Perceptually-Weighted Cnn For 360-Degree Image Quality Assessment Using Visual Scan-Path And Jnd

Abderrezzaq Sendjasni, Mohamed-Chaker Larabi, Faouzi Alaya Cheikh

PERCEPTUALLY-WEIGHTED CNN FOR 360-DEGREE IMAGE QUALITY ASSESSMENT USING VISUAL SCAN-PATH AND JND

Abderrezzaq Sendjasni [1,2]*, Mohamed-Chaker Larabi* [1] *and Faouzi Alaya Cheikh* [2]

[1] CNRS, Univ. Poitiers, XLIM, UMR 7252, France
[2] NTNU, Norwegian Colour and Visual Computing Lab, Gjøvik, Norway

ABSTRACT

Image quality assessment of immersive content and more specifically 360-degree one is still in its infancy. There are many challenges regarding sphere vs. projected representation, human visual system (HVS) properties in a 360-degree environment, etc. In this paper, we propose the use of CNNs to design a no reference model to predict visual quality of 360-degree images. Instead of feeding the CNN with ERPs, visually important viewports are extracted based on visual scan-path prediction and given to a multi-channel CNN using DenseNet-121. Moreover, information about visual fixations and just noticeable difference are used to account for the HVS properties and make the network closer to human judgment. The scan-path is also used to create multiple instances of the database so as to perform a robust generalization analysis and compensate for the lack of databases.

***Index Terms*—** 360-degree images, CNNs, scan-path, JND, blind image quality assessment.

1. INTRODUCTION

Nowadays, immersive technologies are used in many fields, including healthcare, education, and gaming for instance. The ability given to the user to look in any direction is offered by the head-mounted displays (HMDs) using omnidirectional content such as 360-degree images. Two types of impairments can decrease the quality of such content. Those linked to the processing pipeline such as stitching, compression and, transmission, in addition to the ones related to the display device such as the screen door. Together with other factors, these impairments may cause motion- and cyber-sickness that can alter the quality of experience (QoE) of users [1]. Therefore, improving the QoE is crucial for immersive applications. To do so, it is important to study and provide appropriate visual quality assessment approaches.

Image quality assessment (IQA) can be addressed subjectively and objectively. The former remains the most reliable way to evaluate image/video quality while being tedious and time-consuming. Therefore, the latter ensures a trade-off by providing a computational approach for such a task. It combines visual features with the aim to reflect the perceptual quality of components represented by the mean opinion scores (MOS) obtained from subjective experiments.

With the introduction of 360-degree images, a few IQA models have been proposed by extending traditional 2D models such as PSNR or MSE. For example, PSNR-based methods like Spherical PSNR (S-PSNR) [2] which computes the PSNR on a spherical surface instead of the 2D representation. The weighted spherical PSNR (WS-PSNR) [3] uses the scaling factor from a 2D plane to the sphere as a weighting factor for PSNR computation. CPP-PSNR [4] computes PSNR on the craster parabolic projection (CPP) after remapping pixels of the original and distorted images from the spherical domain to CPP. As these models do not account for perceptual aspects, they fail in predicting the visual quality accurately. Besides, well-performing 2D metrics are not suitable for 360-degree images as they neither account for spherical characteristics nor for the specific exploration of the scene made by observers [5]. These limitations push towards the design of specific IQA models accounting for the perceptual peculiarities of 360-degree images.

On another side, the interest for convolutional neural networks (CNNs) for quality assessment tasks is fastly growing. This is mainly due to its architecture, which is capable to extract discriminating features at various levels of abstraction [6, 7], *i.e.* from low-level to high-level features. CNNs are involved in various image processing tasks, such as image segmentation, object detection and, image classification. The inherited models are often exploited to regress the quality scores by means of transfer learning and/or by learning HVS-based features [8, 9].

CNN-based models dedicated to 360-degree IQA are rather few. For instance, a pre-trained model (MC360IQA) is used in [10] to predict the quality on viewports extracted from the cube-map projection of the 360-degree image. Six viewports are extracted and used as inputs of a pre-trained ResNet-34 [7] whose outputs are weighted and concatenated to predict the quality score. The most important component in this model is the pre-trained ResNet that was originally trained on ImagNet [11]. The latter dataset is composed of natural images with distortion occurring in the camera pipeline only, which would not allow the proposed model to predict visual quality for other distortions like compression. A deep learning framework is proposed in [12] where the quality scores are predicted on weighted patches extracted from the equirectangularly projected (ERP) image. ERP images do not sound efficient as the content is geometrically distorted. This problem was tackled in [13] in the development of the SSP-BOIQA metric. Hence, the polar regions are separated from the rest of the sphere when assessing 360-degree image quality. The features are then extracted from both equatorial and polar regions separately. Still, ERP equatorial regions do not necessarily represent nor reflect the actual viewed content by the users.

Different from the aforementioned models, we propose in this paper a no-reference metric based on CNN considering different perceptual characteristics of the human visual system (HVS) represented by the just noticeable difference (JND) and the visual scan-path. First, we extract viewports on the spherical content of 360-degree images according to visual scan-path predictions rather than a projected format. This way, we reproduce the actual viewed content. Then, motivated by the effectiveness of well-known pre-trained CNN models, we use DenseNet-121 [6] to extract visual features from the selected viewports and predicts their visual quality. We use the JND probability map to account for HVS sensitivity to local dis-

tortions. The proposed model estimates the weight of each extracted viewport by fusing JND, extracted visual features, and visual scan-path attributes (fixation duration and fixation order).

2. PROPOSED METHOD

The proposed method involves two steps. The first focuses on data pre-processing including scan-path prediction, viewports extraction, and JND probability maps prediction. The second step consists of an end-to-end training. Details on each step are given below.

2.1. Pre-processing

Fig. 1 provides an illustration of a 360-degree image viewing. Inspired by the way 360-degree images are generally viewed, *i.e.* only portions of the images called viewports are seen by the users through HMDs, we only consider selected viewports to predict the quality. This can be justified by the fact that a user can only see the current rendered field of view (FoV) from the spherical representation. The next viewport depends on his head direction along the x, y, and z axes. This way, quality prediction scenario tends to be in agreement with the viewing experience of 360-degree images and geometric distortions caused by the sphere to plane projection mentioned previously are avoided.

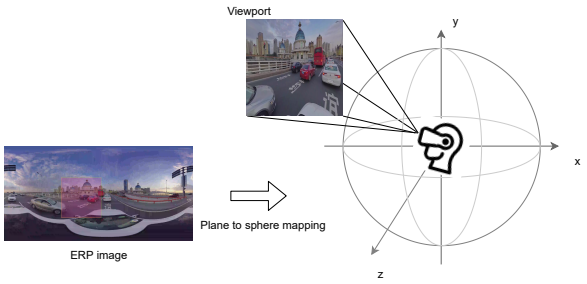

Fig. 1: 360-degree images viewed using head-mounted devices.

It is now widely admitted that when an image is viewed, the HVS gazes on salient details, which translates into eye fixations [14]. In our case, these regions are considered as our viewports and are detected using the visual scan-path model proposed in [15]. This model provides trajectories including the order and duration of fixations. This information giving valuable data about the exploration behavior is fed to the CNN model described in the next section. It corresponds to a sequence of N ordered fixation positions and their corresponding duration, respectively denoted as $[F_{Or}, F_D]$. In our model, the above-mentioned information is predicted for ten different virtual observers representing the diversity of human scan-paths. The predicted scan-paths are considered as data augmentation, not for the training stage but to increase the diversity and robustness of the cross-validation. This will help with the generalization analysis. The motivation behind such an approach is that each virtual observer (VO) will explore the same scene but will probably provide a different rating as in real subjective experiments. So, from each image in the dataset, we extract eight viewports for each VO where fixation points are taken as the center of the viewports with 512×512 resolution. This way, we generate ten different instances of the dataset. During the end-to-end training, each VO is used separately.

With the aim to perceptually account for the sensitivity of viewport content to distortions, and give more cues to our model about distortion visibility, we used just noticeable difference (JND) probability maps. We believe that training the model to learn about HVS sensitivity will perceptually improve the estimation of the weights to be given to each viewport when deciding about the quality of the 360-degree image. Fig. 2 gives samples from extracted viewports and their respective JND probability maps. It shows the impairments detection probability values and their variation depending on the complexity of the region. Flat regions are prone to more visible distortions compared to more complex ones.

Fig. 2: (Top) Examples of extracted viewports and (Bottom) their corresponding JND probability maps.

2.2. Network Architecture

Fig.3 depicts the architecture of the proposed method with its different components. Given a set of viewports Vp_i with $i \in \{0...N\}$ extracted from a 360-degree image, the model takes four inputs for each Vp_i including its visual content, its JND probability map, fixations order and fixations duration. These inputs are fed to the local quality predictor (LQP) (green rectangle) resulting in $N\times$ LQP modules running in parallel. Then, the LQP module fuses different learned features and outputs a weighted quality score for each Vp_i denoted as WQ_{Vp_i}. Finally, the model outputs the weighted arithmetic mean of the local quality scores as follow:

$$PredictedMOS = \sum_{i=1}^{N} {}^{WQ_{Vp_i}} / \sum_{i=1}^{N} W_i. \quad (1)$$

As shown in Fig. 3, the main component is the LPQ which consists of three parts. The first is a visual feature extractor (VFE). Here, we use the DenseNet-121 [6] model with its original weights. The choice of the DenseNet model is made based on a previous comparative study that we conducted and for which it ranked first compared to VGG, ResNet, and Inception architectures. The VFE provides a learned visual feature map Vf_{Vp_i} that goes to a quality estimation module and is used also for the estimation of the weight W_{Vp_i}. The second part consists of JND features extractor that takes the JND probability map $JNDmap_{Vp_i}$ of Vp_i and outputs a feature map that contributes to W_{Vp_i} estimation. The Learned JND features account for the different sensitivities of the HVS toward various distortion types and magnitudes. For the JND probability maps detection, we used the 2D model proposed in [16] as it is applied on the extracted viewports being assimilated to standard 2D images.

The proposed network used for JND features extraction aims to learn from HVS sensitivities [8]. It is composed of three convolutional blocks as illustrated in Fig. 3 (blue rectangle). Each block includes three layers, two convolutions (1×1 and 3×3) kernels followed by a max-pooling layer. By adding a 1×1 convolutional layer before the 3×3 convolution, for the same height and width of the

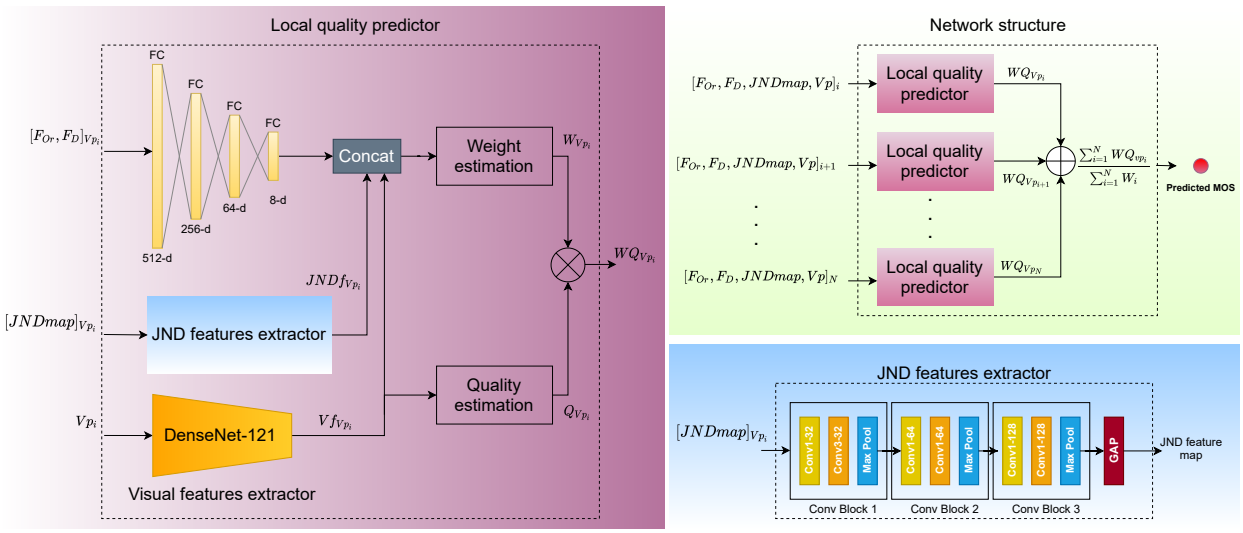

Fig. 3: Architecture of the proposed model: The green rectangle depicts the overall network structure, the magenta rectangle depicts the local quality predictor structure, and the blue rectangle depicts the JND features extractor network.

JND feature map, we reduce the number of operations. It also adds non-linearity to the network and allows to implement a smaller CNN while keeping a higher degree of accuracy [17]. Therefore, we are reducing the computational requirements and being more efficient at the same time.

At the final stage of the network, a global average pooling (GAP) is used according to the recommendation in [17] to generate the feature map. Finally, the third part is a multi-layer perceptron (MLP) that takes as input the duration and order of fixations given by the visual scan-path predictor and encode them to account for the visual exploration behavior. The MLP outputs a visual information vector used for the estimation of W_{Vp_i}. The fixation duration informs about which visual content is more likely to attract the user gaze. It also gives the time spent in visualizing a portion of the scene. As for the fixation order, it informs about the nature of the visual exploration path.

The weight estimation stage considers encoded duration and order of fixations, JND, and visual feature maps. These different features are fused and used to estimate the weights W_{Vpi} of Vp_i using four fully connected (FC) layers. In parallel, the visual feature map is also regressed to predict the quality score Q_{Vpi} of Vp_i. For this, a GAP is performed on the output feature map of DenseNet-121 followed by an FC layer, a dropout layer, and another FC layer for score prediction. The final score is computed using Eq. 1.

For the end-to-end training, we used the L_2 loss function to compute the error between predicted and target scores. The loss function is defined as:

$$loss = (q_{predicted} - q_{target})^2. \quad (2)$$

Three different versions of the proposed model are developed. The first version uses only fused visual features of the 8 extracted viewports from a given 360-degree image. It consists of eight pre-trained DenseNet-121 and a trained quality estimator. The second version accounts for scan-path features F_{Or} and F_D for weights estimation. Finally, the third version is built on top of version two by incorporating the JND probability maps for weight estimations.

3. RESULTS AND DISCUSSION

3.1. Data and implementation:

Dataset: This study is carried out using the CVIQD2018 [18] database. It contains 16 original 360-degree images, compressed using JPEG, H.264/AVC, and H.265/HEVC codecs. It counts in total 544 ERP images and their associated MOS which makes it to this date the largest available database in this field. We used the Pareto principle to split the database into training, validation, and testing sets. The use of a second database is very important for model validation and generalization analysis. We tried to use two different databases [19, 20] to perform a cross-database validation. Unfortunately, we discovered some inconsistencies in terms of subjective scores that could not be explained. So we decided to discard them. To compensate for this lack, we used the strategy discussed in Sec. 2.1 where the proposed model is compared across ten predicted scan-paths.

Implementation: The proposed architecture is implemented using TensorFlow [21] and will be publicly available. The training was performed using NVIDIA Tesla P100-PCIE-16GB and 26GB of RAM. We used the *earlystopping* to stop the training if no performance gain is observed by monitoring the validation loss.

3.2. Performance evaluation

To assess the performance of our model, we used the Pearson Linear Correlation Coefficient (PLCC) and the Spearman Rank Order Correlation Coefficient (SRCC). The predicted scores are fitted using a five-parameter non-linear logistic function. The performance of the proposed model are computed using the ten VOs. The MIN and MAX represent respectively the least and best performance among VOs.

3.2.1. Ablation study

To evaluate the effectiveness of the considered additional inputs (scan-path visual information and JND maps). We provide an ab-

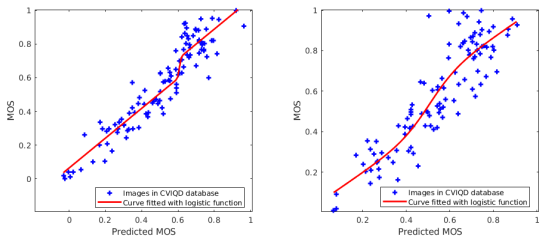

Fig. 4: Scatter plots of predicted quality scores versus MOS of the final model SP360IQA-F-JND (Best performance on the left and worst performance on the right).

lation study. It focuses on performance added to the model by the additional components. First, we predict the quality score using only regressed visual features on 8 viewports extracted based on the virtual observer scan-path denoted as SP360IQA. Second, we add the viewport weight estimation as described in Sec. 2.2 by encoding scan-path visual information thought an MLP (see Fig. 3). This version is denoted as SP360IQA-F. Finally, we optimize the estimation of the weights by exploiting JND probability maps of the selected viewports to account for HVS sensitivity and provide perceptual distortion-ability to the model, denoted as SP360IQA-F-JND. Table. 1 provides the results of the conducted ablation study. The maximum and minimum values of PLCC and SRCC regarding all VOs are given, in addition to standard deviations. One can observe that the proposed weight estimation improves the performance when considering fixations order and duration for each viewport. The minimum PLCC/SRCC shifts from 0.78/0.75 to 0.89/0.86 showing that the model gained significantly in terms of accuracy and monotonicity. The incorporation of the JND further boosted the performances but with a slight shift. Therefore, we conclude that using scan-path visual information and JND features contribute to the prediction accuracy of our model. It also contributes to the generalization of our model as given by the SD values. Indeed, the latter are decreased explaining that VOs are providing better and less spread performances. Scatter plots of the predicted scores versus MOS of the best and least performing VO are given in Fig. 4. It supports the aforementioned discussions and shows consistent distribution of the predictions.

Table 1: Standard deviation, maximum and minimum performance in terms of PLCC and SRCC of virtual observers. Best PLCC values are highlighted in bold and SRCC underlined.

	SP360IQA		SP360IQA-F		SP360IQA-F-JND	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
MAX ↑	0.929	0.911	0.945	0.921	**0.949**	<u>0.928</u>
MIN ↑	0.780	0.750	0.889	0.863	**0.900**	<u>0.866</u>
SD ↓	0.044	0.045	0.020	<u>0.021</u>	**0.019**	0.023

3.2.2. Performance comparison

The performances of our model are compared with state-of-the-art quality models including: 1) 2D full reference (FR) metrics like PSNR and SSIM, 2) Learning-based NR 2D models such as BRISQUE [22], QAC [23], BPRI [24] and DipIQ [25], 3) PSNR-based 360-degree models WS-PSNR, S-PSNR and CPP-PSNR, and 4) learning-based NR 360-degree metrics SSP-BOIQA, MC360IQA$_{origin}$ and MC360IQA$_{mean}$ trained respectively without and with data augmentation. Table. 2 summarizes the performances of aforementioned metrics on the CVIQD database. We can notice that traditional 2D models and their extended versions have significantly lower performance compared to 360-degree models. Therefore, they are not well suited for this type of image as already demonstrated in benchmark studies [5]. SSP-BOIQA slightly improves the correlation with subjective MOS compared to SSIM that measures the structural similarity according to the HVS characteristics. MC360IQA versions provide good results. At its lowest performance (MIN), our model outperformed all state-of-the-art FR, NR, and 360-degree models except MC360IQA. Regarding the latter, the $origin$ version is outperformed by the three versions of the model with the VO providing the maximum performance. The $mean$ version is in turn outperformed by the F and F-JND version of the proposed model.

Table 2: Performance comparison with state-of-the-art quality models in terms of PLCC and SRCC. Best performance is highlighted in bold.

	Metric	PLCC	SRCC
FR	PSNR	0.7662	0.7320
	SSIM	0.8972	0.8857
	S-PSNR	0.7819	0.7574
	WS-PSNR	0.7741	0.7467
	CPP-PSNR	0.7755	0.7498
NR	BRISQUE	0.7641	0.7448
	QAC	0.8681	0.8299
	BPRI	0.8877	0.8576
	DipIQ	0.8065	0.7381
Learning-based 360-degree	SSP-BOIQA	0.9077	0.8614
	MC360IQA$_{origin}$	0.9271	0.9069
	MC360IQA$_{mean}$	0.9391	0.9153
Ours	MIN	0.900	0.866
	MAX	**0.949**	**0.928**

4. CONCLUSION

We presented in this paper a CNN-based model for 360-degree IQA. This model relies on predicted scan-paths for the extraction of adapted viewports. In addition, to account for the HVS properties, fixations order and duration are used together with JND to define weighting factors exploited for quality pooling. This adopted weighting strategy has shown a significant improvement of the prediction performances. Additionally, taking advantage of the variability of visual exploration of 360-degree scenes (visual trajectory) through virtual observers, is a significant added value for model generalization analysis. Our model showed the usefulness of predicting quality on the spherical content rather than projected one. We believe that using additional HVS properties may greatly contribute to the improvement of prediction accuracy. So, a more optimized network for learning HVS properties for 360-degree quality assessment tasks will be investigated.

5. ACKNOWLEDGEMENTS

This work is funded by the Region "Nouvelle Aquitaine" under project SIMOREVA360 2018-1R50112 and by French National Research Agency as part of ANR-FILTER2 project (ANR-16-CE39-0013).

6. REFERENCES

[1] J. J. . Lin, H. B. L. Duh, D. E. Parker, H. Abi-Rached, and T. A. Furness, "Effects of field of view on presence, enjoyment, memory, and simulator sickness in a virtual environment," in *IEEE Virtual Reality*, Orlando, FL, USA, 2002, pp. 164–171.

[2] M. Yu, H. Lakshman, and B. Girod, "A framework to evaluate omnidirectional video coding schemes," in *IEEE International Symposium on Mixed and Augmented Reality*, Fukuoka, Japan, 2015, pp. 31–36.

[3] Yule Sun, Ang Lu, and L. Yu, "Weighted-to-spherically-uniform quality evaluation for omnidirectional video," *IEEE Signal Processing Letters*, vol. 24, pp. 1408–1412, 2017.

[4] V. Zakharchenko, PC. Kwang, and HP. Jeong, "Quality metric for spherical panoramic video," in *Optics and Photonics for Information Processing X*, 2016, vol. 9970, pp. 57 – 65.

[5] A. Sendjasni, MC. Larabi, and FA. Cheikh, "On the improvement of 2d quality assessment metrics for omnidirectional images," in *Electronic Imaging*, Burlingame, California USA, 2020, pp. 287–1.

[6] G. Huanga, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE conference on computer vision and pattern recognition*, Honolulu, HI, USA, 2017, pp. 4700–4708.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Las Vegas, NV, USA, 2016, pp. 770–778.

[8] J. Kim and S. Lee, "Deep learning of human visual sensitivity in image quality assessment framework," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 1969–1977.

[9] S. Seo, S. Ki, and M. Kim, "A novel just-noticeable-difference-based saliency-channel attention residual network for full-reference image quality predictions (early access)," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2020.

[10] W. Sun, W. Luo, X. Min, G. Zhai, X. Yang, K. Gu, and S. Ma, "MC360IQA: The multi-channel CNN for blind 360-degree image quality assessment," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, Sapporo, Japan, Japan, 2019, pp. 1–5.

[11] O. Russakovsky, J. Deng, H. Su, J Krause, and S. Satheesh et al., "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[12] H. G. Kim, H. Lim, and Y. M. Ro, "Deep virtual reality image quality assessment with human perception guider for omnidirectional image," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 4, pp. 917–928, 2020.

[13] X. Zheng, G. Jiang, M. Yu, and H. Jiang, "Segmented spherical projection-based blind omnidirectional image quality assessment," *IEEE Access*, vol. 8, pp. 31647–31659, 2020.

[14] D. Noton and L. Stark, "Scanpaths in saccadic eye movements while viewing and recognizing patterns," *Vision research*, vol. 11, no. 9, pp. 929–IN8, 1971.

[15] W. Sun, Z. Chen, and F. Wu, "Visual scanpath prediction using ior-roi recurrent mixture density network (early access)," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019.

[16] J. Wu, G. Shi, W. Lin, A. Liu, and F. Qi, "Just noticeable difference estimation for images with free-energy principle," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1705–1710, 2013.

[17] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[18] W. Sun, K. Gu, S. Ma, W. Zhu, N. Liu, and G. Zhai, "A large-scale compressed 360-degree spherical image database: From subjective quality evaluation to objective model comparison," in *IEEE 20th international workshop on multimedia signal processing (MMSP)*, Vancouver, BC, Canada, 2018, pp. 1–6.

[19] M. Huang, Q. Shen, Z. Ma, A. C. Bovik, P. Gupta, R. Zhou, and X. Cao, "Modeling the perceptual quality of immersive images rendered on head mounted displays: Resolution and compression," *IEEE Transactions on Image Processing*, vol. 27, no. 12, pp. 6039–6050, 2018.

[20] H. Duan, G. Zhai, X. Min, Y. Zhu, Y. Fang, and X. Yang, "Perceptual Quality Assessment of Omnidirectional Images," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, Florence, Italy, 2018, pp. 1–5.

[21] A. Martín, P. Barham, J. Chen, Z. Chen, and A. Davis et al., "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} Symposium on Operating Systems Design and Implementation*, Savannah, GA, USA, 2016, pp. 265–283.

[22] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[23] W. Xue, L. Zhang, and X. Mou, "Learning without human scores for blind image quality assessment," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 995–1002.

[24] X. Min, K. Gu, G. Zhai, J. Liu, X. Yang, and C. W. Chen, "Blind quality assessment based on pseudo-reference image," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2049–2062, 2018.

[25] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "Dipiq: Blind image quality assessment by learning-to-rank discriminable image pairs," *IEEE Transactions on Image Processing*, vol. 26, no. 8, pp. 3951–3964, 2017.