the HVS guided and able to generate more reliable and interpretable results, which are verified in Section III.

![Fig. 2: Predictive coding theory based hierarchical representation of the sensory input [42].](image)

In this model, the feedback pathways carry the prediction from lower level. In addition, the feedforward pathways return the prediction errors between the prediction and sensory input to update the prediction and get the best hypothesis.

To learn from the detailed predictive coding theory in the visual cortex, the principle of human brain cognition is illustrated in Fig. 2. In the hierarchical predictive coding model, the feedback pathways carry the prediction from lower level and the feedforward pathways return the prediction errors between the prediction and sensory input to update the prediction and get the best hypothesis. To simulate this process, we adopt the encoder-decoder network structure for image compression from [48], [49], which is a hierarchical structure. In Fig. 2 prediction is adjusted with residual errors to obtain better hypothesis, thus, we employ the squared error between predicted (decoded) image and input image [47] as our loss function $l_1$ to pre-train the encoder-decoder network as follows:

$$l_1 = \frac{1}{M \times N \times C \times K} \sum_{x,y,c,k} (I^{(k)}(x, y, c) - \hat{I}^{(k)}(x, y, c))^2,$$

(1)

$$\hat{I}^{(k)} = f_1(I^{(k)}; w_1),$$

(2)

$$w'_1 = \arg\min_{w_1} l_1,$$

(3)

where $I^{(k)}$ and $\hat{I}^{(k)}$ represent the k-th input and predicted image. $M$, $N$ and $C$ are the width, height and channel of $I^{(k)}$ and $\hat{I}^{(k)}$. $K$ is the batch size of a mini-batch training data. $\hat{I}^{(k)}$ is estimated via the encoder-decoder network $f_1$ with weight $w'_1$. During the training stage, the loss which can be considered as prediction error will be backpropagated through the network. Then, we use the gradient descent algorithm to update $w_1$ and generate better prediction. Finally, the prediction error will converge and reach a stable value. The decoded image would not change greatly. Thus, the updating policy of the encoder-decoder network is similar to the prediction coding theory of human brain. Note that the encoder-decoder network is pre-trained on the Waterloo Exploration Database [42], therefore the training data includes both reference and distorted images.

**TABLE I: DETAILED CONFIGURATIONS OF THE ENCODER-DECODER NETWORK.**

| Layer | Input Shape | Output Shape | Kernel size/Stride/Padding |
|-------|-------------|--------------|---------------------------|
| Encoder | Conv1+GDN1 | 3, 256, 256 | 128, 128, 128 | 5x5/2/2 |
| | Conv2+GDN2 | 128, 128, 128 | 128, 64, 64 | 5x5/2/2 |
| | Conv3+GDN3 | 128, 64, 64 | 128, 32, 32 | 5x5/2/2 |
| | Conv4 | 128, 32, 32 | 192, 16, 16 | 5x5/2/2 |
| Decoder | Unconv1+IGDN1 | 192, 16, 16 | 128, 32, 32 | 5x5/2/2 |
| | Unconv2+IGDN2 | 128, 32, 32 | 128, 64, 64 | 5x5/2/2 |
| | Unconv3+IGDN3 | 128, 64, 64 | 128, 128, 128 | 5x5/2/2 |
| | Unconv4 | 128, 128, 128 | 3, 256, 256 | 5x5/2/2 |

1 Conv: Convolution  
2 Unconv: Fractionally-strided convolution  
3 GDN: Generalized divisive normalization  
4 IGDN: Inverse generalized divisive normalization

The structure of the encoder-decoder network is list in Table I. It is composed of convolution and fractionally-strided convolution layers [50], generalized divisive normalization (GDN) transform and inverse generalized divisive normalization (IGDN). GDN is inspired by the neuron models in biological visual system and proves to be effective in density estimation [51], image compression [49] and image quality assessment [41]. The GDN and IGDN operations are given by:

$$y_i(m, n) = \frac{x_i(m, n)}{(\beta_i + \sum_{j} \gamma_{ij} x_j(m, n))^2},$$

(4)

$$\hat{x}_i(m, n) = \tilde{y}_i(m, n) \cdot (\beta_i + \sum_{j} \tilde{\gamma}_{ij} x_j(m, n))^2,$$

(5)

where $x_i$ means the i-th channel of $x$ and it is the input of GDN transform. $y_i$ is the i-th channel of normalized activation feature map and it is the output of GDN operation. Moreover, $\beta$ and $\gamma$ are the parameters to be updated in GDN function.
 Likewise, $\hat{x}_i$, $\hat{y}_i$, $\hat{\beta}$ and $\hat{\gamma}$ share the same meaning as $x_i$, $y_i$, $\beta$ and $\gamma$ for IGDN transform. The goal of our encoder-decoder module is to reconstruct the input sensory instead of compressing it. Therefore, we remove the quantization step in [39] for better reconstruction.

![Simplified schematic of the binocular rivalry mechanism based on predictive coding theory][1]

The black solid arrow is the top-down prediction from high level and the blue dotted arrow is the bottom-up error signals. The error signals correspond to the likelihood and the high-level representation correspond to the prior. Based on the likelihood and prior, the view with higher posterior probability will win during the rivalry and inhibit to each other.

Up till now, we get the high-level representation and reconstructed image from the input sensory. Then, according to the BRM-PC [39], the view with higher posterior probability will win during the rivalry and inhibit to each other as shown in Fig. 3. From the Bayesian perspective, the posterior probability is related to prior and likelihood. In other words, the likelihood is about how well the hypothesis predicts the input and the prior is about how probable the hypothesis is and concern with empirical knowledge [52]. To calculate likelihood, we first obtain the $k$-th squared residual error map $E^{(k)}$ in the mini-batch of a size of $K$ as follows:

$$E^{(k)} = \frac{1}{C} \sum_c (I_c^{(k)} - \hat{I}_c^{(k)})^2,$$

where $I_c^{(k)}$ and $\hat{I}_c^{(k)}$ denote the $c$-th channel of the input and predicted image. $C$ equals 3 in Eq. [6]. The likelihood is used to measure the similarity between $I_c^{(k)}$ and $\hat{I}_c^{(k)}$ and it is inversely proportional to errors. The training stage of encoder-decoder network can be regarded as the procedure of prediction error minimization.

In the Siamese encoder-decoder module, the prior is modeled with the high-level representation of the sensory input since the prior comes from high levels in the cognition system, as assumed in empirical Bayes [39], [53]. Thus, the high-level features are utilized to generate the $k$-th prior map $P^{(k)}$. Before being fed into the quality regression module, the squared error map $E^{(k)}$ and the prior map $P^{(k)}$ are normalized between left and right view as follows:

$$P^{(k)}_{nl} = \frac{P^{(k)}_l}{P^{(k)}_l + P^{(k)}_r} \quad \text{and} \quad P^{(k)}_{nr} = \frac{P^{(k)}_r}{P^{(k)}_l + P^{(k)}_r},$$

$$L^{(k)}_{nl} = \frac{E^{(k)}_l}{E^{(k)}_l + E^{(k)}_r} \quad \text{and} \quad L^{(k)}_{nr} = \frac{E^{(k)}_r}{E^{(k)}_l + E^{(k)}_r},$$

where $P^{(k)}_l$, $P^{(k)}_r$, $E^{(k)}_l$ and $E^{(k)}_r$ are the prior and error maps for the $k$-th left view and right view images in a mini-batch. $P^{(k)}_{nl}$, $P^{(k)}_{nr}$, $L^{(k)}_{nl}$ and $L^{(k)}_{nr}$ indicate the normalized prior and likelihood probability maps. Note that the error is opposite to the likelihood, that is to say, if the error is large, the likelihood will be small. For example, when computing the likelihood map for left view, the error map of right view is adopted and vice versa.

![Detailed configurations of the prior and likelihood generation network.][2]

The detailed structure of prior and likelihood creation is given in Table II. We employ the Softplus activation function [54] in prior generation to avoid square negative values to positive values. It is defined as:

$$s(x) = \log(1 + e^x).$$

The Softplus function can be regarded as the smoothing version of ReLU function which is similar to the way cerebral neurons being activated [55].

### C. Quality Regression Module

Based on the distorted stereoscopic images as well as the obtained prior and likelihood probability maps from the Siamese encoder-decoder module, we fuse them as a 3-channel feature map and further feed the 3-channel feature map into the ResNet-18 quality regression network to extract discriminative features for quality estimation. ResNet-18 is chosen for its excellent ability of feature extraction [33], [56]. The last two layers including average pooling and fully connected layer are removed for regressing the feature map into a quality score. Table III illustrates the architecture of quality regression network. The input stem and basic block of the ResNet-18 structure are shown in Fig. 1.
TABLE III: DETAILED CONFIGURATIONS OF THE QUALITY REGRESSION NETWORK.

| Layer                      | Input Shape | Output Shape | Kernel size /Stride/Padding |
|----------------------------|-------------|--------------|-----------------------------|
| Fusion                     | Conv/GDN9   | 10, 256, 256 | 3, 256, 256, 1x1/1/0        |
| ResNet18 Regression        | ResNet18    | 3, 256, 256  | 512, 8,8, 1x1/8/0           |
| Maxpool10                  | Maxpool10   | 512, 8,8     | 512,1,1, 1x1/0              |
| FC1                        | FC1         | 512          | 1                           |

1 Conv: Convolution
2 GDN: Generalized divisive normalization
3 FC: Fully connected layer

D. Training and Testing

Owing to the limited size of available 3D image quality assessment database, we train the PAD-Net on the sub-image pairs with the resolution of $256 \times 256$, the MOS value for the entire image is assumed as the quality scores for several sub-images as done in [41]. Thus, sub-images coming from the same test image share the same labels. Moreover, transfer learning is adopted to solve the problem of lacking labeled data and enhance the prediction accuracy of the network [57], [58].

In blind stereoscopic image quality assessment, it is difficult to predict the MOS value precisely [41]. Therefore, we divide the training stage into three steps: 1) pre-training of encoder-decoder on the pristine and distorted 2D images; 2) pre-training ResNet-18 regression on the pristine and distorted 2D images; 3) joint optimization on the 3D IQA database.

Firstly, encoder-decoder is trained to minimize the difference between predicted and input images which is described as:

\[ w_1 = \arg \min_{w_1} l_1. \]  

Secondly, we utilize the original and distorted 2D images along with the associated MOS scores to pre-train the ResNet-18 regression network. It is aimed to map the 2D image into a quality score. In addition, ResNet-18 with pre-trained weight on ImageNet is adopted for better initialization. Then, the loss function $l_2$ for second step pre-training is defined as:

\[ l_2 = \frac{1}{K} \sum_k (q_{2d}^{(k)} - \hat{q}_{2d}^{(k)})^2, \]  

\[ \hat{q}_{2d}^{(k)} = f_2(I^{(k)}; w_2), \]  

where $q_{2d}$ and $\hat{q}_{2d}$ indicate the real MOS and predicted score for the $k$-th input 2D sub-image $I^{(k)}$ in a mini-batch. The weight $w_2$ for the ResNet-18 regression network $f_2$ is updated by minimizing $l_2$ as follows:

\[ w_2' = \arg \min_{w_2} l_2. \]  

Finally, the Siamese encoder-decoder and quality regression module are jointly optimized using stereo image pairs. Since the ultimate purpose of PAD-Net is to estimate the perceptual quality of 3D images, we again adopt the $l_2$-norm between the subjective MOS value $q_{3d}$ and predicted score $\hat{q}_{3d}$ as loss function:

\[ l_3 = \frac{1}{K} \sum_k (q_{3d}^{(k)} - \hat{q}_{3d}^{(k)})^2, \]  

where $I_{3d}^{(k)}$ and $\hat{I}_{3d}^{(k)}$ represent the input 3D sub-image pairs. $f$ indicates the PAD-Net with encoder-decoder weight $w_1$, ResNet-18 regression weight $w_2$ and weight $w_3$ training from scratch. At the joint optimization step, $(w_1, w_2)$ are initialized with pre-trained weight $(w_1, w_2)$ and updated with $w_3$ through final loss minimization:

\[ w_1', w_2', w_3' = \arg \min_{w_1, w_2, w_3} l_3. \]  

In the testing stage, the stereo image is divided into sub-image pairs with a stride of $U$ to cover the whole content. The predicted qualities of all sub-image pairs are averaged to compute the final perceptual quality score.

III. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we first introduce the databases and performance measures used in our experiment. Then, the experimental results of the proposed PAD-Net on the entire LIVE databases and individual distortion type are illustrated. Meanwhile, the visualization results are provided for better explanation. Finally, we conduct the ablation study to verify the effectiveness of each component in our model and measure the computation complexity.

A. Databases and Performance Measures

Three benchmark stereoscopic image quality assessment databases are used in our experiment including LIVE Phase I [44], LIVE Phase II [19, 26] and Waterloo IVC Phase I [59].

LIVE Phase I [44]: This database contains 20 original and 365 symmetrically distorted stereo image pairs. Five distortion types are included in this database, namely JPEG2000 compression (JP2K), JPEG compression (JPEG), additive white noise (WN), Gaussian blur (BLUR) and Raleigh fast fading channel distortion (FF). Subjective differential mean opinion score (DMOS) is provided for each degraded stereo image. Higher DMOS value means lower visual quality.

LIVE Phase II [19, 26]: It includes 120 symmetrically and 240 asymmetrically distorted stereopairs derived from 8 reference images. This database contains the same distortion types as LIVE Phase I. For each distortion type, the pristine image pair is degraded to 3 symmetrically and 6 asymmetrically image pairs. Subjective scores are also recorded in the form of DMOS.

Waterloo IVC Phase I [59]: This database originates from 6 pristine stereoscopic image pairs. The reference image is altered by three types of distortions, namely WN, BLUR, and JPEG. Altogether, there are totally 78 symmetrically and 252 asymmetrically distorted stereopairs. Subjective mean opinion score (MOS) and individual scores are provided for each stereoscopic image in this database, while higher MOS value means better visual quality.

Performance Measure: Three commonly used criteria [60] are utilized in our experiment for performance evaluation, including Spearman’s rank order correlation coefficient (SROCC), Pearson’s linear correlation coefficient (PLCC) and

\[ r_{SROCC} = \frac{\sum_{i=1}^{N} (S_i - \bar{S})(O_i - \bar{O})}{\sqrt{\sum_{i=1}^{N} (S_i - \bar{S})^2} \sqrt{\sum_{i=1}^{N} (O_i - \bar{O})^2}}, \]  

\[ r_{PLCC} = \frac{\sum_{i=1}^{N} (S_i - \bar{S})(O_i - \bar{O})}{\sqrt{\sum_{i=1}^{N} (S_i - \bar{S})^2} \sqrt{\sum_{i=1}^{N} (O_i - \bar{O})^2}}, \]

\[ r_{MOS} = \frac{\sum_{i=1}^{N} (S_i - \bar{S})(O_i - \bar{O})}{\sqrt{\sum_{i=1}^{N} (S_i - \bar{S})^2} \sqrt{\sum_{i=1}^{N} (O_i - \bar{O})^2}}, \]
root mean squared error (RMSE). SROCC is a non-parametric measure and independent of monotonic mapping. PLCC and RMSE evaluate the prediction accuracy. Higher SROCC, PLCC and lower RMSE indicate better correlation with human judgements. Before calculating PLCC and RMSE, a five-parameter logistic function \[ f(x) = \frac{1}{1 + \exp(-ax + b)} \] is applied to maximize the correlation between subjective ratings and objective metrics as follows:

\[ y = \beta_1 \left( \frac{1}{2} - \frac{1}{1 + \exp(\beta_2(x - \beta_3))} \right) + \beta_4 x + \beta_5, \]

where \( x \) indicates the predicted score of objective metrics and \( y \) represents the mapped output. \( \beta_1 \) to \( \beta_5 \) are the five parameters to be fitted.

One of the main issues of PLCC and SROCC is that they neglect the uncertainty of the subjective scores \[61\]. Thus, we also employ the Krasula methodology \[61\], which could be used to better assess the capabilities of objective metrics by considering the statistical significance of the subjective scores and getting rid of the mapping functions.

The basic idea of this model is to determine the reliability of objective models by checking whether they are capable of well 1) distinguishing the significantly different stimuli from the similar ones, and 2) indicating whether one stimulus are of better/worse quality than the other. To this end, in the Krasula framework, pairs of stimuli are selected from the database to compute the area under ROC curve of the Different vs. Similar (AUC-DS), area under ROC curve of the Better vs. Worse category (AUC-BW), and percentage of correct classification (CC). Higher AUC-DS and AUC-BW mean more capability to indicate different/similar and better/worse pairs. Higher CC represents better prediction accuracy. Please refer to \[61\] for more details.

B. Performance Evaluation

In the experiment, the distorted stereo pairs are randomly split into 80% training set and 20% testing set according to \[30\]. We adopt the Adam algorithm in the pre-training and joint optimization step. During pre-training, the learning rate \( \alpha \) is set as \(10^{-3}\) and lowered by a factor of 10 every 50 epochs. The pre-trained weights are obtained after 100 epochs. Since encoder-decoder should retain its function, the learning rate \( \alpha_1 \) for encoder-decoder weight \( w_1 \) is set as \(10^{-5}\) to avoid drastic change when conducting joint optimization. Moreover, the learning rate \( \alpha_2 \) for ResNet-18 regression \( w_2 \) is set as half of \( \alpha_3 \) for \( w_3 \). \( \alpha_3 \) is initialized as \(10^{-3}\) and scaled by 0.25 every 50 epochs. The learning rate remains unchanged after 200 epochs. We apply data augmentation by randomly cropping, horizontal and vertical flipping in the training stage \[62\]. The results are obtained after 300 epochs. During testing, the stride \( U \) is set as 192 for width and 104 for height in a slight overlapping manner to cover the whole resolution in LIVE databases as shown in Fig. 4.

We compare the proposed PAD-Net with several classic FR, RR and NR SIQA metrics on the LIVE Phase I and II database. The competing FR and RR models include Gorley’s method \[18\], You’s method \[17\], Benoit’s method \[16\], Lin’s method \[20\], Cyclopean MS-SSIM \[19\], RR-BPI \[21\], RR-RDCT \[22\] and Ma’s method \[23\]. For NR metrics, some hand-crafted features based algorithms including Akhter’s method \[24\], Sazzad’s method \[25\], Chen’s method \[26\], S3D-BLINQ \[27\], DECOISINE \[34\] and deep neural network based models including Shao’s method \[11\], CNN \[63\], DNR-S3DIQE \[28\], DBN \[29\], StereoQA-Net \[30\] are considered in the performance comparison. Note that CNN \[63\] is computed for left and right view images separately and then average the scores for both views. The SROCC, PLCC and RMSE performance for the above metrics and proposed PAD-Net are listed in Table IV where the best results are highlighted in bold. It could be observed from the table that the proposed method outperforms state-of-the-art SIQA metrics, especially on LIVE Phase II database. Since there are more asymmetrically distorted images in LIVE II, the proposed PAD-Net is more effective for the challenging asymmetric distortion which will be explained in Section III C.

**TABLE IV: OVERALL PERFORMANCE COMPARISON ON LIVE PHASE I AND PHASE II DATABASES. THE BEST PERFORMING RESULTS ARE HIGHLIGHTED IN BOLD.**

| Metrics | SROCC | PLCC | RMSE | SROCC | PLCC | RMSE |
|---------|-------|------|------|-------|------|------|
| Gorley  | 0.142 | 0.451| 14.635| 0.146 | 0.515| 9.675 |
| You     | 0.878 | 0.881| 7.746| 0.786 | 0.800| 6.772 |
| Benoit  | 0.899 | 0.902| 7.061| 0.728 | 0.748| 7.490 |
| Lin     | 0.856 | 0.784| -    | 0.638 | 0.642| -    |
| Cyclopean MS-SSIM | 0.916 | 0.917| 6.533| 0.889 | 0.900| 4.987 |

**TABLE V: SROCC, PLCC, RMSE AND KRASULA PERFORMANCE EVALUATION ON THE WATERLOO IVC PHASE I DATABASE.**

| Metrics | SROCC | PLCC | RMSE |
|---------|-------|------|------|
| StereoQA-Net | 0.955 | 0.970 | 4.350 |
| Proposed PAD-Net | 0.974 | 0.979 | 3.696 |

| Metrics | AUC_DS | AUC_BW | CC |
|---------|--------|--------|---|
| StereoQA-Net | 0.894 | 0.998 | 0.979 |
| Proposed PAD-Net | 0.925 | 0.999 | 0.993 |
To employ the Krasula methodology [61], significance analysis of the subjective scores are required. Among the three considered 3D IQA databases [19, 26, 44, 59], only the Waterloo IVC Phase I [59] is equipped with individual scores. Furthermore, the excusable/source code of most of the state-of-the-art NR 3D metrics are not released. Therefore, we could only conduct the Krasula analysis on the Waterloo IVC Phase I dataset and compare the proposed PAD-Net with StereoQA-Net [30], which obtains the best performance on both LIVE Phase I [44] and LIVE Phase II [19, 26] databases. Table V lists the results of SROCC, PLCC, RMSE and Krasula performance evaluation, it can be observed from the table that the proposed PAD-Net achieve the best performance in terms of SROCC, PLCC, RMSE, AUC-DS, AUC-BW and CC. The results of Krasula performance criteria demonstrate that PAD-Net is the most promising metric in distinguishing stereo images with different qualities.

**TABLE VI: RESULTS OF THE T-TEST ON THE LIVE PHASE I DATABASE.**

|                | Cyclopean MS-SSIM [19] | CNN [63] | StereoQA-Net [30] | Proposed PAD-Net |
|----------------|------------------------|----------|-------------------|------------------|
| Cyclopean MS-SSIM [19] | 0                      | -1       | -1                | -1               |
| CNN [63]       | 1                      | 0        | -1                | -1               |
| StereoQA-Net [30] | 1                      | 1        | 0                 | -1               |
| Proposed PAD-Net | 1                      | 1        | 1                 | 0                |

**TABLE VII: RESULTS OF THE T-TEST ON THE LIVE PHASE II DATABASE.**

|                | Cyclopean MS-SSIM [19] | CNN [63] | StereoQA-Net [30] | Proposed PAD-Net |
|----------------|------------------------|----------|-------------------|------------------|
| Cyclopean MS-SSIM [19] | 0                      | 1        | -1                | -1               |
| CNN [63]       | -1                     | 0        | -1                | -1               |
| StereoQA-Net [30] | 1                      | 1        | 0                 | -1               |
| Proposed PAD-Net | 1                      | 1        | 1                 | 0                |

Moreover, we conduct significance t-tests using the PLCC values of 100 runs [28] to verify whether our proposed model is statistically better than other metrics. Table VI and VII list the results of t-tests on LIVE Phase I and II where 1 or -1 indicate that the metric in the row is statistically superior or worse than the competitive metric in the column. The number 0 means that the two metrics are statistically indistinguishable. From Table VI and VII, we can see that our proposed metric is statistically better than other metrics both on LIVE Phase I and II.

**C. Performance Evaluation for Symmetric/Asymmetric Distortion**

Our proposed PAD-Net is based on the predictive coding theory and applies deep neural networks to model binocular rivalry mechanism for better prediction of the stereo image quality. Binocular rivalry seldom happens in symmetrical distortion but plays an important role in asymmetrically distorted image quality assessment. Table VIII presents the SROCC performance for symmetrically and asymmetrically distorted images in LIVE Phase II and Waterloo IVC Phase I databases. PAD-Net demonstrates the extraordinary ability to predict the perceived quality of asymmetrically distorted stereo pairs by well mimicking the visual mechanism in binocular rivalry.

---

**Fig. 5: Visualization results for different kinds of distortion in LIVE II database.** The higher the value (from blue to yellow) is, the larger the probability. Images inside yellow boxes are fed into the quality regression network.
TABLE VIII: SROCC PERFORMANCE FOR SYMMETRICALLY AND ASYMMETRICALLY DISTORTED STEREOSCOPIC IMAGE PAIRS ON LIVE PHASE II AND WATERLOO IVC PHASE I DATABASES. THE BEST PERFORMING RESULTS ARE HIGHLIGHTED IN BOLD.

| Type | Metrics | LIVE Phase II | Waterdo IVC Phase I |
|------|---------|---------------|---------------------|
|      |         | Symmetric     | Asymmetric          | Symmetric | Asymmetric |
| FR   | Godde  | 0.383         | 0.856              | 0.566     | 0.475     |
|      | You    | 0.914         | 0.701              | 0.752     | 0.571     |
|      | Bensen | 0.860         | 0.671              | 0.728     | 0.577     |
|      | Lin    | 0.605         | 0.668              | 0.688     | 0.592     |
|      | Cycloped MS-SSIM | 0.923 | 0.842 | 0.924 | 0.643 |
| NR   | Akhter | 0.420         | 0.517              | 0.934     | 0.907     |
|      | Chen   | 0.918         | 0.834              | 0.934     | -         |
|      | CNN    | 0.590         | 0.633              | -         | -         |
|      | SID-BLING | 0.937 | 0.849 | - | - |
|      | StereoQA-Net | 0.979 | 0.927 | 0.957 | 0.940 |
|      | Proposed PAD-Net | 0.982 | 0.954 | 0.985 | 0.978 |

We provide some visualization results of the PAD-Net for better explanation. In Fig. 5(a), the left view is undamaged while the right view is degraded with white noise. The reconstructed stereo pairs are similar to the input ones. In the normalized prior and likelihood map, lighter color indicates higher probability. For asymmetrically noised image, the view with heavier noise is harder to be reconstructed thus having lower likelihood probability globally. However, at the strong edges that indicate structural information, the undamaged view may be allocated smaller probability as shown in the magnified likelihood map of Fig. 5(a). Moreover, the discrepancy between normalized left and right view prior maps is not very large since we can still recognize the scene for both views. According to [59], 3D image quality is more affected by the poor quality view for noise contamination, as reflected by the magnified likelihood map of Fig. 5(a). This tendency appears around regions containing obvious edges which means subjective rating is more affected by structural information.

Fig. 5(b) presents the asymmetrically blurred image, the left view is undamaged while the right view is blurred. Contrary to white noise, the view with heavier blurring effect is easier to reconstruct thus having larger likelihood probability. In addition, strong edges in the undistorted view tend to gain more probability in binocular rivalry as indicated in the magnified likelihood map of Fig. 5(b). For the same reason that both views are comprehensible, the normalized prior maps do not show huge differences. For image blur, it is reported in [59], [64] that 3D image quality is more affected by the high quality view. In the magnified likelihood map of Fig. 5(b), strong edges in high quality view have higher probability, which again demonstrates that subjective judgments are more affected by structural information. In Fig. 5(c), the stereo pairs are symmetrically distorted. As a result, the normalized prior and likelihood maps for both views share similar probability.

Besides, the fusion maps of distorted images, normalized prior and likelihood from left and right views are depicted in Fig. 6. The colors of fusion map for symmetrically and asymmetrically distorted images are easy to be distinguished. The color of the fusion maps for symmetrically distorted image is gray tone, while asymmetrically distorted images appear green or pink. It is caused by the different probability for left and right view image during fusion. Moreover, for asymmetrically distorted images, white noise is different from the other four distortion types since noise tends to introduce high frequency information while the other four distortion types are apt to remove details that correspond to high frequency information.

D. Performance Evaluation on Individual Distortion Type

We further investigate the capacity of our proposed PAD-Net for each distortion type, the SROCC and PLCC performance are illustrated in Table IX and X. The best performing results across listed metrics are highlighted in boldface. As shown in Table IX and X our proposed model achieves competitive performance for most of the distortion types. In addition, the scatter plots of DMOS values versus objective scores predicted by PAD-Net for each distortion type on LIVE Phase I and II are presented in Fig. 7(a) and 7(b). The linear correlation between DMOS values and predicted scores demonstrates the great monotonicity and accuracy of PAD-Net. DMOS value range of JPEG compressed images is roughly narrower than those of other distortion types, making it more difficult to estimate the perceptual image quality. Thus, the PLCC and SROCC performance for JPEG distortion is generally lower than the other four.

E. Cross Database Tests

We conduct cross database tests to verify the generalization ability of our proposed PAD-Net. Models to be compared are trained on one database and tested on another. Table XI presents the PLCC performance for cross database validation. Although PAD-Net does not show the best performance when trained on LIVE Phase II and tested on LIVE Phase I, it outperforms other metrics in the second round which is a
more challenging task. Since LIVE Phase I only consists of symmetrical distortion while more than half of the 3D pairs in LIVE Phase II are asymmetrically distorted. PAD-Net trained on LIVE Phase I is able to handle the asymmetrical distortion in LIVE Phase II never met before. The PLCC performance on LIVE Phase II not only proves the generalization and robustness of PAD-Net but also demonstrates the effectiveness of the binocular rivalry mechanism based on predictive coding theory for asymmetric distortion in the proposed method.

### F. Effects of Network Structure

To explore the influence of different network structures as quality regression network, VGG-16 [32], ResNet-18, 34 and 50 [55] are adopted to make comparisons. The SROCC, PLCC and RMSE performance on LIVE Phase I and II are reported in Table XI. Firstly, ResNet has superior capability to extract discriminative features for quality prediction than VGG structure. Moreover, with the increased depth of ResNet, the performance does not improve. The possible explanation is that the limited training data requires shallow architecture. Generally, very deep networks need a huge amount of training data.
data to achieve high performance. However, there are only hundreds of distorted images in LIVE Phase I and II, even with data augmentation, it is far from enough for deeper networks. Lack of training data may cause over-fitting problems for deeper neural networks. As a result, ResNet-18 is chosen in this paper to reach better tradeoff.

TABLE XII: PERFORMANCE EVALUATION OF DIFFERENT STRUCTURE AS QUALITY REGRESSION NETWORK. THE BEST PERFORMING RESULTS ARE HIGHLIGHTED IN BOLD.

| Regression Structure | LIVE Phase I | LIVE Phase II |
|----------------------|--------------|---------------|
|                      | SROCC | PLCC | RMSE | SROCC | PLCC | RMSE |
| VGG-16               | 0.913 | 0.924 | 5.945 | 0.858 | 0.869 | 5.506 |
| ResNet-18            | 0.973 | 0.975 | 3.514 | 0.967 | 0.975 | 2.446 |
| ResNet-34            | 0.970 | 0.974 | 3.537 | 0.960 | 0.961 | 2.871 |
| ResNet-50            | 0.968 | 0.974 | 3.588 | 0.963 | 0.965 | 2.727 |

G. Ablation Study

Furthermore, ablation study is conducted to verify the effectiveness of each component in PAD-Net. We first feed the distorted left and right view images into quality regression module as the baseline. Then, normalized likelihood and prior maps are introduced to provide additive information for computing rivalry dominance of both views. Moreover, we compare different fusion methods.

![Fig. 7: Scatter plots of DMOS values against predictions by PAD-Net for individual distortion type on (a) LIVE Phase I and (b) LIVE Phase II.](image)

![Fig. 8: Ablation Study on LIVE Phase I and Phase II Databases.](image)
time for each 3D image. The results in Table XIII show that PAD-Net only needs around 0.906 seconds per image which is significantly lower than other metrics.

TABLE XIII: THE COMPUTATION TIME ON NVIDIA GTX 1080ti GPU. THE BEST PERFORMING RESULTS ARE HIGHLIGHTED IN BOLD.

| Metrics           | CNN [63] | StereoQA-Net [30] | Proposed PAD-Net |
|-------------------|----------|-------------------|------------------|
| Time (sec)        | 8.308    | 2.377             | 0.906            |

IV. CONCLUSIONS

In this paper, we explore a novel deep learning approach for blind stereoscopic image quality assessment according to the binocular rivalry mechanism based on predictive coding theory. Our proposed predictive auto-encoding network is an end-to-end architecture inspired by the human brain cognition process. Specifically, we adopt the Siamese encoder-decoder module to reconstruct binocular counterparts and generate the corresponding likelihood as well as prior maps. Moreover, we incorporate the quality regression module to obtain the final estimated perceptual quality score. The experimental results demonstrate that our proposed PAD-Net correlates well with subjective ratings. In addition, the proposed method outperforms state-of-the-art algorithms for distorted stereoscopic images under a variety of distortion types, especially for those with asymmetric distortions. Furthermore, we also show that the proposed PAD-Net has a promising generalization ability and can achieve lower time complexity. In future work, we intend to extend the method to blind stereoscopic video quality assessment. Except for image visual quality, we plan to investigate other 3D quality dimensions such as depth perception and visual comfort.

REFERENCES

[1] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli et al., “Image quality assessment: from error visibility to structural similarity,” IEEE transactions on image processing, vol. 13, no. 4, pp. 600–612, 2004.
[2] Z. Wang and A. C. Bovik, “A universal image quality index,” IEEE signal processing letters, vol. 9, no. 3, pp. 81–84, 2002.
[3] M. Carne, P. Le Callet, and D. Barba, “An image quality assessment method based on perception of structural information,” in Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429), vol. 3. IEEE, 2003, pp. III–185.
[4] Z. Wang and E. P. Simoncelli, “Reduced-reference image quality assessment using a wavelet-domain natural image statistic model,” in Human Vision and Electronic Imaging X, vol. 5666. International Society for Optics and Photonics, 2005, pp. 149–159.
[5] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multiscale structural similarity for image quality assessment,” in The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, vol. 2. Ieee, 2003, pp. 1398–1402.
[6] L. Zhang, L. Zhang, X. Mou, and D. Zhang, “Fsim: A feature similarity index for image quality assessment,” IEEE transactions on Image Processing, vol. 20, no. 8, pp. 2378–2386, 2011.
[7] H. Talebi and P. Milanfar, “Nima: Neural image assessment,” IEEE Transactions on Image Processing, vol. 27, no. 8, pp. 3998–4011, 2018.
[8] C.-C. Su, A. K. Moorthy, and A. C. Bovik, “Visual quality assessment of stereoscopic image and video: challenges, advances, and future trends,” in Visual Signal Quality Assessment. Springer, 2015, pp. 185–212.
[9] B. Series, “Subjective methods for the assessment of stereoscopic 3d tv systems,” 2012.
[10] Z. Chen, W. Zhou, and W. Li, “Blind stereoscopic video quality assessment: From depth perception to overall experience,” IEEE Transactions on Image Processing, vol. 27, no. 2, pp. 721–734, 2017.
[11] F. Shao, W. Tian, W. Lin, G. Jiang, and Q. Dai, “Toward a blind deep quality evaluator for stereoscopic images based on monocular and binocular interactions,” IEEE Transactions on Image Processing, vol. 25, no. 5, pp. 2059–2074, 2016.
[12] F. Shao, Y. Gao, Q. Jiang, G. Jiang, and Y.-S. Ho, “Multistage pooling for blind quality prediction of asymmetric multiply-distorted stereoscopic images,” IEEE Transactions on Multimedia, vol. 20, no. 10, pp. 2605–2619, 2018.
[13] J. Wang, S. Wang, K. Ma, and Z. Wang, “Perceptual depth quality in distorted stereoscopic images,” IEEE Transactions on Image Processing, vol. 26, no. 3, pp. 1200–1215, 2016.
[14] J. Chen, J. Zhou, J. Sun, and A. C. Bovik, “Visual discomfort prediction on stereoscopic 3d images without explicit disparities,” Signal Processing: Image Communication, vol. 51, pp. 50–60, 2017.
[15] P. Campisi, P. Le Callet, and E. Marinin, “Stereoscopic image quality assessment,” in 2007 15th European Signal Processing Conference. IEEE, 2007, pp. 2110–2114.
[16] A. Benoit, P. Le Callet, P. Campisi, and R. Rousseau, “Quality assessment of stereoscopic images with application to virtual reality,” EURASIP journal on image and video processing, vol. 2008, no. 1, p. 659024, 2009.
[17] J. You, L. Xing, A. Perkis, and X. Wang, “Perceptual quality assessment for stereoscopic images based on 2d image quality metrics and disparity analysis,” in Proc. Int. Workshop Video Process. Quality Metrics Conf. Electron, vol. 9, 2010, pp. 1–6.
[18] P. Gorley and N. Holliman, “Stereoscopic image quality metrics and compression,” in Stereoscopic Displays and Applications XIX, vol. 6803. International Society for Optics and Photonics, 2008, p. 680305.
[19] M.-J. Chen, C.-C. Su, D.-K. Kwon, L. K. Cormack, and A. C. Bovik, “Full-reference quality assessment of stereopairs focusing on rivalry,” Signal Processing: Image Communication, vol. 28, no. 9, pp. 1145–1155, 2013.
[20] Y.-H. Lin and J.-L. Wu, “Quality assessment of stereoscopic 3d image compression by binocular integration behaviors,” IEEE transactions on Image Processing, vol. 23, no. 4, pp. 1527–1542, 2014.
[21] F. Qi, D. Zhao, and W. Gao, “Reduced reference stereoscopic image quality assessment based on binocular perceptual information,” IEEE Transactions on multimedia, vol. 17, no. 12, pp. 2338–2344, 2015.
[22] L. Ma, X. Wang, Q. Liu, and K. N. Ngan, “Reorganized dct-based image representation for reduced reference stereoscopic image quality assessment,” Neurocomputing, vol. 215, pp. 21–31, 2016.
[23] J. Ma, P. An, L. Shen, and K. Li, “Reduced-reference stereoscopic image quality assessment using scene natural statistics and structural degradation,” IEEE Access, vol. 6, pp. 2768–2780, 2017.
[24] R. Akhter, Z. P. Sazzad, Y. Horita, and J. Baltes, “No-reference stereoscopic image quality assessment,” in Stereoscopic Displays and Applications XXI, vol. 7524. International Society for Optics and Photonics, 2010, p. 75240T.
[25] Z. Sazzad, R. Akhter, J. Baltes, and Y. Horita, “Objective no-reference stereoscopic image quality prediction based on 2d image features and relative disparity,” Advances in Multimedia, vol. 2012, p. 8, 2012.
[26] M.-J. Chen, L. K. Cormack, and A. C. Bovik, “No-reference quality assessment of natural stereopairs,” IEEE Transactions on Image Processing, vol. 22, no. 9, pp. 3379–3391, 2013.
[27] C.-C. Su, L. K. Cormack, and A. C. Bovik, “Oriented correlation models of distorted natural images with application to natural stereopair quality evaluation,” IEEE Transactions on image processing, vol. 24, no. 5, pp. 1685–1699, 2015.
[28] H. Oh, S. Ahn, J. Kim, and S. Lee, “Blind deep s3d image evaluation via local to global feature aggregation,” IEEE Transactions on Image Processing, vol. 26, no. 10, pp. 4923–4936, 2017.
[29] J. Yang, Y. Zhao, Y. Zhu, H. Xu, W. Lu, and Q. Meng, “Blind assessment for stereo images considering binocular characteristics and deep perception map based on deep belief network,” Information Sciences, vol. 474, pp. 1–17, 2019.
[30] W. Zhou, Z. Chen, and W. Li, “Dual-stream interactive networks for no-reference stereoscopic image quality assessment,” IEEE Transactions on Image Processing, 2019.
[31] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in neural information processing systems, 2012, pp. 1097–1105.
[32] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.
