# 3D saliency guided deep quality predictor for no-reference stereoscopic images

Messai Laboratoire Arc, Université Des Frères Mentouri Constantine 1, Algeria Oussama, Aladine Chetouani, Fella Hachouf, Zianou Ahmed Seghir

Version of Record: https://www.sciencedirect.com/science/article/pii/S0925231222000029
Manuscript_3d910ea09a7fa5a6a1e6610d8929b595



# 3D Saliency guided Deep Quality predictor for No-Reference Stereoscopic Images

Oussama Messai +*, Aladine Chetouani •, Fella Hachouf +, Zianou Ahmed Seghir **
+ Laboratoire ARC, Université des Frères Mentouri Constantine 1, Algérie.
* Univ Lyon, Lyon 2, LIRIS, F-69676 Lyon, France
• PRISME Laboratory, University of Orleans, France.
** Computing Department, University of Abbes laghrour Khenchela, Algeria.

***Abstract*—The use of 3D technologies is growing rapidly, and stereoscopic imaging is usually used to display the 3D contents. However, compression, transmission and other necessary treatments may reduce the quality of these images. Stereo Image Quality Assessment (SIQA) has attracted more attention to ensure good viewing experience for the users and thus several methods have been proposed in the literature with a clear improvement for deep learning-based methods. This paper introduces a new deep learning-based no-reference SIQA using cyclopean view hypothesis and human visual attention. First, the cyclopean image is constructed considering the presence of binocular rivalry that covers the asymmetric distortion case. Second, the saliency map is computed considering the depth information. The latter aims to extract patches on the most perceptual relevant regions. Finally, a modified version of the pre-trained Convolutional Neural Network (CNN) is fine-tuned and used to predict the quality score through the selected patches. Five distinct pre-trained models were analyzed and compared in term of results. The performance of the proposed metric has been evaluated on four commonly used datasets (3D LIVE phase I and phase II databases as well as Waterloo IVC 3D Phase 1 and Phase 2). Compared with the state-of-the-art metrics, the proposed method gives better outcomes. *The implementation code will be made accessible to the public at: https://github.com/o-messai/3D-NR-SIQA***

***Index Terms*—Stereoscopic Image Quality Assessment (SIQA), No-reference, Cyclopean view, 3D Saliency, Convolutional Neural Network (CNN), Deep learning.**

## I. INTRODUCTION

THE use of 3D technologies is becoming increasingly attractive for various academic and industrial applications (e.g., entertainments, 3D visualization, robotic navigation, medical surgery, etc.) [1], [2]. However, treatments usually applied to capture, transmit or display the content (i.e. compression, transmission, etc.) may impact the perceived quality. However, where these distortions are unavoidable, subjective quality evaluation cannot provide an autonomous system with real-time feedback. With the rapid development of 3D digital systems over the last decade, the Image Quality Assessment (IQA) approach has played an important role in providing relevant information to test, improve, benchmark, and monitor the process. Therefore, efficient metrics are critical for improving the 3D technologies and ensuring a high quality of experience. Stereo imaging, often known as stereoscopy, is a technique used in most 3D monitors to create the illusion of depth in an image using stereopsis for binocular vision. Stereoscopic Image Quality Assessment (SIQA) approaches, which differ from 2D IQA methods, are then developed for this type of content.

In general, both IQA and SIQA metrics can be divided into two classes: subjective and objective methods. Subjective assessment is based on opinion scores given by human observers. It is mostly expressed in terms of Mean Opinion Score (MOS) or Difference Mean Opinion Score (DMOS). This approach is effective and reliable for assessing perceptual quality, but it has some drawbacks such as time consuming, high cost and it is not applicable for online applications. In the meantime, objective evaluation provides an automated assessment that addresses the limitations of subjective assessment. Therefore, a lot of efforts has been dedicated to design accurate IQA/SIQA methods. However, the existing objective methods are divided into three categories: Full-Reference (FR), Reduced-Reference (RR), and No-Reference (NR) metrics. FR methods utilize the reference stereo image, RR methods use only partial information of the reference image, while the reference stereo image is completely unavailable for NR metrics. As a result, in practice, NR models are more useful in most applications.

In the majority of NR-SIQA model designs, quality indicators, also known as quality-aware features of image structure, play essential roles. The distortions added to images generate changes in structural features which can be captured by structural feature statistics. Based on how these quality-aware features are calculated, NR models can be further categorized into machine learning-based methods and training-free based methods. Training-free approaches have an internal generalization potential, and yet, their performances are currently inferior to machine learning-based methods. Instead, using machine learning techniques such as Support Vector Regression (SVR) and Random Forest (RF), image feature values can be simply mapped to the image quality index, assisting machine learning-based NR-IQA models to obtain comparatively higher evaluation performance. Furthermore in latest years, deep-learning-based algorithms that directly map an image or image structure to a quality index have achieved promising results. But, there are several flaws to this latter, such as fixed input pixel resolution, pixel attack sensitivity, and large scale training data requirement.

Regardless of the learning approach employed in the SIQA system, modeling Human Visual System (HVS) is important to simulate the visual judgment. However, the HVS is a complex

visual process and still an open question for researchers. For SIQA development, many researchers have used fusion hypothesizes of the perceived left and right eye signals called cyclopean view [3], [4]. Meanwhile, in most of the suggested SIQA approaches the human visual attention is not explored.

In this paper, we focus on the use of 3D saliency information which can be a step forward to the best human visual system simulation. However, since that the majority of NR-SIQA approaches do not explore the human visual attention information, we introduce a new NR-SIQA metric that exploits the benefit of 3D saliency map integrated with cyclopean view and deep feature learning. The proposed method has the following distinctive features:

- 1) The cyclopean image is well-known for its good performance, specifically with respect to asymmetric distortion, where a conventional gray tone of cyclopean image is being used. In our model, we compute a cyclopean image using RGB (Red, Green, Blue) color channels rather than gray one to maintain the same spatial domain being viewed by the observer.
- 2) The use of 3D saliency map combined with a well defined deep quality predictor help to improve the overall performance and outperforms state-of-the-art metrics.
- 3) Compared to similar deep-learning-based metrics, our method is fully blind and uses the most salient region patches rather than using all scene patches. This optimizes the complexity and run time of the suggested process. In addition, the simplicity of the proposed approach is a benefit for implementation or even integration with other algorithms, such as quality enhancement.

The rest of this paper is structured as follows. Section II presents related work. The proposed approach is described in Section III. Section IV shows experimental results. Finally, section V concludes the paper.

## II. RELATED WORK

With the rise of stereoscopic imaging in different applications which demand Quality of Experience (QoE) assessment. Several metrics based on the use of 2D Image Quality Assessment (IQA) metrics have been proposed in the literature. Generally, 2D IQA metrics can be extended to stereoscopic images. These 2D-extended metrics usually extract feature vectors separately from the left and right images. They are weight-averaged to obtain the final feature vector for training. In the meanwhile, other improved 2D IQA metrics tend to use the disparity/depth map either by adding it in the feature extraction process or by incorporating it into the original design. However, the following discusses 2D-IQA measures that have mostly been tested on stereoscopic images :

- **2D-IQA metrics**: For instance, Gorley *et al*. [5] did not use or measure disparity/depth information. They compute quality scores on matched feature points delivered by SIFT (Scale-Invariant Feature Transform) [6] and RANSAC (RANdom Sample Consensus) [7] applied to the left and right views. Different strategies have been applied to derive a quality score through these 2D metrics (mean, weight, etc.) [8], [9]. This kind of metrics, however, is still gaining popularity among researchers for the growing number of multimedia on Internet, where novel 2D IQA measures is being developed and potentially are applicable to stereoscopic images. For instance, Gu et *et al* [10] applied convolution operations at multi-scale to the input image, where gradient magnitude, and color information similarity are extracted as features. In [11], a training-free metric has been proposed for contrast distorted images using histogram information and salient regions. Moreover, we have recently noticed that CNN algorithms are increasingly being used to address quality evaluation problem. For example, a deep regression model is used after identifying the most salient patches from the input image to perform 2D IQA without reference in [12]. While in [13], a handcrafted quality features are combined with deep CNN extracted features to propose a 2D IQA metric.

However, most of these approaches do not consider the asymmetrical distortion case and thus fail to estimate the quality of the latter. The overall results of these extended metrics are not as good as for 2D images, which motivates to have metrics dealing with 3D perception. The issue of asymmetric distortion in stereoscopic images is related to binocular rivalry/suppression, which occurs when the eyes of a viewer see different scenes. This phenomenon often causes fatigue and visual discomfort to the observers [14]. Recently, metrics that exploit HVS characteristics are more introduced and showed better results. For instance, methods in [15], [3], [16] use cyclopean view hypothesis that seems consistent for the assessment, especially for asymmetric distortions. Overall, there are very few SIQA methods that explore visual saliency information. However, these state-of-the-art SIQA metrics are discussed by their category in the following:

- **FR-SIQA metrics**: Chen *et al.* [3] have proposed an FR quality assessment model that utilizes the linear expression of cyclopean view influenced by binocular rivalry between left and right views. In [17], the author has also used the cyclopean image hypothesis and proposed a new metric based on 2D FR-IQA fusion. Authors of [18] have proposed a model that combines two measures. They first measure the difference between the left and right reference images and the distorted ones. Then, they compute the difference between the pure stereo image disparity map and the deformed ones. You *et al.* [8] have developed model where they incorporate 2D-IQA metrics with disparity information. Another metric named Binocular Energy Quality Metric (BEQM) has been proposed by Bensalma *et al* [4]. They have measured the stereoscopic image quality by calculating the binocular energy variation between the reference and distorted stereo-pairs. The authors of [19], [5] have proposed a PSNR-based stereo IQA models. Hewage *et al*. [19] have extracted edge maps from the disparity maps of the reference and distorted stereo-pairs. The PSNR is then computed between the reference and test edge maps to assess the quality. Instead of using the disparity/depth

information, Gorley *et al*. [5] calculate quality scores on corresponded feature points of the left and right images provided by SIFT [6] and RANSAC [7].
However, full-reference metrics continue to pique the interest of researchers. For instance, metrics based on binocular receptive field properties have been proposed in [20], [21]. In [20], to determine the best features that imitate the reactions of basic cells in the brain, an Independent Component Analysis (ICA) technique was applied. While the scheme in [21] tends to learn a multi-scale dictionary from the training database. In the quality estimation phase, they calculate a sparse feature similarity index based on the estimated sparse coefficient vectors. The latter (e.g coefficient vectors) is built with phase and amplitude differences in mind as well as a global luminance similarity index that takes luminance changes into account. A similar FR method by human binocular perception was proposed in [22]. More precisely, the binocular perceptual properties of simple and complex cells are simulated. For simple cells simulation, which is assumed to represent a monocular cue, the authors have used a push–pull combination of receptive fields response. While for complex cells, which are used to represent a binocular cue, are simulated by using binocular energy response and binocular rivalry response. Following the simulation phase, quality-aware characteristics are extracted from the responses using a self-weighted histogram, and similarity measurement is used to determine the quality score. Furthermore, another recent metric based on monocular and binocular visual features in [23]. First, the authors suggested a segmentation strategy to find occluded and non-occluded areas in the scene by using disparity information and Euclidean distance between stereo pairs. The occluded regions are considered to represent the monocular vision while non-occluded regions to reveal the binocular vision of the HVS. Global and local features are then extracted from the regions and used to predict the visual quality.

- **RR-SIQA metrics**: Authors in [24] have utilized binocular perceptual information to perform an RR quality measurement, while Ma et al [25] have characterized the statistical properties of stereoscopic images in the reorganized Discrete Cosine Transform (RDCT) domain. In [26], another RR method based on Natural Scene Statistics (NSS) and structural degradation has been also proposed.
- **NR-SIQA metrics**: The reference-less SIQA metrics also have attracted researchers. For instance, Akhter *et al.* [27] have designed a reference-less SIQA algorithm. They extract features from the disparity map and the stereo-pairs. In [28], the authors proposed a new NR framework based on a degradation identification and fusion steps of features. Zhou *et al.* [29] have simulated binocular phenomenon and they used the well known k-Nearest Neighbors (KNN). While Fang *et al.* [30] have proposed an unsupervised model for stereoscopic images. From the monocular and cyclopean view patches, they extract quality indicators in spatial and frequency domains. Then, Bhattacharya distance has been used to get a quality score. Appina recently introduced another unsupervised measure in [31], where saliency information is incorporated in the creation of a cyclopean image. The quality score is then estimated using a model based on multi-orient subband decomposition of the cyclopean image. Deep convolutional network predictors also have being used. For example in [32], a local patches are extracted and then combined to obtain global features using an aggregation layer in the network. While authors in [33] have modeled the human visual cortex using the deep auto-encoder. The auto-encoder is also used in [34] to achieve high-level features, where the authors first compute a gray level cyclopean image, difference and summation image from the input stereoscopic view. Then, series of feature extraction have been conducted from these images. In [35], the authors have considered the deep perception map and binocular weight model to predict the perceived stereo image quality. Meanwhile, Zhou *et al.* have suggested a metric called StereoQA-Net [36] using a novel end-to-end dual convolutional network. Xu *et al.* [37] have simulated our human brain cognition process to propose NR metric using the deep encoder-decoder network. Other recent metrics have been proposed that utilize deep models. For instance, authors in [38] have used deep sub-networks in a single model to extract primary, local and global features from the input left and right images. These features are eventually concatenated for quality score regression. Similarly, the authors of [39] proposed another end-to-end CNN-based measure that uses three sub-networks for monocular feature encoding, binocular feature fusion and a third network for quality prediction. Meanwhile, authors in [40], deployed saliency information to determine salient and non-salient patches for local features extraction. The authors used reference stereoscopic images to compute local quality maps that are then used as labels to train deep networks. Image segmentation technique also has been deployed for NR-SIQA methods. Where in [41] a superpixel segmentation is used based on K-mean clustering approach [42]. Then, from these superpixel regions, a spatial entropy and NSS features are extracted to obtain quality ratings using regression model.

## III. PROPOSED METHOD

The general framework of the proposed method is summarized in Fig. 1. From a given stereo image, the cyclopean image is first calculated, allowing to consider the binocular rivalry phenomenon as mentioned above [14]. Then, the 3D saliency map of the stereo image is computed. It aims to focus on regions that attract more our perception. After having thresholded the obtained 3D saliency map, small patches are extracted and fed into a CNN model in order to predict the overall quality of the stereo image. Each of these steps is described in coming subsection.

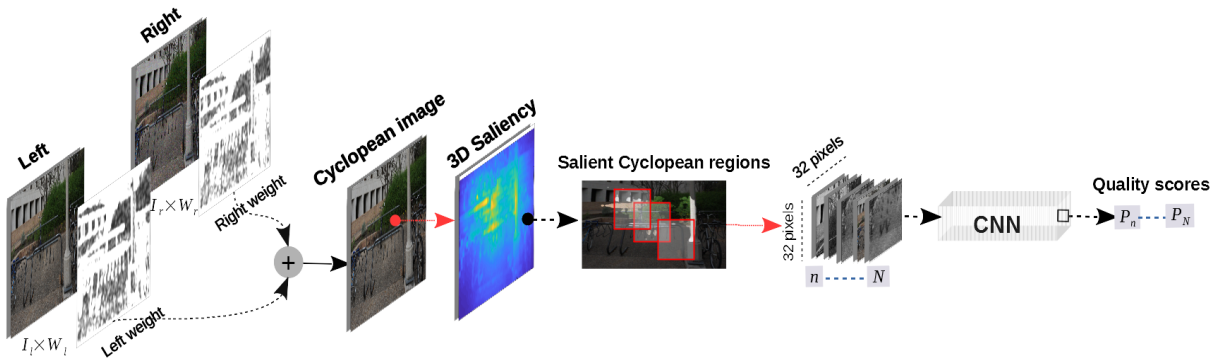
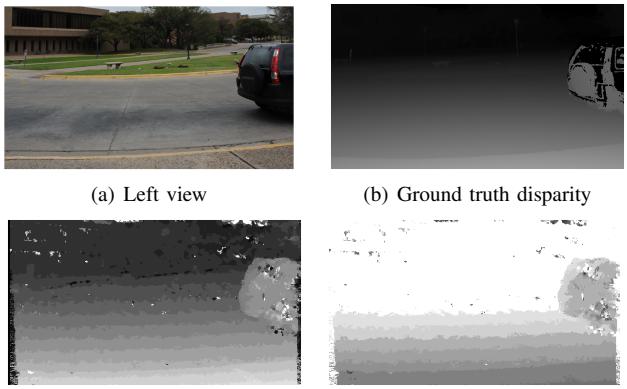



Fig. 1: Flowchart of the proposed metric.



(a) Left view (b) Ground truth disparity

(c) Estimated disparity (d) Estimated depth

Fig. 2: An example of disparity and depth map estimation from stereoscopic image.

### A. *disparity/depth map*

Disparity and depth maps are important in many applications such as augmented reality [43], 3D object detection and mapping [44], [45]. These maps are also essential to develop efficient SIQA metrics, since degradation on disparity/depth maps may cause visual discomfort/fatigue that definitely influences the overall quality of the 3D stereo images [46].

The disparity map is here computed using an SSIM-based method which is an upgraded version of the Sum of Absolute Differences (SAD) stereo matching algorithm [47]. It consists of selecting the best matches through the SSIM metric [48] instead of SAD. More precisely, SSIM scores between the current block from left image and right image blocks along the horizontal direction are maximized and the disparity map is given by the difference between the current pixel and the best SSIM location. The block size was fixed to $7 \times 7$ and the maximum disparity distance was set to 25. Fig. 2 shows an example of obtained disparity map and its correspond depth map conversion.

### B. *Cyclopean image*

The cyclopean image purpose is to simulate the human brain fusion of the perceived signal from the left and right eyes. A study has been conducted in [16] to exhibit the benefit of using cyclopean image for SIQA. Where the use and non-use of cyclopean hypothesis has been analyzed. The comparison results indicated better accuracy when cyclopean image is being deployed.

Inspired by the model used in previous work [16], we construct a cyclopean image over three channels Red, Green, and Blue (RGB) rather than one gray channel to maintain the distortion effects on the stereo image. The formula used is as follows:

$$C(x,y)_n = w_l(x,y)_n \times I_l(x,y)_n + w_r(x+d,y)_n \times I_r(x+d,y)_n \quad (1)$$

where $C$ refers to the cyclopean image and $n$ for the color channel number in-which $n \in \{R, G, B\}$. Left and right views are represented by $I_l$ and $I_r$, respectively. $d$ is the disparity index that matches pixels from left image $I_l$ with those in right image $I_r$. While $w_l$ and $w_r$ are the weights of the left and right eyes, respectively. The weights $w_l$ and $w_r$ are given by:

$$w_l(x,y) = \frac{GI_l(x,y)}{GI_l(x,y) + GI_r(x+d,y)} \quad (2)$$

$$w_r(x+d,y) = \frac{GI_r(x+d,y)}{GI_l(x,y) + GI_r(x+d,y)} \quad (3)$$

where $GI_l$ and $GI_r$ are the summation of Gabor filter the magnitude responses over eight orientations for left and right views respectively. The use of weight coefficients helps to consider asymmetric distortions. However Gabor filter has a wide range of applications and is theoretically related to the function of primary human visual cortex cells in primates because it extracts features of luminance and chromatic channels [49]. Thus, this type of filter is commonly utilized for HVS modeling in several IQA metrics. As an example presented in Fig. 3, sub-Fig. (a) shows RGB cyclopean image formed from the left image in sub-fig (b) that is not distorted and the right image in sub-fig (c) that suffers from WN distortion. It is worth noticed that this asymmetric distortion is stated clearly onto the cyclopean image (the red boxes).

### C. *3D Saliency map*

Visual attention/saliency is an important characteristic of our HVS since it represents the regions of the image in

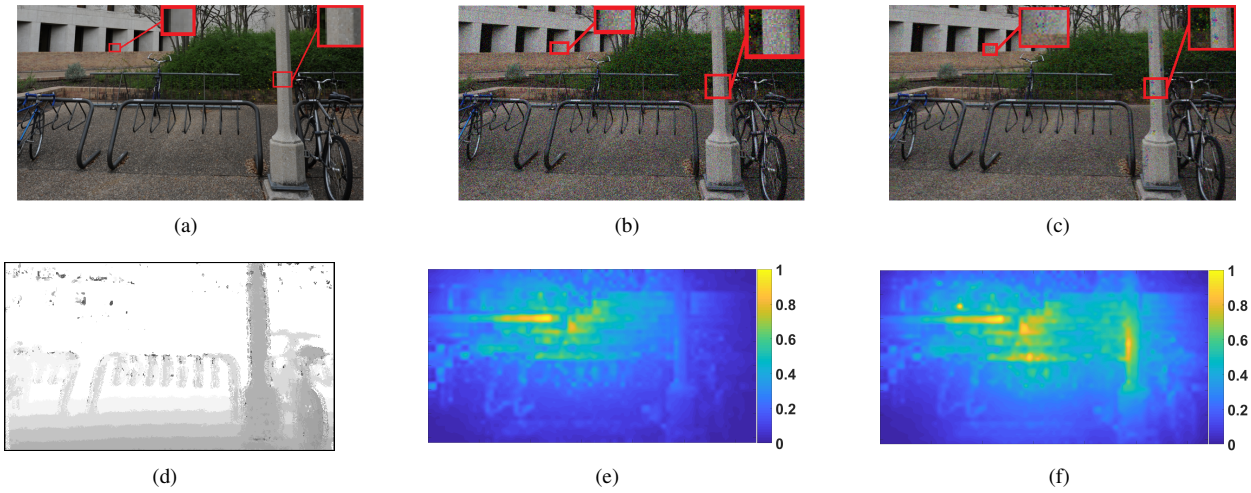



Fig. 3: Saliency of a stereo image: (a) Left view without distortion, (b) Right view with White Noise (WN) distortion, (c) Synthesized RGB cyclopean image, (d) estimated depth using disparity map, (e) 2D saliency map and (f) the used 3D saliency map.

which the observer focus the most. Hence, salient regions impact more the subjective scores given by the observers and thus the quality of a given image is highly related to these regions. However, the visual saliency in stereoscopic images is gaining more and more attraction among researchers. For example, a recently published method in [50] that predicts salient objects in stereoscopic images using an end-to-end Deep Convolutional Residual Autoencoder (DCRA) that takes stereoscopic image and disparity map as inputs.

Despite its recent popularity and potential step forward in HVS simulation, the 3D saliency map is still not given much consideration in the NR-SIQA metrics. According to this observation, 3D saliency map has been used in this study to extract perceptual relevant patches instead of all patches. The 3D saliency method suggested in [51] has been here used for its low complexity and ease of use. Also, this method is based on the integration of the depth information and 2D saliency maps. The saliency map of the luminance, color and texture from one view are first computed [52]. Then, the depth map is calculated through the left and right views as shown in Fig.3.d. After a normalization step, the 3D saliency map is finally given by averaging the achieved maps. For comparison example, we compute non-depth saliency map (i.e. 2D saliency) and depth saliency map (i.e. 3D saliency) displayed in Fig. 3.e and Fig. 3.f, respectively. We can see that the 3D saliency map gives more importance to near objects than the 2D one because the algorithm incorporates the depth map.

The 3D saliency map is then normalized (using min-max normalization) and thresholded to extract patches of size 32x32x3 from the cyclopean image allowing thus to focus only on the most salient regions. The extracted patches are then fed to a CNN model to predict the quality. After several tests, the threshold has been fixed to 0.3. The impact of the threshold value on the performance is presented in Section IV-C.

### *D. Quality prediction model*

Several CNN models with different architectures have been proposed in the literature. In this paper, performances of five pre-trained models widely used have been compared, briefly described above:

- **AlexNet [53]:** Developed in 2012, the AlexNet model is one of the pioneering models proposed by Alex Krizhevsky. This model highlighted the relevance of using CNN models for classification tasks. Composed of 5 convolutional layers and 3 FC layers, the authors stressed three main points: the use of the Relu (Rectified Linear Units) function, the exploitation of the dropout to prevent the over-fitting and overlap during the pooling step.
- **VGG16 and VGG19 [54]:** have been proposed in 2014. VGG models were developed by the Oxford Visual Geometry Group. To increase the ability of the model to discriminate between objects, the authors integrated more non-linearities by using convolutional layers with 3x3 filters instead of 7x7 filters. Several versions were proposed with 11 (VGG11), 13 (VGG13), 16 (VGG16) and 19 (VGG19) layers. Here, VGG16 and VGG19 are used and compared.
- **ResNet18 and ResNet50** [55]: In 2015, a Residual Neural Network (ResNet) model was proposed. This model stands out by its integration of a residual module. The idea developed by the authors is to reformulate the output (H(x)=F(x)) of each series of Conv-ReLu-Conv by adding the input $\boldsymbol{x}$ as information ($\boldsymbol{H(x) = F(x)+x}$). Different versions are available: ResNet18 (18 layers), ResNet34 (34 layers), ResNet50 (50 layers), ResNet152 (152 layers) and so on. ResNet18 and ResNet50 are used in this study.

The use of these models allows to compare different depths (from a shallow model i.e. AlexNet to deeper models i.e.

TABLE I: Pre-trained models descriptions.

| Model | Size | Learnable parameters (Millions) | Depth |
|---|---|---|---|
| AlexNet | 227 MB | 61.0 | 8 |
| VGG-16 | 528 MB | 138 | 16 |
| VGG-19 | 549 MB | 144 | 19 |
| ResNet18 | 44 MB | 11.7 | 18 |
| ResNet50 | 98 MB | 25.6 | 50 |

the other models), different architectures (ResNet and VGG) as well as same architecture with different depths (VGG16 against VGG19 and ResNet18 against ResNet50).

Each of these models has its specificities as shown in Table I that compares the used pre-trained models in terms of memory size and amount of learned parameters. The network depth refers to the largest number of sequential convolution or fully connected layers on the path from the input layer to the output layer. They have a distinct number of learnable parameters and different depth sizes. This diversity will drive us to the best architectures that are suited for quality assessment. It is worth noticed that these models were modified and fine-tuned to adapt their learnable parameters to our context.

## IV. EXPERIMENTAL RESULTS

### A. Datasets and Training Protocol

To examine the consistency and effectiveness of our method, four databases have been used to evaluate the performance of our metric. These datasets are listed in Table II and briefly described below:

**3D LIVE phase I (LIVE-P1) [56]**: It consists of 365 distorted stereo images of size 360 x 640 pixels generated from 20 stereo image scenes. Five degradation types are considered (White Noise: WN, JPEG2000: JP2K, JPEG, Fast Fading: FF and Blur). All the distortions are carried out symmetrically. The subjective evaluation scores are given in the term of DMOS within the range of [-10,70].

**3D LIVE phase II (LIVE-P2) [3]**: It contains 360 distorted stereo images with the same size and distortion types as phase I. This database includes symmetric as well as asymmetric distortions. Subjective evaluation scores are given within the range of [20,80] in the DMOS term.

**Waterloo IVC 3D Phase 1 (P-1) [57]**: It has 330 full HD (1920 x 1080 pixels) distorted stereo images derived from six pristine stereo images collected from the Middlebury Stereo 2005 Datasets. Three forms of distortion are present in this database: additive white Gaussian noise, Gaussian blur, and JPEG compression. Subjective evaluation scores are given in term of MOS and distributed in the interval of [10,100].

**Waterloo IVC 3D Phase 2 (P-2) [58]**: It contains 460 full HD stereo images created from 10 pristine stereo image pairs. The stereo images carry the same distortion types as Phase 1, and both of them include symmetric and asymmetric distortions. Subjective assessment scores are in term of MOS and the range is the same of Waterloo-P1 ([10,100]).

It is worth noting that the asymmetric degradations in the Waterloo P-1 and P-2 databases are different from those in the LIVE-II database. LIVE-II uses only one type of distortion to perform the asymmetry, while the two Waterloo databases consider the possibility of multiple types of degradation in which the left and the right images are affected by different distortions.

Generally, the above-described SIQA databases have small-limited labelled images. To increase the amount of data, data augmentation is often applied. The available data augmentation techniques except horizontal flipping, affects the subjective quality ratings. The rotation and re-sizing approaches often applied change the observers perception of spatial details and are thus not appropriate for SIQA methods. Therefore in this work, neither rotation nor translation or re-sizing were applied. Instead of, we allow a maximum of 80% overlapping between patches. During the learning, each pre-trained CNN model is fine-tuned for 50 epochs using a learning rate of 0.01. Stochastic Gradient Descent (SGD) with a momentum equals to 0.9 is used as optimization function. The human scores are normalized in the form of DMOS/MOS to min-max normalization [0,1]. The closer to 0 the better quality of the stereo image is for DMOS and the opposite for MOS. Afterall, the expected quality rating for each scene is the average of quality scores obtained from patches, described as follows:

$$Q = \frac{1}{N}\sum_{n=1}^{N} P_n \tag{4}$$

Where $P_n$ is the predicted score for the $n^{th}$ patch, $N$ is the number of patches, and $Q$ is the final quality score. We have carried out 10-fold validation test by randomly splitting the dataset into training (80%) and test (20%) at each time. The average result is then used as evaluation criterion. We also evaluate the generalization ability of our method by applying a cross-dataset evaluation.

### B. Evaluation Criteria

The performance has been measured across three well-known metrics [59]: The *RMSE*, the Spearman's rank-order correlation coefficient (*SROCC*), and the Pearson linear correlation coefficient (*PLCC*). The metrics have been computed between the predicted quality scores (objective scores) and the subjective ones (*DMOS/MOS*). *PLCC* and *RMSE* measure the assessment accuracy, while *SROCC* evaluates the prediction notability. Higher values for *PLCC* and *SROCC* (closer to 1) and lower values for *RMSE* (closer to 0) indicate superior linear rank-order correlation and better precision with respect to human quality judgments, respectively. Objective scores are fitted to the subjective ones using logistic function [60]. This function is based on five parameters ($\theta_1$, $\theta_2$, $\theta_3$, $\theta_4$ and $\theta_5$). The logistic mapping function used for the nonlinear regression is introduced by the following equation:

$$Q_{map} = \theta_1 \left( \frac{1}{2} - \frac{1}{\exp\left(\theta_2 \left(Q - \theta_3\right)\right)} \right) + \theta_4 Q + \theta_5 \tag{5}$$

Where $Q$ and $Q_{map}$ are the objective quality scores before and after the nonlinear mapping, respectively. $\theta_i$ ($i$ = 1 to 5) are selected for the most excellent fit.

TABLE II: Summary of the four databases.

| Database | # of Reference scenes | Resolution | # of images (Sym., Asym.) | Distortions |
|---|---|---|---|---|
| 3D LIVE P-I | 20 | 360 x 640 | 365 (365, 0) | JP2K, JPEG, WN, Blur, FF |
| 3D LIVE P-II | 8 | 360 x 640 | 360 (120, 240) | JP2K, JPEG, WN, Blur, FF |
| Waterloo IVC 3D P-I | 6 | 1080 x 1920 | 330 (180, 150) | JPEG, WN, Blur |
| Waterloo IVC 3D P-II | 10 | 1080 x 1920 | 460 (210, 250) | JPEG, WN, Blur |

TABLE III: PLCC results of different deep models versus saliency threshold on LIVE-P2.

| Saliency threshold | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|---|---|---|---|---|---|---|---|---|---|
| Number of patches | 82277 | 68900 | 46300 | 23968 | 12276 | 5466 | 1692 | 557 | 369 |
| AlexNet | 0.960 | 0.968 | 0.970 | 0.969 | 0.959 | 0.906 | 0.870 | 0.832 | 0.801 |
| VGG-16 | **0.977** | **0.984** | **0.985** | **0.983** | **0.981** | **0.945** | **0.907** | **0.891** | **0.881** |
| VGG-19 | **0.977** | 0.983 | 0.984 | 0.982 | 0.980 | 0.943 | **0.907** | 0.890 | **0.881** |
| ResNet18 | 0.970 | 0.976 | 0.975 | 0.974 | 0.968 | 0.926 | 0.889 | 0.794 | 0.500 |
| ResNet50 | 0.966 | 0.974 | 0.976 | 0.975 | 0.972 | 0.931 | 0.882 | 0.823 | 0.675 |

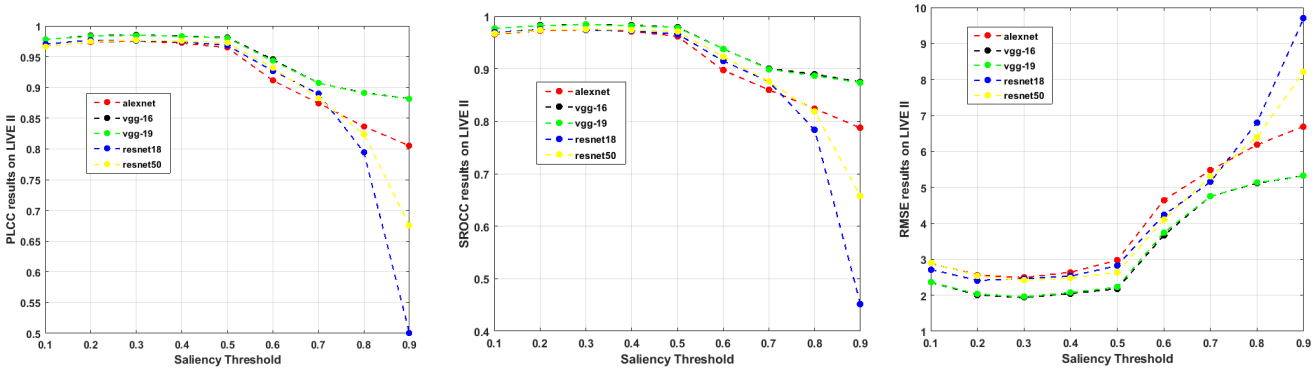



Fig. 4: PLCC, SROCC and RSME comparison results of pre-trained models versus different thresholds on LIVE-P2.

TABLE IV: Impact of the saliency-guided patch selection and the RGB cyclopean image on the performance using VGG-16 and a saliency threshold of 0.3. The tests were carried-out on LIVE-P2 dataset.

| Method | Input Stereoscopic image | LIVE-P2 SROCC | LIVE-P2 PLCC | LIVE-P2 RMSE |
|---|---|---|---|---|
| Saliency guided | RGB | **0.984** | **0.985** | **1.938** |
| | Gray | 0.953 | 0.960 | 3.829 |
| Without saliency | RGB | 0.958 | 0.961 | 3.814 |
| | Gray | 0.931 | 0.942 | 4.011 |

### *C. Different Saliency thresholds and predictors analysis*

In this section, many tests have been conducted to define the best network architecture and to identify suitable saliency threshold. Saliency-based patches are extracted with regard to threshold value. The five pre-trained models are adjusted and tested using the same train configurations as discussed in section III-D. Starting with value of 0.1, we update the threshold and notice the performance using the LIVE P-2 database in Table III. The Table also includes the number of patches extracted at each threshold. Obtained PLCC results show that the VGG-16 and VGG-19 architectures are better for mapping the extracted patches to quality scores. From plots in Fig. 4, we notice that using different saliency-based cropping thresholds influence the quality prediction with best threshold value of 0.3. As we increase the starting value, we get better results for all models. After threshold of 0.3, the coefficient correlations decrease while saliency thresholds cropping increase. The fact that higher saliency threshold gives smaller datasets, it may limit the model to learn best quality prediction from the salient regions. For instance, 0.3 gives 46 300 patches for training, while only 12 276 patches for 0.5. This is a trade-off between the saliency threshold and the training dataset size that need to be balanced. For example, although using a threshold of 0.1 that yields more training sets (i.e. 82277 patches), the better precision results are still obtained with a threshold equal to 0.3. Based on these results, the saliency-guided cropping step allows to considerably improve the performance. Notice that the performance drops for thresholds which offer small train datasets, such as the 0.6 threshold.

Moreover, AlexNet gives the lowest correlation performance among all models. VGG-16 and VGG-19 yield similar correlation performance with little differences since they have nearly the same architecture. These models contain more series of convolutional layers and thus extract higher and better quality indicators for prediction. In the meantime, going deeper than VGG-16 model, ResNet18 and ResNet50 regressors appear to slightly diverge from the path toward the best quality predictions. For instance, using the best saliency threshold of value 0.3, AlexNet model with performance of RMSE = 2.491 comes in the last place compared to the the other four networks.

VGG-16 and ResNet18 behave slightly better compared to deeper ones; VGG-19 and ResNet50, respectively. The RMSE is 1.938 for VGG-16 and 2.416 for ResNet18, while the error values for VGG-19 and ResNet50 are 1.957 and 2.459, respectively. Meanwhile, analyzing the same architectures and different depths, VGG-16 performs better than VGG-19. Also ResNet18 provides better results than ResNet50. Despite that going deeper with convolutions improves the accuracy in object recognition/classification tasks, for regression problems it might not perform well. Allowing the network to perform more convolutions does not necessary imply extraction of more precision quality-features. Since the majority of degradations affects the stereoscopic image at the pixel level. The deployed $3\times3$ convolutional layers in VGG and ResNet extract discriminative features from these local distortions. Therefore, we believe that deeper models are more likely to lose quality-aware features across the network. This could explain why increasing the depth of ResNet and VGG architectures has a negative impact on performance.

After the selection of the best pre-trained model and saliency threshold, we evaluated the impact of the saliency-based patch selection and the RGB cyclopean image. For the no saliency test, all possible patches of the cyclopean image were sequentially extracted by sliding over the whole scene from left to right with a stride of 32 pixel (i.e. without overlap). This creates 220 patches for every scene in the LIVE P-2 database, while 128 patches are approximately cropped for the saliency-guided extraction. Table IV shows PLCC, SROCC and RSME results over LIVE P-2 database with and without saliency-guided patches as well as the grayscale cyclopean versus RGB cyclopean as inputs. As can be seen, the saliency-guided patch selection considerably improves the performance with a quality prediction error decrease of 49% in term of RMSE. The use of RGB cyclopean image allows also to increase quality prediction efficiency in both cases (i.e. with and without the saliency-guided patch selection). During subjective assessments, the ratings are given based on RGB stimulus. The RGB cyclopean is therefore closer to reflect the distorted spatial information experienced by the observer. The best result is reached when both are considered. This experiment supports the use of the saliency map and the RGB cyclopean image for the SIQA. Moreover besides accuracy improvement, the saliency guidance approach may also decrease the cost and run-time, since the approach uses recommended patches rather than using all patches of the scene.

### D. Patch-size effect on quality evaluation

TABLE V: Performance versus patch size on Waterloo-P1 and LIVE-P2 databases.

| | Waterloo-P1 | | | LIVE-P2 | | |
|---|---|---|---|---|---|---|
| Patch size (in pixels) | SROCC | PLCC | RMSE | SROCC | PLCC | RMSE |
| 32 x 32 | 0.946 | 0.960 | 4.376 | **0.984** | **0.985** | **1.938** |
| 64 x 64 | **0.967** | **0.973** | **3.592** | 0.975 | 0.978 | 2.342 |
| 96 x 96 | 0.964 | 0.971 | 3.757 | 0.975 | 0.977 | 2.389 |
| 128 x 128 | 0.956 | 0.967 | 4.009 | 0.969 | 0.972 | 2.628 |

The metric implementation needs a fixed size patch for the deep CNN regression/classification stage. In most classification tasks, the model takes the entire image as input, which is typically 224 x 224 pixels. However, the stereo images for our proposed patch-based CNN regression have different aspects and resolutions. Such change would have an impact on the salient selection regions, and the 32 x 32 patch might not be ideal in this situation. In particular, LIVE-P1 and P2 stereo images have 360 x 640 pixels size, while Waterloo-P1 and P2 stereo images have higher resolution with size of 1920 x 1080 pixels. Tests have been conducted for this manner using the VGG-16 and 0.3 saliency threshold for their best fit. We increase the patch size by 32 x 32 pixels each time and notice the effect on quality prediction performance using the three indexes; PLCC, SROCC, and RSME. Table V show the results of these tests.

Performance results demonstrate that increasing patch size can improve the performance for higher resolution stereo images such in Waterloo-P1 and P2 databases. The best patch size for LIVE-P2 is 32 x 32 and 64 x 64 for Waterloo-P1. Typically higher resolution images give the viewer a larger salient region, and increase the number of extracted patches for 32 x 32 pixels cropping.

The number of patches extracted must be balanced by the resolution of the stereo image. Therefore, the patches size relies on the resolution of salient region seen by the observer.

### E. Comparison with the State-of-the-Art

Obtained results have been compared with several FR, RR and NR SIQA. Among them, there are recent blind metrics based on the use of CNN models, namely PAD-Net [35], Chen [62], Zhou [39] and Sun [40].

Table VI shows the results of these methods on both LIVE-P1 and P2 datasets. Best metric of each category (FR, RR and NR) is represented on bold and the best one whatever the category is with a gray background. As can be seen, our metric outperforms all the compared NR metrics on both databases. The top best FR metrics are the ones proposed in *et al* [22], [23], while the method proposed by Ma et al [26] achieved the best performance among all the compared RR methods. On LIVE-P1, compared to the best metrics in each category (i.e. Chen for FR and Ma for RR) the improvements in term of PLCC are 7% for FR and 5.6% for RR. While on LIVE-P2, the improvements are 8.8% for FR and 6.3% for RR.

Overall, the performance of our method from both LIVE datasets is somewhat equivalent with slight advantage for LIVE-P2. Indeed, the PLCC and SROCC values obtained for LIVE-P1 are respectively 0.982 and 0.981, while those obtained for LIVE-P2 are 0.984 and 0.985, respectively. Furthermore, we report the performance of our method according to the size of the training set. Table VIII shows the correlations achieved for a training set of size 50%, 70% and 80% using LIVE databases. The partition ratio has a slight impact on the performance. And it does not suffer from an over-fitting problem. The diminution is similar for both datasets. Meanwhile, performance evaluation on Waterloo datasets are not reported in several metrics papers. Table VII shows the state-of-the-art comparison using Waterloo-P1 and Waterloo-P2 databases. In comparison with two FR metrics and four NR

TABLE VI: Overall performance comparison on LIVE-P1 and LIVE-P2.

| | | LIVE-P1 | | | LIVE-P2 | | |
|---|---|---|---|---|---|---|---|
| Type | Metrics | SROCC | PLCC | RMSE | SROCC | PLCC | RMSE |
| FR | Benoit [18] | 0.899 | 0.902 | 7.061 | 0.728 | 0.748 | 7.490 |
| | You [8] | 0.878 | 0.881 | 7.746 | 0.786 | 0.800 | 6.772 |
| | Gorley [5] | 0.142 | 0.451 | 14.635 | 0.146 | 0.515 | 9.675 |
| | Chen [3] | 0.916 | 0.917 | 6.533 | 0.889 | 0.900 | 4.987 |
| | Hewage [19] | 0.501 | 0.558 | 9.364 | 0.501 | 0.558 | 9.364 |
| | Bensalma [4] | 0.874 | 0.887 | 7.558 | 7.558 | 0.769 | 7.203 |
| | Geng [20] | 0.932 | 0.943 | 5.514 | 0.919 | 0.921 | 5.400 |
| | Ma [22] | 0.934 | **0.946** | **5.211** | 0.921 | **0.930** | **4.123** |
| | Si [23] | **0.942** | 0.944 | 5.312 | **0.924** | 0.927 | 6.193 |
| RR | RR-BPI [24] | - | - | - | 0.867 | 0.915 | 4.409 |
| | RR-RDCT [25] | 0.905 | 0.906 | 6.954 | 0.809 | 0.843 | 6.069 |
| | Ma [26] | **0.929** | **0.930** | **6.024** | **0.918** | **0.921** | **4.390** |
| NR | Akhter [27] | 0.383 | 0.626 | 14.827 | 0.543 | 0.568 | 9.294 |
| | Zhou [29] | 0.901 | 0.929 | 6.010 | 0.819 | 0.856 | 6.041 |
| | Fang [30] | 0.877 | 0.880 | 7.191 | 0.838 | 0.860 | 5.767 |
| | Appina [31] | 0.801 | 0.829 | 9.149 | 0.669 | 0.729 | 7.813 |
| | DNR-S3DIQE [32] | 0.935 | 0.943 | - | 0.871 | 0.863 | - |
| | Fezza [61] | - | - | - | 0.925 | 0.908 | 3.018 |
| | 3D-AdaBoost [16] | 0.930 | 0.939 | 5.605 | 0.913 | 0.922 | 4.352 |
| | DBN [35] | 0.944 | 0.956 | 4.917 | 0.921 | 0.934 | 4.005 |
| | Chen [62] | 0.943 | 0.959 | 4.838 | 0.922 | 0.936 | 3.667 |
| | Sun [40] | 0.959 | 0.951 | 4.573 | 0.918 | 0.938 | 3.809 |
| | DECOSINE [33] | 0.953 | 0.962 | - | 0.941 | 0.950 | - |
| | Zhou [39] | 0.954 | 0.962 | 5.243 | 0.946 | 0.957 | 3.671 |
| | Yang [34] | 0.949 | 0.961 | - | 0.928 | 0.938 | - |
| | StereoQA-Net [36] | 0.965 | 0.973 | 4.711 | 0.947 | 0.957 | 3.270 |
| | Shen [38] | 0.962 | 0.972 | - | 0.951 | 0.953 | - |
| | PAD-Net [37] | 0.973 | 0.975 | 3.514 | 0.967 | 0.975 | 2.446 |
| | Proposed | **0.981** | **0.982** | **3.086** | **0.984** | **0.985** | **1.938** |

TABLE VII: Overall performance comparison on Waterloo-P1 and Waterloo-P2.

| | | Waterloo-P1 | | | Waterloo-P2 | | |
|---|---|---|---|---|---|---|---|
| Type | Metrics | SROCC | PLCC | RMSE | SROCC | PLCC | RMSE |
| FR | Benoit [18] | 0.332 | 0.332 | - | 0.165 | 0.320 | - |
| | Chen [3] | 0.457 | 0.631 | - | 0.272 | 0.442 | - |
| | Ma [22] | 0.911 | 0.925 | 5.876 | - | - | - |
| NR | Fezza [61] | 0.904 | 0.898 | - | 0.890 | 0.866 | - |
| | DECOSINE [33] | 0.924 | 0.943 | - | 0.914 | 0.933 | - |
| | Yang [34] | 0.911 | 0.940 | - | 0.866 | 0.899 | - |
| | Chen [62] | 0.923 | 0.931 | 5.989 | 0.922 | 0.925 | 7.119 |
| | Sun [40] | - | - | - | 0.835 | 0.840 | - |
| | Proposed | **0.967** | **0.973** | **3.592** | **0.977** | **0.981** | **3.617** |

TABLE VIII: Performance of the proposed metric using VGG-16 under different train-test partitions on LIVE-P1 and LIVE-P2.

| | LIVE-P1 | | | LIVE-P2 | | |
|---|---|---|---|---|---|---|
| Partition | SROCC | PLCC | RMSE | SROCC | PLCC | RMSE |
| 80%-20% | 0.981 | 0.982 | 3.086 | 0.984 | 0.985 | 1.938 |
| 70%-30% | 0.980 | 0.980 | 3.189 | 0.982 | 0.983 | 2.061 |
| 50%-50% | 0.976 | 0.977 | 3.432 | 0.977 | 0.978 | 2.327 |

metrics including two recently published methods (i.e. Chen [62] and Sun [40]),the proposed approach again outperforms both NR and FR metrics on both Waterloo datasets.

To exhibit the prediction responses against human score (objective scores predicted by our method vs. subjective scores), we show in Fig. 5 the scatter plots obtained on the four databases. For all datasets, the distribution of the predicted scores is in accordance with the MOS/DMOS for all the considered degradation types.

### *F. Performance on individual distortions*

The overall performance on the four databases has shown good performance and remarkable consistency. Furthermore, the proposed scheme has been examined on individual distortion types. The performance indexes are computed for each distortion individually. Performance in Tables IX, X and XI indicates that the proposed metric predicts perceptual quality well regardless of types of distortion. Overall, the proposed metric delivers stable performance. On FF subsets, the best accuracy in term of PLCC is achieved by PAD-net metric. In term of SROCC on LIVE-P2, the performance of our metric has achieved the state-of-the-art on all distortion subsets. For Waterloo databases, both the PLCC and SROCC indexes are observed to be above 0.9 on the three distortions JPEG, WN,

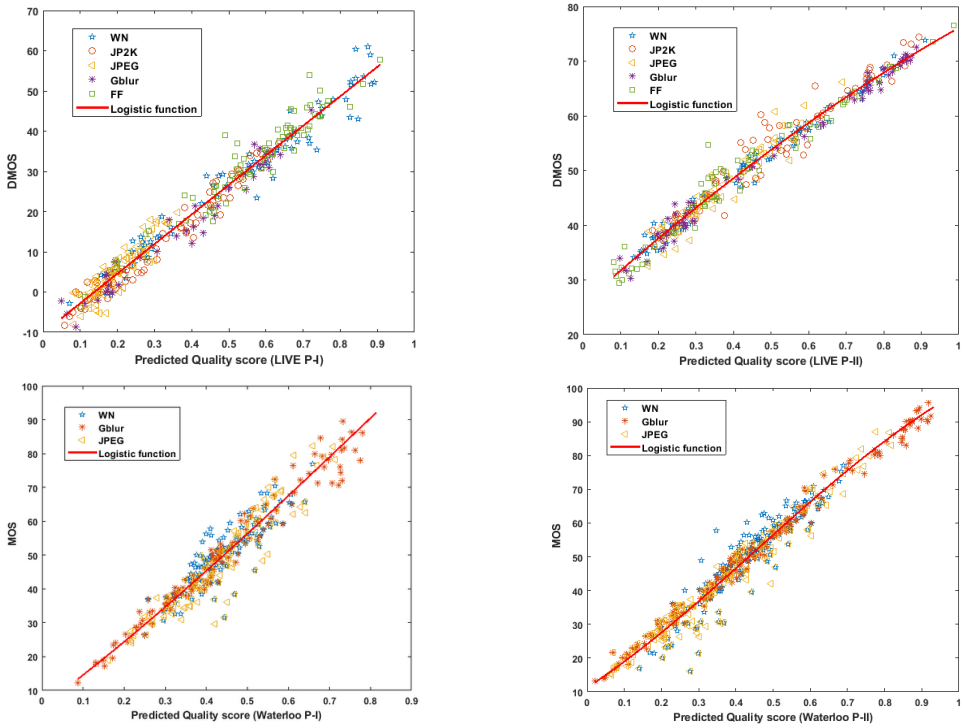



Fig. 5: Scatter plot of subjective scores against objective scores from the proposed metric on the used four databases.

TABLE IX: PLCC results over five types of Distortions using LIVE-P1 and LIVE-P2.

| | | LIVE-P1 | | | | | LIVE-P2 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Type | Metrics | JP2K | JPEG | WN | BLUR | FF | JP2K | JPEG | WN | BLUR | FF |
| FR | Benoit [18] | 0.939 | 0.640 | 0.925 | 0.948 | 0.747 | 0.784 | 0.853 | 0.926 | 0.535 | 0.807 |
| | You [8] | 0.877 | 0.487 | 0.941 | 0.919 | 0.730 | 0.905 | 0.830 | 0.912 | 0.784 | 0.915 |
| | Gorley [5] | 0.485 | 0.312 | 0.796 | 0.852 | 0.364 | 0.372 | 0.322 | 0.874 | 0.934 | 0.706 |
| | Chen [3] | 0.912 | 0.603 | 0.942 | 0.942 | 0.776 | 0.834 | 0.862 | 0.957 | 0.963 | 0.901 |
| | Hewage [19] | 0.904 | 0.530 | 0.895 | 0.798 | 0.669 | 0.664 | 0.734 | 0.891 | 0.450 | 0.746 |
| | Bensalma [4] | 0.838 | 0.838 | 0.914 | 0.838 | 0.733 | 0.666 | 0.857 | 0.943 | 0.907 | 0.909 |
| RR | RR-BPI [24] | - | - | - | - | - | 0.858 | 0.871 | 0.891 | 0.981 | 0.925 |
| | RR-RDCT [25] | 0.918 | 0.722 | 0.913 | 0.925 | 0.807 | 0.897 | 0.748 | 0.810 | 0.969 | 0.910 |
| | Ma [26] | 0.940 | 0.720 | 0.935 | 0.936 | 0.843 | 0.880 | 0.765 | 0.932 | 0.913 | 0.906 |
| NR | Akhter [27] | 0.905 | 0.729 | 0.904 | 0.617 | 0.503 | 0.776 | 0.786 | 0.722 | 0.795 | 0.674 |
| | Fang [30] | 0.911 | 0.547 | 0.900 | 0.903 | 0.718 | 0.740 | 0.764 | 0.961 | 0.968 | 0.867 |
| | DNR-S3DIQE [32] | 0.913 | 0.767 | 0.910 | 0.950 | 0.954 | 0.865 | 0.821 | 0.836 | 0.934 | 0.915 |
| | Fezza [61] | - | - | - | - | - | 0.936 | 0.905 | 0.953 | 0.974 | 0.957 |
| | 3D-AdaBoost [16] | 0.926 | 0.668 | 0.941 | 0.935 | 0.845 | 0.835 | 0.859 | 0.953 | 0.978 | 0.925 |
| | DBN [35] | 0.942 | 0.824 | 0.954 | 0.963 | 0.789 | 0.886 | 0.867 | 0.887 | 0.988 | 0.916 |
| | PAD-Net [37] | 0.982 | **0.919** | 0.978 | 0.985 | **0.994** | **0.981** | 0.898 | 0.973 | **0.997** | **0.986** |
| | Proposed | **0.986** | 0.906 | **0.979** | **0.986** | 0.963 | 0.969 | **0.964** | **0.992** | **0.997** | 0.982 |

and BLUR. The highest score has been reached on BLUR distortion. From the used Waterloo and LIVE databases, the metric has reached it highest performance on BLUR. This is also observed in other metrics scores. Usually, the BLUR distortions are easy to detect and they are compared to other forms of distortion such as JPEG one. In the proposed model, the well tuned convolutional layers have given a step further to capture this distortion. On BLUR's distortion over the four datasets, the accuracy of quality assessment was found to be 98% in terms of PLCC.

Table XII shows the performance of our metric on symmetric and asymmetric distorted stimuli. As can be seen, some metrics totally fail to predict the quality for asymmetric distorted images. They give high correlations for symmetric distorted images (Benoit, You and Bensalma). PAD-Net yields the best performance for symmetric distorted images. The first and the second best correlations for symmetric and asymmetric distorted images have been produced by the proposed approach. According to the Table VI; our metric achieves the best global results. Moreover, high accuracy on asymmetric distortions is more challenging, since most of the existing methods fail.

### *G. Cross database performance*

Cross-database experiments have been conducted in order to verify the generalization ability of the proposed approach. The

TABLE X: SROCC results over five types of distortions using LIVE-P1 and LIVE-P2.

| | | LIVE-P1 | | | | | LIVE-P2 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Type | Metrics | JP2K | JPEG | WN | BLUR | FF | JP2K | JPEG | WN | BLUR | FF |
| FR | Benoit [18] | 0.910 | 0.603 | 0.930 | 0.931 | 0.699 | 0.751 | 0.867 | 0.923 | 0.455 | 0.773 |
| | You [8] | 0.860 | 0.439 | 0.940 | 0.882 | 0.588 | 0.894 | 0.795 | 0.909 | 0.813 | 0.891 |
| | Gorley [5] | 0.015 | 0.569 | 0.741 | 0.750 | 0.366 | 0.110 | 0.027 | 0.875 | 0.770 | 0.601 |
| | Chen [3] | 0.888 | 0.530 | 0.948 | 0.925 | 0.707 | 0.814 | 0.843 | 0.940 | 0.908 | 0.884 |
| | Hewage [19] | 0.856 | 0.500 | 0.940 | 0.690 | 0.545 | 0.598 | 0.736 | 0.880 | 0.028 | 0.684 |
| | Bensalma [4] | 0.817 | 0.328 | 0.905 | 0.915 | 0.915 | 0.803 | 0.846 | 0.938 | 0.846 | 0.846 |
| RR | RR-BPI [24] | - | - | - | - | - | 0.776 | 0.736 | 0.904 | 0.871 | 0.854 |
| | RR-RDCT [25] | 0.887 | 0.616 | 0.912 | 0.879 | 0.696 | 0.879 | 0.737 | 0.732 | 0.876 | 0.895 |
| | Ma [26] | 0.907 | 0.660 | 0.928 | 0.921 | 0.792 | 0.868 | 0.791 | 0.954 | 0.923 | 0.944 |
| NR | Akhter [27] | 0.866 | 0.675 | 0.914 | 0.555 | 0.640 | 0.724 | 0.649 | 0.714 | 0.682 | 0.559 |
| | Zhou [29] | 0.856 | 0.562 | 0.921 | 0.897 | 0.771 | 0.647 | 0.737 | 0.936 | 0.911 | 0.798 |
| | Fang [30] | 0.880 | 0.523 | 0.883 | 0.523 | 0.650 | 0.714 | 0.709 | 0.955 | 0.807 | 0.872 |
| | DNR-S3DIQE [32] | 0.885 | 0.765 | 0.921 | 0.930 | 0.944 | 0.853 | 0.822 | 0.833 | 0.889 | 0.878 |
| | Fezza [61] | - | - | - | - | - | 0.927 | 0.886 | 0.947 | 0.928 | 0.952 |
| | 3D-AdaBoost [16] | 0.899 | 0.625 | 0.941 | 0.887 | 0.777 | 0.842 | 0.837 | 0.943 | 0.913 | 0.925 |
| | DBN [35] | 0.897 | 0.768 | 0.929 | 0.917 | 0.685 | 0.859 | 0.806 | 0.864 | 0.834 | 0.877 |
| | PAD-Net [37] | 0.969 | 0.889 | 0.968 | 0.917 | **0.996** | 0.959 | 0.882 | 0.962 | 0.867 | 0.945 |
| | Proposed | **0.975** | **0.906** | **0.978** | **0.967** | 0.950 | **0.963** | **0.957** | **0.988** | **0.983** | **0.972** |

TABLE XI: Performance comparison of the proposed metric on individual distortions using Waterloo-P1 and Waterloo-P2 database.

| | Waterloo-P1 | | | Waterloo-P2 | | |
|---|---|---|---|---|---|---|
| Distortion type | SROCC | PLCC | RMSE | SROCC | PLCC | RMSE |
| JPEG | 0.951 | 0.954 | 4.084 | 0.968 | 0.970 | 4.075 |
| WN | 0.915 | 0.916 | 3.756 | 0.940 | 0.941 | 4.178 |
| BLUR | 0.985 | 0.987 | 2.715 | 0.988 | 0.995 | 2.017 |

TABLE XII: SROCC performance for symmetric and asymmetric distorted images on LIVE-P2. Best result of each category is highlighted in bold.

| | | LIVE-P2 | | Waterloo-P1 | | Waterloo-P2 | |
|---|---|---|---|---|---|---|---|
| Method | Type | Symmetric | Asymmetric | Symmetric | Asymmetric | Symmetric | Asymmetric |
| Benoit [18] | FR | 0.860 | 0.671 | - | - | - | - |
| You [8] | | 0.914 | 0.701 | 0.752 | 0.571 | - | - |
| Gorley [5] | | 0.383 | 0.056 | 0.566 | 0.475 | - | - |
| Chen [3] | | 0.923 | 0.842 | 0.924 | 0.643 | - | - |
| Hewage [19] | | 0.656 | 0.496 | - | - | - | - |
| Bensalma [4] | | 0.841 | 0.721 | - | - | - | - |
| Akhter [27] | NR | 0.420 | 0.517 | - | - | - | - |
| Fezza [61] | | 0.928 | 0.882 | 0.902 | 0.869 | 0.915 | 0.804 |
| 3D-AdaBoost [16] | | 0.898 | 0.917 | - | - | - | - |
| PAD-Net [37] | | **0.982** | 0.954 | 0.985 | **0.978** | - | - |
| Proposed | | 0.973 | **0.987** | **0.987** | 0.967 | **0.987** | **0.976** |

implemented tests are shown in Table XIII. Metrics shown are all NR methods. They have been trained in the former database and tested on the latter.

Comparing with the NR metrics, our method has competitive prediction about the quality of stereo pairs despite cross-database tests. DECOSINE, Sun and PAD-net algorithms deliver decent performance in the four cross-database tests, but Sun is the only algorithm which gives performance over 0.9 in term of PLCC in the L1/L2 test. From LIVE datasets, the performance of the other NR algorithms is not as good as the performance of the individual database tests. For instance, Chen and DBN metrics showed good results on the individual database tests where Pearson correlations (PLCCs) of 0.959 and 0.956 have been achieved on LIVE P-1 for Chen and DBN, respectively. They gave low performance scores in the L1/L2 test. PLCC of 0.869 and 0.827 are reported for Chen and DBN respectively. Waterloo datasets have shown lower correlations than LIVE datasets. It is important to notice that which makes Waterloo databases more challenging than LIVE is that they not only include both symmetric and asymmetric distorted pairs like LIVE phase-II. Also, the left and right views of a stereo pair may be distorted by different distortion types. The cross-database tests revealed that the proposed approach ranks third after the two metrics DECOSINE and PAD-net. However on LIVE datasets, the correlation gaps are not profound, 0.003 and 0.005 are the difference values of our metric with PAD-net and DECOSINE respectively.

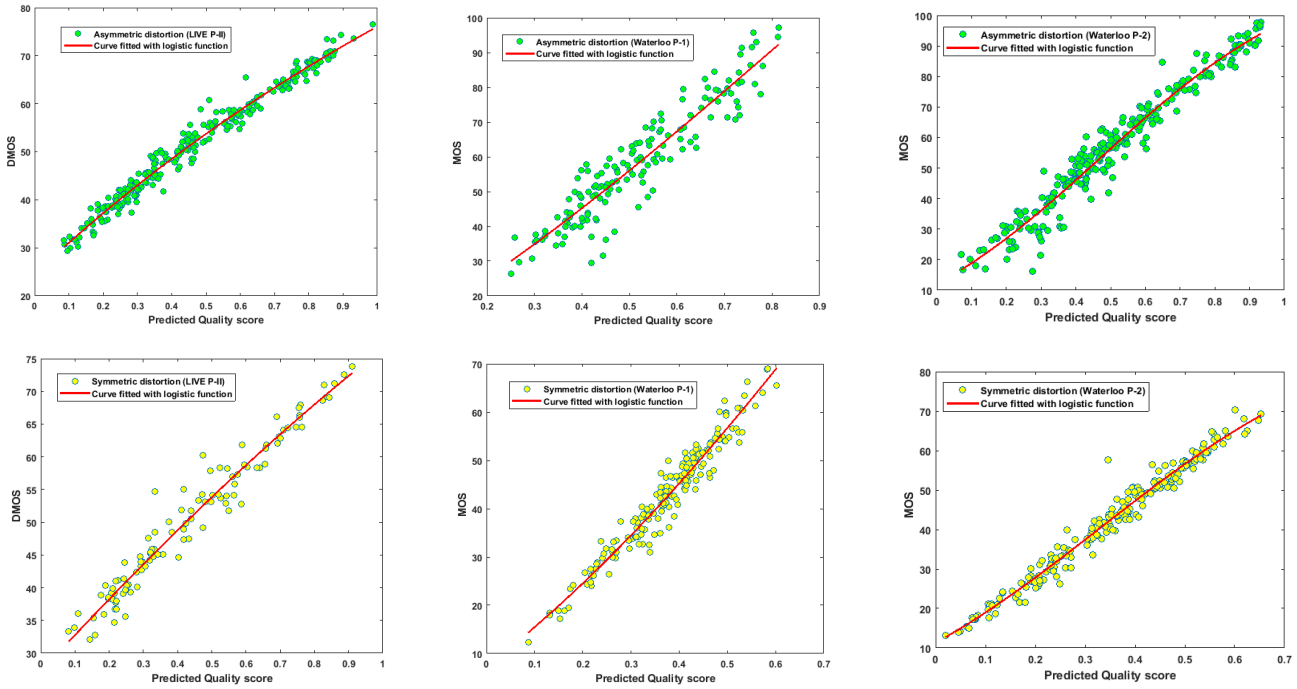



Fig. 6: Asymmetric and symmetric distortion plots from the four databases using the proposed method.

TABLE XIII: PLCC Performance of cross database tests using the four databases. (Expressed as: Train database/Test database.)

| Metrics | L-P2/L-P1 | L-P1/L-P2 | W-P1/W-P2 | W-P2/W-P1 |
|---|---|---|---|---|
| DBN [35] | 0.869 | 0.852 | - | - |
| DECOSINE [33] | **0.916** | 0.846 | **0.842** | **0.873** |
| 3D-AdaBoost [16] | 0.892 | 0.824 | - | - |
| Chen [62] | 0.827 | 0.812 | 0.806 | 0.846 |
| Sun [40] | 0.899 | **0.919** | - | - |
| Yang [34] | 0.860 | 0.861 | 0.781 | 0.864 |
| Shen [38] | 0.915 | 0.848 | - | - |
| PAD-Net [37] | 0.915 | 0.854 | - | - |
| Proposed | 0.911 | 0.851 | 0.826 | 0.848 |

### *H. Statistical test performance*

In order to verify whether our proposed model is statistically better than other metrics. We conducted the T-test against the state-of-the-art metrics with confidence at 90% applied over 10 trials for PLCC and SROCC. This test is one of numerous statistical tests [63]. It questions whether the difference between the groups represents a true difference in the study or if it is more likely a meaningless statistical difference. The results is statistically superior or worse than the competitive metric in the column, respectively. The value of 1 indicates the superiority of the proposed method, and -1 indicates the opposite. While 0 means that the two metrics are statistically similar. From the tabulated results, we notice that our metric performs statistically better than other NR-SIQA metrics both on LIVE Phase I and II.

TABLE XIV: T-test results with confidence of 90% of the proposed metric against the others using PLCC, SROCC on LIVE I and II

| Database | Index | 3D-AdaBoost [16] | Chen [62] | Shen [38] | PAD-Net [37] |
|---|---|---|---|---|---|
| LIVE I | PLCC | 1 | 1 | 1 | 1 |
| | SROCC | 1 | 1 | 1 | 1 |
| LIVE II | PLCC | 1 | 1 | 1 | 1 |
| | SROCC | 1 | 1 | 1 | 1 |

### *I. Computational complexity*

We compare here computational time with the most recent NR-SIQA metrics that incorporate deep learning into their designs.The working platform uses the MATLAB2020a on a computer equipped with Intel(R) Xeon(R) CPU E5-2620 v4 processor at 2.10GHz, 64GB of memory and a NVIDIA Quadro P5000 GPU - 16GB of memory. It should be noted that the other approaches have been tested on various hardware. The test was performed on a stereo image from the LIVE phase II database with a resolution of 640 x 360 pixels.

The run time (in seconds) tests are listed in Table XV. It is worth noting that for our model we record the time around 17 seconds for predicting quality score. The results show that PAD-Net [37] only needs around 1 second per image which is

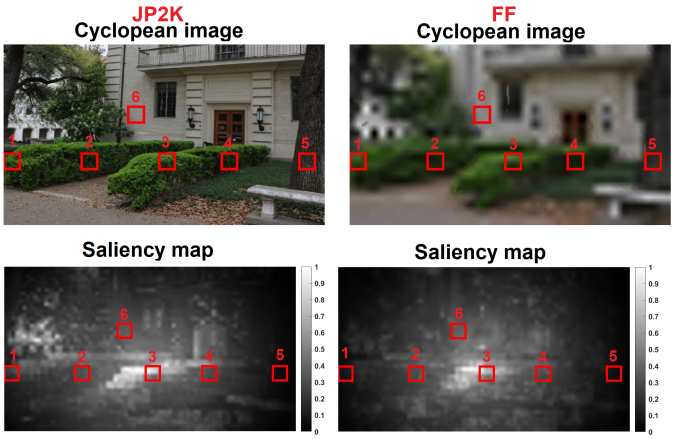



Fig. 7: Examples of synthesized cyclopean image and 3D saliency map on two different types of distortion.

significantly lower than other metrics, while metrics in [38], [36] require around 9 and 3 seconds, respectively, to deliver quality ratings. In our approach, the most computationally expensive stage is the cyclopean image construction, since it involves weights computation of the left and right views by performing a multi-scale Gabor filter. Note that the metric in [34] also includes a cyclopean image computation, where this metric records higher run time, around 20 seconds. Therefore, we can observe that metrics which do not require considerable pre-processing, such as cyclopean image computation, are more likely to be faster than others because they mostly use the stereoscopic image directly as input.

### *J. Influence of distortions on the 3D saliency map*

To investigate the impact of the distortions on the computed 3D saliency map from the cyclopean image, we observe the 3D saliency map generated over two different types of distortion, namely JP2K and FF. The cyclopean image is also being spotted on these distortions. Fig. 7 displays the computation outputs. As can be seen, in each of the synthesized cyclopean image, the quality deformation is clearly stated. It depends on the type of distortion. Meanwhile, the computed 3D saliency maps are very similar despite the variation of distortion. This latter indicates consistency against the degradations that occur in the stereoscopic images. Furthermore, relationship of the saliency value and the error quality prediction are studied. Six patches of the same locations have been selected from each cyclopean and 3D saliency maps as shown in Fig. 7. Quality prediction error of each patch $P_e$ and its saliency average $P_s$ computations are as follows:

$$P_e = abs(y - \hat{y}) \tag{6}$$

where $y$ is the ground truth quality and $\hat{y}$ is the predicted quality using the proposed metric. The patch saliency average $S_a$ is defined by:

$$S_a = \frac{1}{m.n} \sum_{i=1}^{m} \sum_{j=1}^{n} M(i,j) \tag{7}$$

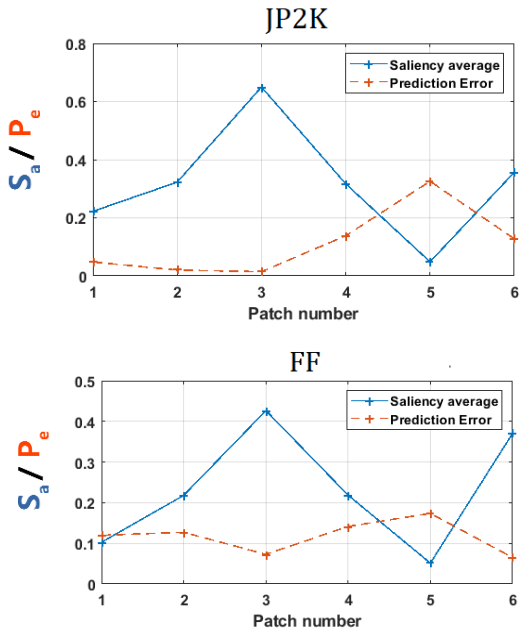



Fig. 8: Saliency patch average versus quality prediction error for patches from 1 to 6 under JP2K, and FF distortion shown in Fig. 7.

where $M$ is the computed 3D saliency map from previous steps.

Fig. 8 shows the obtained curves. On both distortions, it is remarkable to observe the changes of prediction error derived by the saliency. Curves show that the prediction error drops when the saliency patch average increases and vice-versa. In the case of JP2K distortion, patch number three shows that the highest saliency (0.63) is visible at lowest quality prediction error of values (0.004). For FF distortion with the same patch, we note the lowest error (0.068) at the highest saliency value (0.42). Generally, for saliency values above the 0.3 threshold, we find consistency quality prediction errors below 0.15. From these findings, we conclude that the human visual selectivity influences the quality evaluation. This quality evaluation can be improved by saliency information for objective methods.

### *K. Feature map visualization*

In this section, we take a look at what deep CNN sees from degraded images. We also analyzed the learned convolutional filters and their activation functions that yield feature maps. We examine which parts of the cyclopean image are most important for our CNN models. To ensure independence output, we have preferred the model trained on LIVE-P2 to observe its behavior on new cyclopean images from LIVE-P1. The test cyclopean images are shown in Fig. 7, where only the patch number six is fed to the network. The synthesized cyclopean views were formed under different types of distortion: JP2K, WN and FF. The patches are fed to the CNN and then inspect the outputs of activation functions (ReLU) after the first and second convolutional layers. The first two convolution layers produce 64 channels each. Among the 64 channels output

TABLE XV: The computation time comparison using NVIDIA P4000 GPU for the proposed method.

| Metrics | Shen [38] | StereoQA-Net [36] | PAD-Net [37] | Yang [34] | Proposed |
|---|---|---|---|---|---|
| Time (sec.) | 8.822 | 2.377 | **0.906** | 19.882 | 16.335 |

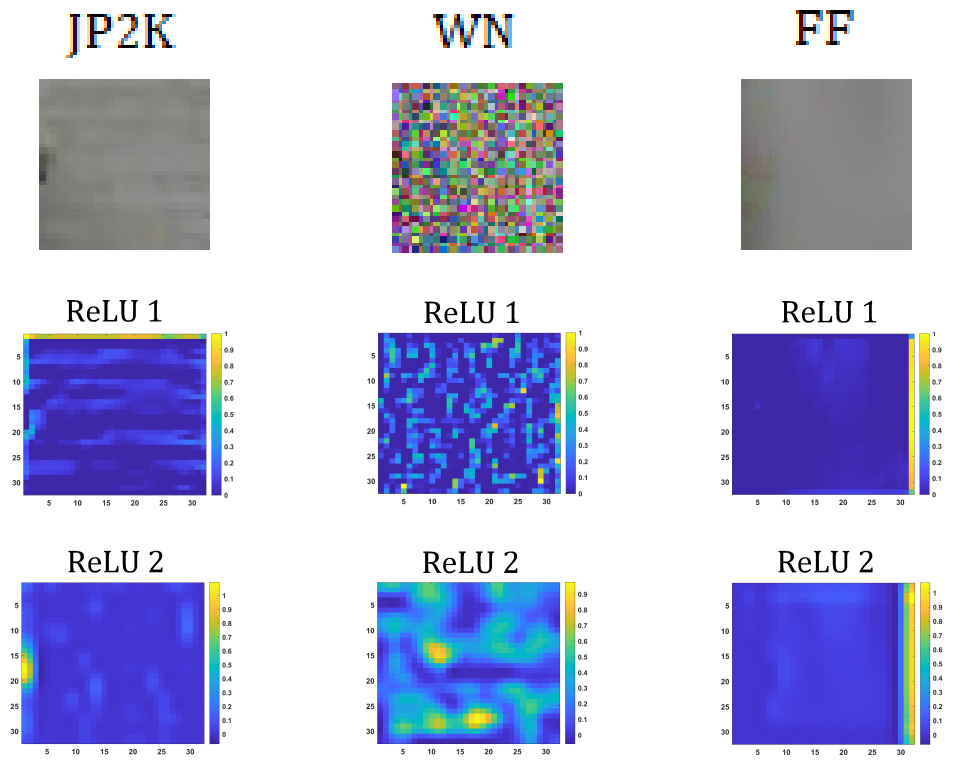



Fig. 9: The first and second feature map (ReLU activation layer) outputs from a test cyclopean patch for three degradation types.

from ReLU layer, their mean values are computed and the strongest channel has been selected by indexing the maximum. Fig. 9 despites the first and second ReLU layer responses for the input cyclopean patch. As can be seen, where the warmer (closer to 1) regions activate the ReLU function and thus influence the decision of the network. It is remarkable that the first activation function reflects the presence of pixel deformation. The JP2K compression is well known artifact that causes undesirable blocks in the image due to the quantization. This issue is stated in ReLU 1 activation map of JP2K patch that shows the selection of these blocks as a highly important information to pass through the network. As well as for WN and FF cyclopean patches, the ReLU 1 activation function has succeeded to focus on noise and blur artifacts.

While the second activation function (ReLU layer) is controlled by a deeper representation that makes it harder to fully comprehend the outputs. However, for JP2K cyclopean patch, deformed regions cover most of the patch that captures peace of house wall on the scene. For WN, the deformed regions are located around everywhere the wall. From the second ReLU output maps, the warmer regions are somewhat distributed according to the most infected regions in the scene. Meanwhile for FF patch, the spatial information of the wall is less effected since FF is considered as high frequency distortion. Interestingly, the ReLU 2 responses show that the degradation covers the entire wall, which is often the case for FF degradation. It is worth noting that the activation functions for each patch differ as the type of degradation differs.

Overall, we can see how the model learns to focus on pixel deformations in order to extract complex quality indicators. As a result, the model can distinguish between various types of distortion. Based on our findings, we conclude that the deep network retrieves high-quality features that are influenced by the form and degree of distortion.

## V. CONCLUSION

In this paper, a no-reference stereoscopic IQA based on the use of cyclopean image and saliency map has been proposed. The simplicity of the proposed scheme is a benefit for an easy implementation in the multimedia software. Cyclopean image has been introduced to consider asymmetrical distortion, while the saliency aims to focus on the most perceptual relevant regions by selecting relevant patches from the cyclopean image. These patches are then fed as input to a modified version of a pre-trained CNN model to estimate the quality. We compared five pre-trained models (i.e. AlexNet, VGG16, VGG19, ResNet18 and resent50) and we also show the impact of the saliency selection. The best performance has been

obtained with VGG16 for a saliency threshold equals to 0.3. Experimental results have demonstrate the efficiency of the proposed metric since it outperforms all the compared FR and NR SIQA of the state-of-the-art on LIVE and Waterloo databases. Also, the capacity of our method to predict the quality of unknown stereo images has been evaluated.

As future work, by incorporating an adaptive automatic adjustment for saliency threshold and patch size, the quality prediction can be further enhanced.

## REFERENCES


[1] Jose Weber Vieira de Faria, Manoel Jacobsen Teixeira, Leonardo de Moura Sousa Júnior, Jose Pinhata Otoch, and Eberval Gadelha Figueiredo, "Virtual and stereoscopic anatomy: when virtual reality meets medical education," *Journal of neurosurgery*, vol. 125, no. 5, pp. 1105–1111, 2016.

[2] Hanz Cuevas-Velasquez, Antonio-Javier Gallego, and Robert B Fisher, "Segmentation and 3d reconstruction of rose plants from stereoscopic images," *Computers and electronics in agriculture*, vol. 171, pp. 105296, 2020.

[3] M. Chen, Che-Chun Su, Do-Kyoung Kwon, Lawrence K Cormack, and Alan C Bovik, "Full-reference quality assessment of stereopairs accounting for rivalry," *Signal Processing: Image Communication*, vol. 28, no. 9, pp. 1143–1155, 2013.

[4] R. Bensalma and Mohamed-Chaker Larabi, "A perceptual metric for stereoscopic image quality assessment based on the binocular energy," *Multidimensional Systems and Signal Processing*, vol. 24, no. 2, pp. 281–316, 2013.

[5] P. Gorley and Nick Holliman, "Stereoscopic image quality metrics and compression," in *Electronic Imaging 2008*. International Society for Optics and Photonics, 2008, pp. 680305–680305.

[6] David G Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*. Ieee, 1999, vol. 2, pp. 1150–1157.

[7] Martin A Fischler and Robert C Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," in *Readings in computer vision*, pp. 726–740. Elsevier, 1987.

[8] J. You, Liyuan Xing, Andrew Perkis, and Xu Wang, "Perceptual quality assessment for stereoscopic images based on 2d image quality metrics and disparity analysis," in *Proc. of International Workshop on Video Processing and Quality Metrics for Consumer Electronics, Scottsdale, AZ, USA*, 2010.

[9] Anush Krishna Moorthy and Alan Conrad Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, 2011.

[10] Ke Gu, Leida Li, Hong Lu, Xiongkuo Min, and Weisi Lin, "A fast reliable image quality predictor by fusing micro-and macro-structures," *IEEE Transactions on Industrial Electronics*, vol. 64, no. 5, pp. 3903–3912, 2017.

[11] Ke Gu, Weisi Lin, Guangtao Zhai, Xiaokang Yang, Wenjun Zhang, and Chang Wen Chen, "No-reference quality metric of contrast-distorted images based on information maximization," *IEEE transactions on cybernetics*, vol. 47, no. 12, pp. 4559–4565, 2016.

[12] Aladine Chetouani and Leida Li, "On the use of a scanpath predictor and convolutional neural network for blind image quality assessment," *Signal Processing: Image Communication*, vol. 89, pp. 115963, 2020.

[13] Aladine Chetouani, Maurice Quach, Giuseppe Valenzise, and Frédéric Dufaux, "Combination of deep learning-based and handcrafted features for blind image quality assessment," in *9th European Workshop on Visual Information Processing (EUVIP 2021)*, 2021.

[14] R. Blake, David H Westendorf, and Randall Overton, "What is suppressed during binocular rivalry?," *Perception*, vol. 9, no. 2, pp. 223–231, 1980.

[15] Aladine Chetouani, "Full reference image quality metric for stereo images based on cyclopean image computation and neural fusion," in *2014 IEEE Visual Communications and Image Processing Conference*. IEEE, 2014, pp. 109–112.

[16] Oussama Messai, Fella Hachouf, and Zianou Ahmed Seghir, "Adaboost neural network and cyclopean view for no-reference stereoscopic image quality assessment," *Signal Processing: Image Communication*, p. 115772, 2020.

[17] A. Chetouani, "Full reference image quality metric for stereo images based on cyclopean image computation and neural fusion," in *2014 IEEE Visual Communications and Image Processing Conference*, 2014, pp. 109–112.

[18] Alexandre Benoit, Patrick Le Callet, Patrizio Campisi, and Romain Cousseau, "Quality assessment of stereoscopic images," *EURASIP journal on image and video processing*, vol. 2008, no. 1, pp. 1–13, 2009.

[19] CTER Hewage, Stewart T Worrall, Safak Dogan, and AM Kondoz, "Prediction of stereoscopic video quality using objective quality models of 2-d video," *Electronics letters*, vol. 44, no. 16, pp. 963–965, 2008.

[20] Xianqiu Geng, Liquan Shen, Kai Li, and Ping An, "A stereoscopic image quality assessment model based on independent component analysis and binocular fusion property," *Signal Processing: Image Communication*, vol. 52, pp. 54–63, 2017.

[21] Feng Shao, Kemeng Li, Weisi Lin, Gangyi Jiang, Mei Yu, and Qionghai Dai, "Full-reference quality assessment of stereoscopic images by learning binocular receptive field properties," *IEEE Transactions on Image Processing*, vol. 24, no. 10, pp. 2971–2983, 2015.

[22] Jian Ma, Ping An, Liquan Shen, and Kai Li, "Full-reference quality assessment of stereoscopic images by learning binocular visual properties," *Applied optics*, vol. 56, no. 29, pp. 8291–8302, 2017.

[23] Jianwei Si, Huan Yang, Baoxiang Huang, Zhenkuan Pan, and Honglei Su, "A full-reference stereoscopic image quality assessment index based on stable aggregation of monocular and binocular visual features," *IET Image Processing*, 2021.

[24] Feng Qi, Debin Zhao, and Wen Gao, "Reduced reference stereoscopic image quality assessment based on binocular perceptual information," *IEEE Transactions on multimedia*, vol. 17, no. 12, pp. 2338–2344, 2015.

[25] Lin Ma, Xu Wang, Qiong Liu, and King Ngi Ngan, "Reorganized dct-based image representation for reduced reference stereoscopic image quality assessment," *Neurocomputing*, vol. 215, pp. 21–31, 2016.

[26] Jian Ma, Ping An, Liquan Shen, and Kai Li, "Reduced-reference stereoscopic image quality assessment using natural scene statistics and structural degradation," *IEEE Access*, vol. 6, pp. 2768–2780, 2017.

[27] R. Akhter, ZM Parvez Sazzad, Yuukou Horita, and Jacky Baltes, "No-reference stereoscopic image quality assessment," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2010, pp. 75240T–75240T.

[28] Aladine Chetouani, "Toward a universal stereoscopic image quality metric without reference," in *Advanced Concepts for Intelligent Vision Systems*, Cham, 2015, pp. 604–612, Springer International Publishing.

[29] Wujie Zhou, Weiwei Qiu, and Ming-Wei Wu, "Utilizing dictionary learning and machine learning for blind quality assessment of 3-d images," *IEEE Transactions on Broadcasting*, vol. 63, no. 2, pp. 404–415, 2017.

[30] Meixin Fang and Wujie Zhou, "Toward an unsupervised blind stereoscopic 3d image quality assessment using joint spatial and frequency representations," *AEU-International Journal of Electronics and Communications*, vol. 94, pp. 303–310, 2018.

[31] Balasubramanyam Appina, "A 'complete blind'no-reference stereoscopic image quality assessment algorithm," in *2020 International Conference on Signal Processing and Communications (SPCOM)*. IEEE, 2020, pp. 1–5.

[32] Heeseok Oh, Sewoong Ahn, Jongyoo Kim, and Sanghoon Lee, "Blind deep s3d image quality evaluation via local to global feature aggregation," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4923–4936, 2017.

[33] Jiachen Yang, Kyohoon Sim, Xinbo Gao, Wen Lu, Qinggang Meng, and Baihua Li, "A blind stereoscopic image quality evaluator with segmented stacked autoencoders considering the whole visual perception route," *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1314–1328, 2018.

[34] Jiachen Yang, Kyohoon Sim, Wen Lu, and Bin Jiang, "Predicting stereoscopic image quality via stacked auto-encoders based on stereopsis formation," *IEEE Transactions on Multimedia*, vol. 21, no. 7, pp. 1750–1761, 2018.

[35] Jiachen Yang, Yang Zhao, Yinghao Zhu, Huifang Xu, Wen Lu, and Qinggang 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.

[36] Wei Zhou, Zhibo Chen, and Weiping Li, "Dual-stream interactive networks for no-reference stereoscopic image quality assessment," *IEEE Transactions on Image Processing*, vol. 28, no. 8, pp. 3946–3958, 2019.

[37] Jiahua Xu, Wei Zhou, Zhibo Chen, Suiyi Ling, and Patrick Le Callet, "Predictive auto-encoding network for blind stereoscopic image quality assessment," *arXiv preprint arXiv:1909.01738*, 2019.

[38] Lili Shen, Xiongfei Chen, Zhaoqing Pan, Kefeng Fan, Fei Li, and Jianjun Lei, "No-reference stereoscopic image quality assessment based on global and local content characteristics," *Neurocomputing*, vol. 424, pp. 132–142, 2021.

[39] Wujie Zhou, Jingsheng Lei, Qiuping Jiang, Lu Yu, and Ting Luo, "Blind binocular visual quality predictor using deep fusion network," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 883–893, 2020.

[40] Guangming Sun, Bufan Shi, Xiaodong Chen, Andrey S Krylov, and Yong Ding, "Learning local quality-aware structures of salient regions for stereoscopic images via deep neural networks," *IEEE Transactions on Multimedia*, 2020.

[41] Yun Liu, Chang Tang, Zhi Zheng, and Liyuan Lin, "No-reference stereoscopic image quality evaluator with segmented monocular features and perceptual binocular features," *Neurocomputing*, vol. 405, pp. 126–137, 2020.

[42] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[43] T. Bertel, M. Mühlhausen, M. Kappel, P. M. Bittner, C. Richardt, and M. Magnor, "Depth augmented omnidirectional stereo for 6-dof vr photography," in *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 2020, pp. 660–661.

[44] A. Mukherjee, S. Sarkar, and S. K. Saha, "Object mapping from disparity map by fast clustering," in *2020 IEEE Calcutta Conference (CALCON)*, 2020, pp. 74–79.

[45] C. Zhou, Y. Liu, P. Lasang, and Q. Sun, "Vehicle detection and disparity estimation using blended stereo images," *IEEE Transactions on Intelligent Vehicles*, pp. 1–1, 2021.

[46] Jiheng Wang, Shiqi Wang, Kede Ma, and Zhou Wang, "Perceptual depth quality in distorted stereoscopic images," *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1202–1215, 2016.

[47] Karsten Mühlmann, Dennis Maier, Jürgen Hesser, and Reinhard Männer, "Calculating dense disparity maps from color stereo images, an efficient implementation," *International Journal of Computer Vision*, vol. 47, no. 1-3, pp. 79–88, 2002.

[48] Z. Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[49] John G Daugman, "Two-dimensional spectral analysis of cortical receptive field profiles," *Vision research*, vol. 20, no. 10, pp. 847–856, 1980.

[50] Wujie Zhou, Junwei Wu, Jingsheng Lei, Jeng-Neng Hwang, and Lu Yu, "Salient object detection in stereoscopic 3d images using a deep convolutional residual autoencoder," *IEEE Transactions on Multimedia*, 2020.

[51] Junle Wang, Matthieu Perreira Da Silva, Patrick Le Callet, and Vincent Ricordel, "Computational model of stereoscopic 3d visual saliency," *IEEE Transactions on Image Processing*, vol. 22, no. 6, pp. 2151–2165, 2013.

[52] Yuming Fang, Zhenzhong Chen, Weisi Lin, and Chia-Wen Lin, "Saliency detection in the compressed domain for adaptive image retargeting," *IEEE Transactions on Image Processing*, vol. 21, no. 9, pp. 3888–3901, 2012.

[53] A. Krizhevsky, "One weird trick for parallelizing convolutional neural networks," *CoRR*, vol. abs/1404.5997, 2014.

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

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[56] A.K. Moorthy, Che-Chun Su, Anish Mittal, and Alan Conrad Bovik, "Subjective evaluation of stereoscopic image quality," *Signal Processing: Image Communication*, vol. 28, no. 8, pp. 870–883, 2013.

[57] Jiheng Wang, Kai Zeng, and Zhou Wang, "Quality prediction of asymmetrically distorted stereoscopic images from single views," in *2014 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2014, pp. 1–6.

[58] Jiheng Wang, Abdul Rehman, Kai Zeng, Shiqi Wang, and Zhou Wang, "Quality prediction of asymmetrically distorted stereoscopic 3d images," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3400–3414, 2015.

[59] Kjell Brunnstrom, David Hands, Filippo Speranza, and Arthur Webster, "Vqeg validation and itu standardization of objective perceptual video quality metrics [standards in a nutshell]," *IEEE Signal processing magazine*, vol. 26, no. 3, pp. 96–101, 2009.

[60] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on image processing*, vol. 15, no. 11, pp. 3440–3451, 2006.

[61] Sid Ahmed Fezza, Aladine Chetouani, and Mohamed-Chaker Larabi, "Using distortion and asymmetry determination for blind stereoscopic image quality assessment strategy," *Journal of Visual Communication and Image Representation*, vol. 49, pp. 115–128, 2017.

[62] Yong Chen, Kaixin Zhu, and Liu Huanlin, "Blind stereo image quality assessment based on binocular visual characteristics and depth perception," *IEEE Access*, vol. 8, pp. 85760–85771, 2020.

[63] ITUT Rec, "P. 1401, methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models," *International Telecommunication Union, Geneva, Switzerland*, 2012.