https://doi.org/10.2352/EI.2022.34.9.IQSP-395

Patch-based CNN Model for 360 Image Quality Assessment with Adaptive Pooling Strategies

Abderrezzaq Sendjasni [1,2]***, Mohamed-Chaker Larabi*** [1] ***and Faouzi Alaya Cheikh*** [2]
[1] ***CNRS, Xlim UMR 7252, Université de Poitiers, France***
[2] ***NTNU, Norwegian Colour and Visual Computing Lab, Gjøvik, Norway***

Abstract

Patch-based training for 360-degree images allows to significantly reduce the complexity compared to multichannel models while maintaining good performances. Differently from multichannel models where multi neural networks are trained in parallel to predict the score of the whole 360-degree image, a pooling stage is required to map local qualities to the global one. This step is often neglected by using a simple arithmetic mean, which does not account for (i) the non-uniformity distribution of quality and (ii) the variability among local qualities. In this paper, we analyze several pooling strategies, including basic statistic methods and adaptive pooling ones. Additionally, we propose a pooling strategy based on scene exploration behavior relying on visual scan-path. The performance analysis showed the benefit of using adaptive pooling over arithmetic mean, as well as the incorporation of perceptual properties during the pooling stage. Besides, the comparison with state-of-the-art multichannel models asserts the effectiveness of patch-based training compared to multichannel models.

Keywords: *Image quality assessment, Convolutional neural networks, 360-degree images, Perceptual quality, Adaptive pooling.*

Introduction

360-degree images are gaining more popularity. They are used in several applications, ranging from social media to virtual reality (VR). Such images enable the viewer to explore a scene in an omnidirectional way using head-mounted displays (HMD). This offers an immersive experience to the user, but only a portion of the image, called viewport, is viewed at a time. It consists of a rendered window from the sphere in a given direction based on the yaw, pitch and roll of the user. To ensure the improvement of quality of experience (QoE) and immersiveness, it is important to dispose of tools allowing to assess it. The literature is rich of objective quality metrics, but most of them address standard 2D images and are not appropriate for immersive content, including 360-degree.

Since 360-degree images are different from existing 2D images, a few IQA models have been proposed by extending traditional 2D models such as PSNR or mean squared error (MSE) to account for this difference. For example, PSNR-based methods like Spherical PSNR (S-PSNR) [1], weighted spherical PSNR (WS-PSNR) [1], and craster parabolic projection PSNR (CPP-PSNR) [2]. These metrics either are proposed for a specific projection format, like the CPP-PSNR, or are computed on the sphere like S-PSNR. As these models do not account for perceptual aspects, they fail in predicting the visual quality accurately. This motivates the use of data-driven approaches to achieve better and accurate IQA models, and more specifically, convolutional neural networks (CNNs).

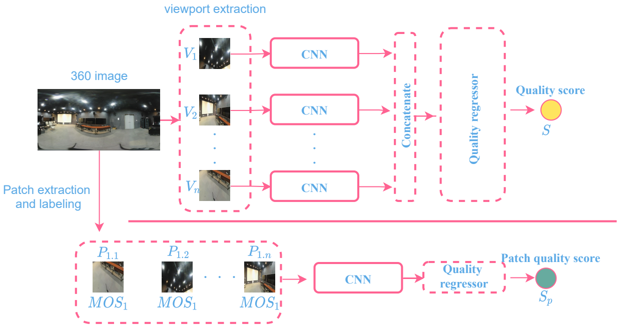

Figure 1: CNN models for IQA. (top) multichannel vs. (bottom) patch-based CNN.

Image quality evaluation using CNNs has demonstrated good performances over the past decade, especially for 2D images. For 360-degree IQA, the multichannel paradigm is usually adopted, where multiple CNNs are used to extract features from different regions in parallel [3, 4, 5]. The extracted features from each CNN are concatenated and regressed to a single quality score (*see.* Fig. 1 top). This allows to train and optimize the model to the mean opinion score of the whole 360-degree image. Hence, the model learns to fuse the different feature maps together to predict a quality score. However, the computational complexity induced by this paradigm is significant, and may have a negative effect on the overall optimization of the model as it makes the latter hard to train. In addition, the concatenation of feature maps must be guided in order to give importance to features extracted from important regions. For instance, Xu *et al.* [5] used twenty ResNet-18 [6] in parallel to extract visual features from twenty viewports. The output from each one is then used by a graph CNN to learn the dependencies among the selected viewports. Additionally, a subnetwork that takes equirectangular (ERP) images as input is used to account for the global quality. The subnetwork is composed of a VGG-16 [7] and the deep-bilinear CNN [8]. Sun *et al.* [3] used six pre-trained hyper ResNet-34 [6] by combining features from intermediate layers with later ones on each channel separately. Zhou *et al.* [9] used six Inception-V3 [10] through a shared weight strategy. The proposed model is trained to predict the quality score in addition to the classification of distortions. By applying the multichannel paradigm, the aforementioned models become extremely complex in terms of trainable parameters as well as floating operations needed. Therefore, their training requires significantly more processing resources.

In contrast, the patch-based training takes individual regions separately. Here, a single CNN is used (*see.* Fig. 1 bottom) which implies less complexity and leads to faster training. In a patch-based IQA framework, two important aspects must be carefully considered. The first one corresponds to patches' selection and extraction. This is usually performed by using some criteria such as saliency. In the extraction of these patches, the use of radial content (*i.e.* from the sphere) is highly recommended compared to the projected one [11, 12]. This way, the geometric distortion induced by the sphere-to-plane projection can be avoided. The second aspect focuses on the aggregation of local qualities to a global quality score that should account for (i) the non-uniformity distribution of quality, and (ii) the variation among quality scores of individual patches.

In the literature, several works adopted the patch-based training for 2D IQA [13, 14, 15, 16], and good performances have been reported. The interest of such an approach lies in its proven performance in various image processing tasks such as medical imaging, image classification and recognition. Also, the quality prediction tends to agree with the scene exploration by focusing on prominent parts of it that are translated into patches. However, the unavailability of mean opinion scores (MOS) for individual patches is considered as the main issue of this approach. Existing models label all patches extracted from the same image with the same MOS. Despite this limitation, the achieved results are quite interesting, and could be due to the adopted pooling strategy. Therefore, one may consider adopting appropriate strategies would provide better performances.

Pooling strategies have been investigated for 2D images [17, 18, 19, 20, 21], with the aim to map quality and distortions maps to a final score. Several strategies are considered, ranging from basic statistics and percentile pooling to content-based and information weighted spatial pooling. In [22] temporal pooling methods are compared for video quality assessment, where individual scores from the different frames are pooled to a single quality score of the video. It is known that quality scores pooling is paramount in IQA frameworks, especially for patch-based CNNs. To the best of our knowledge, there is no study featuring quality scores pooling strategies with CNNs for IQA, either with 2D nor 360-degree images.

In this paper, we present a comparative study of pooling techniques for CNN-based 360 IQA models. The considered methods include basic statistics and adaptive pooling. To this end, a patch-based CNN model is designed by fine-tuning ResNet-50 [6]. To select relevant patches, a visual scanpath is used to mimic the exploration behavior of human observers by predicting possible visual trajectories. The latter are used to define patches locations on the spherical content rather than the projected one. The visual trajectories are also incorporated at the pooling stage.

Adaptive Pooling Strategies

A patch-based CNN model is basically trained on individual patches extracted from the input images. This means that the model is trained only on these patches, without having access to the whole 360-degree images. Therefore, N scores associated to N patches are predicted, and mapping of these individual scores to a single quality score is challenging. This operation must be performed by adaptive pooling to improve the correlation with the human judgement quality. Fig. 2 illustrates the pooling operation, where for each 360-degree image I, N predicted scores $S = \{S_{p_1}, S_{p_2}, ..., S_{p_n}\}$ corresponding to N patches $P = \{P_1, P_2, ..., P_n\}$ are pooled together using the function $Pooling(.)$.

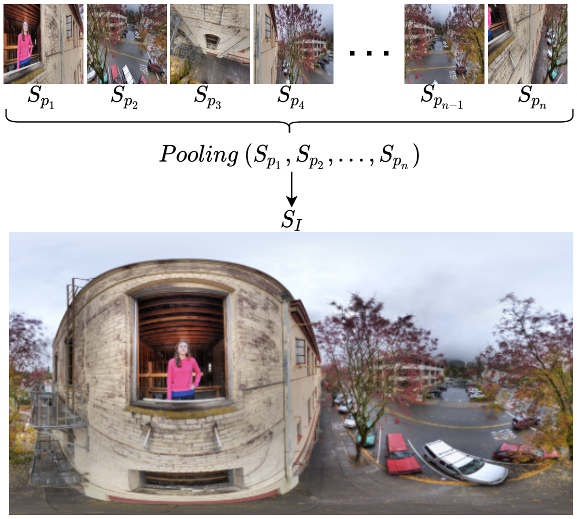

Figure 2: Mapping of predicted local qualities S_{P_i} (per patch) to global quality S_I (per 360-degree image).

Basic Statistic Methods

Arithmetic Mean

The arithmetic mean is a straightforward method for pooling local qualities to a global one. By simply averaging the quality scores, the local qualities will contribute equally to the final score as follows:

$$S_I = \frac{1}{N}\sum_{i=1}^{N} S_{p_i}. \tag{1}$$

Harmonic Mean

The harmonic mean is one of the Pythagorean means. It is calculated by dividing the number of scores by the reciprocal of each score S_{p_i} in S. Hence, the harmonic mean is the reciprocal of the arithmetic mean of the reciprocals. It is known to emphasize the impact of small scores [23] reflecting the fact that subjective ratings are influenced by worst regions in terms of visual quality. The harmonic mean is calculated as follows:

$$S_I = (\frac{1}{N}\sum_{i=1}^{N} S_{p_i}^{-1})^{-1}. \tag{2}$$

Geometric Mean

The geometric mean is the third Pythagorean mean. It signifies the central tendency or typical values of S by taking the root of the product of their values, as given below:

$$S_I = (\prod_{i=1}^{N} S_{p_i})^{\frac{1}{N}}. \tag{3}$$

Five-Number Summary

This method provides a description of S using various descriptive statistics [24]. The five-number summary makes use of information on (i) the location given by the median, (ii) the spread of the scores given by the $Q1$ and $Q3$ quartiles representing the 25% and 75% percentile respectively, and (iii) the range of values expressed by the minimum and maximum of S. Therefore, the five-number summary is computed as follows:

$$S_I = \frac{min + Q1 + median + Q3 + max}{5}. \quad (4)$$

Minkowski Mean

The Minkowski pooling has been widely used for IQA [17, 25]. The P parameter emphasizes the lowest scores among S, *i.e.* the highly distorted patches. To understand the influence of the latter, we set it values to the most commonly used ones in the literature, including 1/4, 1/2, 2, 4, 8 and 16.

$$S_I = (\frac{1}{N}\sum_{i=1}^{N} S_{p_i}^{p})^{\frac{1}{P}}. \quad (5)$$

Percentile Pooling

The percentile pooling is considered as one of the most effective pooling methods. It is based on the fact that perceived quality is strongly affected by the most distorted regions [18]. This is accomplished by considering only the quality scores from S that are lower than a $k-$th percentile. In order for us to study the impact of this latter, five percentiles are used as threshold including 5%, 10%, 20%, 25%, and 50%.

$$S_I = \frac{1}{\mid S \downarrow k\% \mid} \sum_{i \in S \downarrow k\%} S_{p_i}. \quad (6)$$

Scene Exploration Based Pooling

It is known that pooling strategies based on basic statistics tend to have poor correlations. It is especially the case of simple mean pooling that enforces an equal contribution of all patches to the global quality scores. By doing so, the non-uniformity distribution of quality is not taken into account. For this reason, a weighted mean pooling can reproduce this behavior by weighting each local score according to the importance of the patch's content. The estimation of these weights are usually based on perceptual properties such as visual attention [26], equator-bias [27] to incorporate the way the human gaze is biased toward the equator, making the computation of these weights handcrafted. Others opted for data-driven based estimation of the weights by adding subnetworks within a CNN model [13, 28].

Differently, we adopted a weighting strategy based on visual exploration. This is motivated by the fact that quality metrics are tuned and compared against the MOS collected by psychophysical experiments. By incorporating the way observers explore a scene before rating its quality could improve the pooling performance. Thus, an observer explores a visual scene by focusing on certain regions and usually not all parts of the scene. This behavior can be modeled using visual scanpaths by predicting possible visual trajectories based on head and gaze movements. As the exploration of a scene is different from one observer to another, we use multiple and different visual scanpaths to account for this diversity. To do so, the scanpath model proposed in [29] is used to predict possible visual trajectories. Ten scanpaths, composed of eights gaze fixations for each 360-degree image I are generated. Two important information associated with each fixation are considered as weights for each patch P_i extracted from I. The first is the order of fixations, expressing the temporal progress of the visual trajectory. The second is the duration, representing the amount of time a region is likely to be focused on. The longer the gaze, the greater the influence on the observers' judgment. Finally, the pooling is performed as shown by Eq. 7, with W_i is either the fixation duration or fixation order associated with patch P_i.

$$S_I = \frac{\sum_{i=1}^{N} W_i S_{p_i}}{\sum_{i=1}^{N} W_i}. \quad (7)$$

Furthermore, to account for previous observations about the impact of most distorted regions highly on perceived quality, we combine the fixation-based pooling with the percentile threshold as given in Eq. 8.

$$S_I = \frac{\sum_{i \in |S \downarrow k\%|} W_i S_{p_i}}{\sum_{i \in |S \downarrow k\%|)} W_i}. \quad (8)$$

Patch-based Model

The process of selection and extraction of patches is depicted on Fig. 3. Hence, relevant patches are obtained using the visual scanpath discussed previously. The predicted eye fixations by the scanpath model are considered as the center of patches. By taking the content surrounding these fixations, we extract patches of 256 × 256 pixels. Selected patches are then extracted on the sphere in order to avoid any geometric distortions due to the sphere-to-plane projection [30, 11]. In total, eighty patches are extracted from the 360-degree image I. Each patch extracted from the same image receives the MOS of the 360 image as a label.

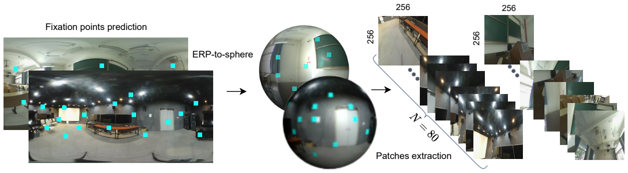

Figure 3: Selection and extraction of patches.

In this paper, a patch-based CNN is designed using ResNet-50 [6] with the ImageNet [31] weights as the base model to extract visual features from selected patches. Fig. 4 depicts its architecture. We replaced the top layers with a regression block in order to regress the learned features into a single quality score. The extracted features $F_{W \times H \times C}$ where H, W, and C stand for the height, width, and dimension, are fed to a global average pooling (GAP) so to reduce the spatial dimensions of the extracted feature maps and to avoid overfitting. The GAP outputs a feature vector F' of size $1 \times 1 \times C$ that is in turn fed to a fully connected (FC) layer with dimension of 512 followed by a rectified linear unit (ReLU) [32] activation function and a dropout regularization [33] with ratio of 0.2. The output of the latter is sent to a FC layer with a single node followed by a linear activation function to deliver the quality score S_{P_i}. The weights for the quality regressor are initialized according to the method proposed by He *et al.* [34]. For the end-to-end training, we used the L_2 loss function to compute the error between predicted and target scores.

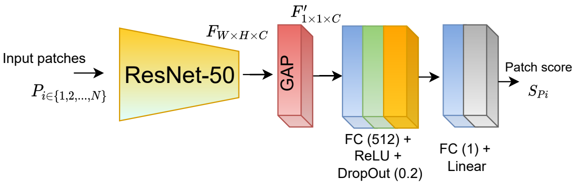

Figure 4: Architecture of the adopted model including global average pooling (GAP) and fully connected layer (FC).

Experiments

Experimental Setup

We selected two benchmark 360-degree image databases, namely OIQA [35] and CVIQ [3], to evaluate the pooling strategies described previously. OIQA consists of 320 distorted images generated from 16 pristine ones by applying five levels of JPEG, JPEG 2000 (JP2K), Guassian blur (GB), and Gaussian white noise (GWN). CVIQ contains 528 distorted images obtained from 16 pristine ones by applying eleven levels of JPEG, H.264/AVC, and H.265/HEVC.

The model is trained on a server with Intel Xeon Silver 4208 2.1GHz, 192G RAM and a GPU Nvidia Telsa V100S 32G. The batch size was set to 32 and the Adam optimizer [36] is used with a learning rate of $1e-4$, first parameter $\beta_1 = 0.9$ and second parameter $\beta_2 = 0.999$. We used the early stopping by monitoring the validation loss to stop the training once no improvement is observed and retain the best state of the model. Five-fold cross-validation is performed for a complete evaluation with each database. During training, the databases are split using the well known Pareto principle, 80% for training and 20% for testing. To ensure a complete separation of the training and testing sets, the distorted images associated to the same pristine image are allocated to the same set.

Results and Discussion

The performance evaluation is performed by computing the Pearson linear correlation coefficient (PLCC) for accuracy, the Spearman rank order correlation coefficient (SRCC) for monotonicity, and the root-mean-square error (RMSE) for prediction errors between the ground truth MOS and the predicted scores. The provided performances are computed as the median of the five-fold cross-validation.

We summarize the performances of all pooling strategies in Table 1. Overall, one can notice that the widely used arithmetic mean ranks among the worst approach on both databases, demonstrating its weakness when it comes to quality pooling. Pooling strategies accounting for the variability among quality scores should be considered in this case, as shown by the performance results in Table 1. One can observe that harmonic and geometric means outperformed the arithmetic one on both databases. For instance, the harmonic mean performances are approx. 0.8% PLCC, 1.0% SRCC, and 4.2% RMSE better than the arithmetic mean on OIQA, and approx. 0.6% PLCC, 1.2% SRCC, and 4.0% RMSE on CVIQ. The Minkowski mean and the five-number summary did not perform well compared to arithmetic mean, a slight improvement can be observed with the Minkowski mean, whereas the five-number summary did not appear to express the nature of the variability among the local qualities scores. The percentile pooling achieved the best performance in terms of PLCC and SRCC on OIQA, and competitive results when combined with fixation orders and fixation durations on CVIQ. This shows that expressing the phenomena of perceived quality being impacted by the most distorted content improves the final quality pooling.

In the following, we analyze closely the performance of the Minkowski mean and the Percentile pooling. The evaluation of PLCC/SRCC scores is given in Fig. 5 and 6, respectively. For the Minkowski mean, one can observe a decrease in both accuracy and monotonicity with the increase of P. This observation is valid on both databases, with a significant margin on CVIQ, approximately 6% with PLCC and 10% with SRCC. As for the Percentile Pooling, an increase of performances can be observed with a saturation at $k = 25$ on OIQA and $k = 10$ on CVIQ, followed by a decease of performance. Based on these observations, the parameter for both methods should be carefully chosen, as it is dependent on the variability and span of local qualities. In addition, the difference among OIQA and CVIQ is due to the nature and diversity of their content, as shown in [30]. This is also depicted by the provided curves, where an important gap between PLCC and SRCC values can be observed on CVIQ compared to OIQA independently of the used pooling methods.

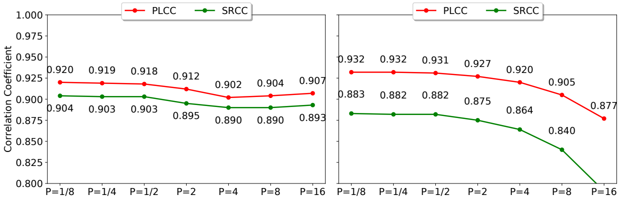

Figure 5: Performance of Minkowski mean in terms of PLCC/SRCC on OIQA (left) and CVIQ (right).

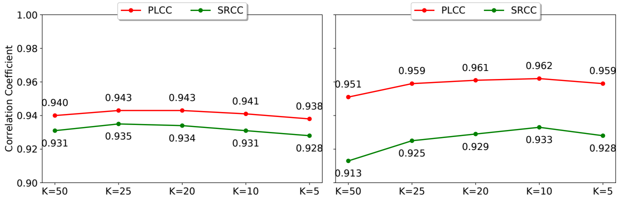

Figure 6: Performance of Percentile pooling in terms of PLCC/SRCC on OIQA (left) and CVIQ (right).

With the intent to show the effectiveness of the patch-based CNN over multichannel models, we provide in Table 2 a performance comparison with three state-of-the-art models. These models adopt a multichannel paradigm using different strategies. From the table, one can observe that patch-based CNN with a simple arithmetic mean achieved competitive results compared to Sun *et al.* and Zhou *et al.*. However, it scored worse than Xu *et al.* (approx. 3.8% PLCC and 4.5% SRCC) on OIQA and (approx. 3.0% PLCC and 8.7% SRCC) on OIQA. When the adaptive pooling is used, a different behavior is observed. The patch-based

Table 1: Performance evaluation of the pooling strategies in terms of PLCC, SRCC, and RMSE. The best performance is highlighted in **bold** and second-best underlined

Database	OIQA			CVIQ		
Metric	PLCC	SRCC	RMSE	PLCC	SRCC	RMSE
Arithmetic Mean	0.9162	0.9017	5.8185	0.9297	0.8786	5.0537
Harmonic Mean	0.9235	0.9105	5.5685	0.9352	0.8891	4.8582
Geometric Mean	0.9200	0.9066	5.6876	0.9326	0.8841	4.9537
Five-number summary	0.9061	0.8971	6.1415	0.9233	0.8721	5.2683
Minkowski Mean	0.9196	0.9045	5.7034	0.9322	0.8833	4.9660
Percentile Pooling	**0.9434**	**0.9340**	**4.8156**	0.9623	**0.9329**	3.6790
Fixation Order	0.9063	0.8931	6.1357	0.9305	0.8787	5.0273
Percentile Fixation Order	0.9392	0.9296	4.8265	0.9621	**0.9329**	**3.6287**
Fixation Duration	0.9164	0.9028	5.8096	0.9296	0.8792	5.0564
Percentile Fixation Duration	0.9403	0.9291	4.8658	**0.9625**	0.9324	3.6883

Table 2: Performance comparison with state-of-the-art mutlichannel-based models

	Database		OIQA		CVIQ	
	Multichannel	Number (Bacckbone)	PLCC	SRCC	PLCC	SRCC
Xu *et al.* [5]	✓	20 (Resnet-18)	**0.952**	**0.944**	0.959	**0.953**
Sun *et al.* [3]	✓	6 (ResNet-34)	0.924	0.918	0.950	0.914
Zhou *et al.* [9]	✓	6 (Inception-V3)	0.899	0.923	0.902	0.911
Ours $_{\textit{Arethmetic Mean}}$	✗	1 (ResNet-50)	0.916	0.902	0.930	0.879
Ours $_{\textit{Adaptive Pooling}}$	✗	1 (ResNet-50)	0.943	0.935	**0.963**	0.932

model outperformed Sun *et al.* and Zhou *et al.* on both databases, and scored slightly lower compared to Xu *et al.* on OIQA and achieved the best accuracy on CVIQ. This slight difference of performance could be considered as insignificant when weighted by the complexity generated by the multichannel architecture. These performances support the previous observation regarding the usefulness of adaptive pooling of local qualities on the one hand. On the other hand, patch-based CNN is as effective as multichannel networks, and sometimes even better if proper training techniques are adopted.

Conclusions

In this paper, we compared various pooling strategies for 360-degree IQA using patch-based CNN with adaptive pooling. We found that the use of a simple arithmetic mean does not account for the variability among the quality scores, and therefore, the correlation performance tends to drop. Adaptive pooling strategies are seen as a good answer to cope this limitation, especially when IQA-specific characteristics are incorporated. Moreover, patch-based CNN with adaptive pooling achieved competitive performances compared to state-of-the-art multichannel models. As patch-based CNNs introduce less complexity compared to multichannel, it makes it more appropriate with an adequate training strategy for 360-degree IQA.

Acknowledgments

This work is partially funded by the Nouvelle Aquitaine regional Council under project SIMOREVA360 2018-1R50112 and Project CPER/FEDER e-immersion.

References

[1] JVET. Algorithm description of joint exploration test model 6 (JEM6). Technical Report JVET-F1001, ITU-T VCEG (Q6/16) and ISO/IEC MPEG (JTC 1/SC 29/WG 11), 2017.

[2] V. Zakharchenko, P. Kwang, and H. Jeong. Quality metric for spherical panoramic video. In *Optics and Photonics for Information Processing X*, volume 9970, pages 57 – 65, 2016.

[3] W. Sun, X. Min, G. Zhai, K. Gu, H. Duan, and S. Ma. MC360IQA: A multi-channel CNN for blind 360-degree image quality assessment. In *IEEE Journal of Selected Topics in Signal Processing*, volume 14, pages 64–77, 2020.

[4] HG. Kim, H. Lim, and YM. Ro. Deep virtual reality image quality assessment with human perception guider for omnidirectional image. *IEEE Trans. on Circuits and Systems for Video Technology*, 30(4):917–928, 2020.

[5] J. Xu, W. Zhou, and Z. Chen. Blind omnidirectional image quality assessment with viewport oriented graph convolutional networks. *IEEE Trans. on Circuits and Systems for Video Technology*, 31(5):1724–1737, 2021.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE CVPR*, pages 770–778, Las Vegas, NV, USA, 2016.

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

[8] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Trans. on Circuits and Systems for Video Technology*, 30(1):36–47, 2020.

[9] Y. Zhou, Y. Sun, L. Li, K. Gu, and Y. Fang. Omnidirectional image quality assessment by distortion discrimination assisted multi-stream network. *IEEE TCSVT (Early Access)*, pages 1–1, 2021.

[10] C. Christian, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vi-

sion. In *IEEE CVPR*, pages 2818–2826, Las Vegas, NV, USA, 2016.

[11] A. Sendjasni, MC. Larabi, and FA. Cheikh. Perceptually-weighted CNN for 360-degree image quality assessment using visual scan-path and JND. In *IEEE ICIP*, pages 1439–1443, Anchorage, AK, USA, 2021.

[12] A. Sendjasni, MC. Larabi, and FA. Cheikh. Convolutional neural networks for omnidirectional image quality assessment: Pre-trained or re-trained? In *IEEE ICIP*, pages 3413–3417, Anchorage, AK, USA, 2021.

[13] S. Bosse, D. Maniry, K. Müller, T. Wiegand, and W. Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Trans. on Image Processing*, 27(1):206–219, 2018.

[14] L. Po, M. Liu, W. Yuen, Y. Li, and X. Xu et al. A novel patch variance biased convolutional neural network for no-reference image quality assessment. *IEEE Trans. on Circuits and Systems for Video Technology*, 29(4):1223–1229, 2019.

[15] J. Kim and S. Lee. Fully deep blind image quality predictor. *IEEE Journal of selected topics in signal processing*, 11(1):206–220, 2016.

[16] H. Wen and J. Tingting. From image quality to patch quality: an image-patch model for no-reference image quality assessment. In *IEEE ICASSP*, pages 1238–1242, USA, 2017.

[17] Z. Wang and X. Shang. Spatial pooling strategies for perceptual image quality assessment. In *IEEE ICIP*, pages 2945–2948, Atlanta, GA, USA, 2006.

[18] AK. Moorthy and AC. Bovik. Visual importance pooling for image quality assessment. *IEEE journal of selected topics in signal processing*, 3(2):193–201, 2009.

[19] Z. Wang and Q. Li. Information content weighting for perceptual image quality assessment. *IEEE Trans. on image processing*, 20(5):1185–1198, 2010.

[20] D. Temel and G. AlRegib. A comparative study of quality and content-based spatial pooling strategies in image quality assessment. In *IEEE GlobalSIP*, pages 732–736, FL, USA, 2015.

[21] G. Mingming and P. Marius. Spatial pooling for measuring color printing quality attributes. *Journal of Visual Communication and Image Representation*, 23(5):685–696, 2012.

[22] Z. Tu, C. Chen, L. Chen, and *et al.* A comparative evaluation of temporal pooling methods for blind video quality assessment. In *IEEE ICIP*, pages 141–145, UAE, 2020.

[23] Z. Li, C. Bampis, J. Novak, A. Aaron, and et al. VMAF: The journey continues. *Netflix Technology Blog*, 25, 2018.

[24] C. Zewdie, M. Pedersen, and Z. Wang. A new pooling strategy for image quality metrics: Five number summary. In *5th EUVIP*, pages 1–6, Paris, France, 2014.

[25] Z. Wang and AC. Bovik. Modern image quality assessment. *Synthesis Lectures on Image, Video, and Multimedia Processing*, 2(1):1–156, 2006.

[26] K. Jongyoo, N. Anh-Duc, and L. Sanghoon. Deep cnn-based blind image quality predictor. *IEEE Trans. on neural networks and learning systems*, 30(1):11–24, 2018.

[27] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein. Saliency in VR: How do people explore virtual environments? *IEEE Trans. on Visualization and Computer Graphics*, 24(4):1633–1642, 2018.

[28] S. Seo, S. Ki, and M. Kim. A novel just-noticeable-difference-based saliency-channel attention residual network for full-reference image quality predictions. *IEEE Trans. on Circuits and Systems for Video Technology*, 31(7):2602–2616, 2021.

[29] W. Sun, Z. Chen, and F. Wu. Visual scanpath prediction using ior-roi recurrent mixture density network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6):2101–2118, 2021.

[30] A. Sendjasni, MC. Larabi, and FA. Cheikh. Visual scanpath based data-augmentation for cnn-based 360-degree image quality assessment. In *LIM*, volume 2021, pages 21–26. IS&T, 2021.

[31] O. Russakovsky, J. Deng, H. Su, J Krause, and S. Satheesh et al. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3):211–252, 2015.

[32] V. Nair and G. Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010.

[33] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[34] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE ICCV*, pages 1026–1034, Santiago, Chile, 2015.

[35] H. Duan, G. Zhai, X. Min, Y. Zhu, Y. Fang, and X. Yang. Perceptual Quality Assessment of Omnidirectional Images. In *IEEE ISCAS*, pages 1–5, Florence, Italy, 2018.

[36] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.