Convolutional Neural Networks For Omnidirectional Image Quality Assessment: A Benchmark

Abderrezzaq Sendjasni, Mohamed-Chaker Larabi, Faouzi Alaya Cheikh

Convolutional Neural Networks For Omnidirectional Image Quality Assessment: A Benchmark

Abderrezzaq Sendjasni, *Student Member, IEEE,* Mohamed-Chaker Larabi, *Senior Member, IEEE,* Faouzi Alaya Cheikh, *Senior Member, IEEE,*

***Abstract*—In this paper, we conduct an extensive study on the use of pre-trained convolutional neural networks (CNNs) for omnidirectional image quality assessment (IQA). To cope with the lack of available IQA databases, transfer learning from seven pre-trained CNN models is investigated over retraining on standard 2D databases. In addition, we explore the influence of various image representations and training strategies on the model's performance. A comparison of the use of projected versus radial content, and multichannel CNN versus patch-wise training is also covered. The experimental results on two publicly available databases are used to draw conclusions about which strategy best fits the visual quality prediction and at which computational cost. The analysis shows that retraining CNN models on 2D IQA databases improves the prediction accuracy. The latter and the required computational time are found to be significantly affected by the training strategy. Cross-database evaluations demonstrate that the nature and variety of the content impact the generalization ability of the models. Finally, we show that conclusions coming from other image processing communities may not hold for IQA. The provided discussion shall provide insights and recommendations when using pre-trained CNNs for omnidirectional IQA.**

***Index Terms*—Benchmark, Convolutional neural networks, Omnidirectional images, Image quality assessment.**

I. Introduction

In recent years, virtual reality (VR) applications have known an impressive growth. It is used in many fields including education, entertainment, health-care, etc. One of the most used type of content for VR is omnidirectional images (*a.k.a.* 360-degree). It provides users with a visual experience of real-world scenes as well as synthesized ones. By means of commercial head-mounted displays (HMDs), users get an immersive experience, where they are able to explore the displayed content by head rotations in three degrees of freedom, *i.e.* pitch, yaw, and roll. This guarantees an interaction with the virtual environment by exploiting gaze and head movements. Since the viewing context for this type of content varies from traditional 2D, the user experience in terms of perception, expectation, immersiveness, possible sickness can be entirely different. In order to understand and improve the quality of experience (QoE) of such content, it is important to develop adapted paradigms and algorithms for images quality assessment (IQA).

IQA is an important topic in image processing and computer vision. It refers to the process of measuring the weighted combination of all visual attributes that reflect the perceived quality of a given image. This is performed in a way that should be consistent with human subjective opinion [1] considered as the ground truth obtained by means of psychophysical experiments. Depending on the availability of pristine images, IQA can be considered as full reference (FR) or no-reference (NR). It could be qualified as reduced reference (RR) if only partial information of the source image is used [2]. NR or blind approaches are widely adopted as they correspond to real-world scenarios where original images are most likely unavailable. At present, IQA is widely studied for 2D/3D content, and comprehensive reviews are provided in [3], [4]. However, for omnidirectional scenes, it is still in its infancy and not fully investigated. As the omnidirectional content is generally processed, encoded, and transmitted using a 2D plane representation, a straightforward solution is to use the large literature on 2D quality metrics directly on the 2D representation. Still, these metrics do not account for the non-uniform sampling density at pixel locations from the sphere to plane projection [5]. Their performances are lacking in terms of correlation with human quality judgment, as shown in [6]. Besides, most of the available 360-degree images are projected using the equirectangular (ERP) format. This projection suffers from geometric distortion and therefore, does not represent with fidelity the content viewed by the users. Having accurate metrics dedicated to omnidirectional images accounting for their characteristics either spherical- or projections-based becomes of a major importance in order to meet the challenges related to this type of content.

With omnidirectional images, a few IQA models have been proposed by extending traditional 2D metrics such as PSNR or MSE. For example, PSNR-based methods like Spherical PSNR (S-PSNR) [5] which computes the PSNR on a spherical surface instead of the 2D representation. The weighted spherical PSNR (WS-PSNR) [7] uses the scaling factor from a 2D plane to the sphere as a weighting factor for PSNR computation. CPP-PSNR [8] computes PSNR on the craster parabolic projection (CPP) after re-mapping 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 accurately predicting visual quality. One of the possible ways to reach reliable and accurate solutions for omnidirectional IQA, is the use of machine learning techniques.

Machine learning and more precisely deep learning is widely used for various image processing tasks in general and quality assessment in particular. Convolutional neural network (CNN) is a class of deep neural networks applied to

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

analyze and extract visual features from scenes. It is a trainable architecture inspired by biology that can learn invariant characteristics [9]. A CNN model can learn multi-level hierarchies of features. A better representation of such features, allows the development of more accurate IQA models [10]. This task was traditionally performed manually, resulting in handcrafted features, based on natural scene statistics and often distortion type specific. They are derived using different algorithms and represent details contained in visual contents, such as edges and shapes. Differently from handcrafting such features, CNNs learn how to extract and represent them automatically. By means of fusing and regressing extracted features, visual quality scores can be predicted for a given image [11], [12].

The absence of accurate and representative datasets and associated mean opinion scores to help in the design of IQA metrics and their validation is a significant issue when dealing with IQA for omnidirectional images. The construction of such databases require important efforts in terms of scenes acquisition, device calibration, paradigm definition, subjective testing and data analysis [6], [13], [14]. The use of well-known pre-trained models such as ResNet-18/34/50 [15], Vgg-16 [16] and DenseNet-121 [17] appears as good alternatives. This can be performed by the mean of transfer-learning (TL) as they have been trained for different tasks. In omnidirectional IQA literature, each work adopted and fine-tuned a well-known pre-trained model for a different task within the IQA framework. The main reason behind is that, the used models are trained on very large datasets, allowing them to reach a significant learning level. Besides, TL could be considered as a solution for the lack of data. However, at this stage several important questions could be raised regarding the use of pre-trained models on omnidirectional images, as well as exploiting the rich state-of-the-art work dedicated to 2D content, such as:

- Prediction accuracy of pre-trained models for omnidirectional images quality? And which model is performing the best?
- Radial vs. projected content-based training?
- Performance of projected format: CMP versus ERP?
- Performance of Patch-based training schemes?
- Would 2D quality databases improve the performance of CNN models?

In this paper, we intend to answer the above-mentioned questions by conducting an empirical and extensive analysis using different and widely used CNN models including ResNet-18/34/50 [15], Vgg-16/19 [16], DenseNet-121 [17] and Inception-V3 [18]. These models are compared under different configurations related to omnidirectional and spherical characteristics. The novelty of this work lies in the fact of providing answers to the above questions, for a very challenging type of content *i.e.* omnidirectional images, as one cannot rely on conclusions drawn from standard 2D benchmarks [19], [20], not taking into account the targeted characteristics.

II. Related Work

In this section, we provide a literature review of works featuring the use of different architectures of CNNs for IQA in general and omnidirectional IQA in particular

A. 2D-IQA

CNN models achieved great performances in various image processing tasks compared to conventional methods including object detection, classification, and segmentation. Foreseeing its remarkable performance, CNN-based approaches for IQA are proposed for 2D images. In particular, the authors in [19] benchmarked four well-known pre-trained CNN models *i.e.* Vgg-16/19 [16], ResNet-50 [15] and AlexNet [21] by means of transfer learning and fine-tuning. Vgg-16 outperformed the other models on 2D IQA databases, and was adopted as a baseline to their proposed architecture. Another comparison of pre-trained models was performed in [20] on 2D IQA databases, where eleven models (a mix of widely used and not very common models) such as AlexNet, ResNet-50/101, Vgg-16 and NASNet [22] were included. The NASNet model has been retained by the authors and fine-tuned for IQA tasks as it outperformed the selected pre-trained models. To the best of our knowledge, these are the only two works investigating the performances of pre-trained models for IQA. Besides, several authors have adopted such models as backbones of their approaches based on their popularity [23], [24], [25], while others opted for a self-defined model with a training from scratch [26], [27], [28]. However, available IQA databases are of insufficient size compared to the widely used image recognition databases [12]. Building large-scale perceptual quality databases is a much more difficult and time-consuming challenge than for other image processing tasks such as image classification, for example.

B. Omnidirectional IQA

Impressive results have been achieved by the aforementioned works on 2D content. Still, the conclusions made by the benchmarks in [19], [20] may not be applicable to omnidirectional images. This could be explained by the very specific nature of this type of content, introducing new features not existing for 2D. For instance, the equator bias explained by the attraction of the human gaze towards the equator [29] and the geometric distortions due to the projection of the sphere on a 2D plane. Furthermore, users with HMDs see just portions of the omnidirectional image known as viewports, which are represented by the actual displayed field of view (FoV) from the spherical representation. Their head movement determines the next viewport, making this exploration behavior unique to omnidirectional and VR environments.

Omnidirectional-IQA models are rather limited. In addition to the FR metrics mentioned previously, some learning-based models have been proposed in [30], [31], [32], [33], [34]. In these work, well-known CNN architectures such as Vgg [16] and ResNet [15] are used either as backbones or as part of the whole architecture. For instance, the authors in [30], use the ResNet architecture, where they propose a viewports-based approach with a multichannel CNN. First, they transpose the content from the equirectangular projection to the cube-map (CMP) projection. Then, each face of the CMP is used as a viewport, hence six viewports for each image. These viewports are used as input to six parallel ResNet-34 [15] sharing the same weights obtained based on the ImageNet

[35] database. Finally, the output of the six channels is fed to the image quality regressor, which concatenates the extracted features and derives a quality score. Here, the authors compute the quality on the projected format, which unfortunately does not represent the actual viewed content. In addition, the most important component in this model is the pre-trained ResNet that was originally trained on ImageNet. The latter dataset is composed of natural images with distortions occurring in the camera pipeline only. This would not allow the proposed model to predict visual quality for other distortions like compression, for instance. Another viewport based approach is proposed in [31]. Here, the authors take benefit from the spatial mutual dependencies among the extracted viewports by using a graph CNN. In addition, they use the Deep Bilinear CNN proposed in [25] to compute the global quality with a down-sampled ERP image as an input. Both outputs are combined to predict the quality score. The same shortcoming holds here as for the previous work. The selected regions are extracted from a projected format (ERP), which do not correspond to what a user would observe on the HMD. Kim *et al.* used the positions of selected patches from the ERP along with their content to estimate their weights [32]. A total of 32 patches is obtained. The position features are fused with visual features, extracted by the ResNet-50 model, to predict the quality score of the selected patches. The overall quality score of all 32 patches is then pooled and fed to a perception quality guider. This latter makes use of the reference image, putting *de facto* this approach in the FR class of models. The use of ERP images do not sound efficient as the content is geometrically distorted. Besides, 32 patches require 32 instances of ResNet-50 in parallel. This results in a much heavier model in terms of complexity. In response to these limitations, a CNN-based model is proposed in [33] to predict visual quality based on the spherical content of selected viewports rather than the projected one. Viewports are selected using visual scan-path predictions. Furthermore, the just-noticeable difference map is used to account for the perceptual characteristics of the human visual system (HVS) along with features produced from scan-paths in order to estimate the weight of each viewport. The visual features are extracted using the DenseNet-121 model [17]. Differently from the previous mentioned works, a patch-based approach is adopted in [34]. Patches are of size 64×64 and extracted from the ERP with a focus on the equatorial region. Each patch inherits the mean opinion score (MOS) of its 360-degree image, which may be questionable as it is too small to represent a 2K+ scene.

III. The Proposed Benchmark: Design and Architecture

With the intent to provide a holistic study as well as recommendations on the use of CNNs for omnidirectional IQA and to answer the questions raised in Sec. I, we designed a benchmark taking multiple considerations into account, related to the use of: 1) Content based splitting criteria for selecting training and validation sets, 2) Projected images as ERPs and CMPs, 3) Radial content rather than projected one, 4) multichannel CNN architecture, 5) Patch-based learning scheme, and 6) 2D benchmark IQA databases to train the selected models.

A. Pre-trained CNN models

In this study, seven among the widely used models are exploited and compared. A brief description of their architecture is provided below in addition to Table I giving their number of parameters and output size. All used models are fine-tuned by replacing the original top layers used for classification, by a quality regressor block (*see* Fig. 1). The latter is composed of: 1) a global average pooling (GAP) used to reduce the spatial dimensions of the extracted feature maps and to minimize overfitting, 2) a fully connected (FC) layer used with a rectified linear unit (ReLU) [36] activation function, 3) a dropout (DO) layer, which is a very effective regularization method to reduce overfitting and improve generalization error in deep neural networks [37], and 4) a FC layer with a single node and a linear activation function to deliver the quality score. The weights of the quality regressor are initialized according to [38]. During training, all layers of the pre-trained models are freezed to rely on the weights from ImageNet [35], and only the quality regressor block is trained for IQA.

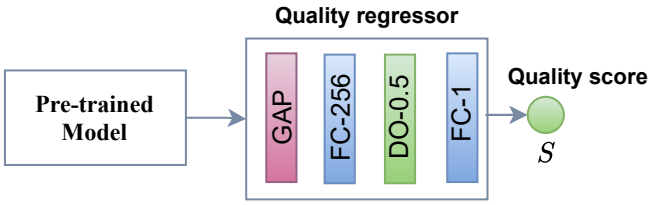

Fig. 1: Architecture of the CNN models: Top layers replaced by a regression block composed of a global average pooling (GAP) layer, a fully connected layer (FC), a dropout layer (DO) and a final FC layer to output the predicted score.

In the following, we describe the used CNN models by giving the most important characteristics, leaving the readers to refer to the original cited works.

ResNet : residual networks are artificial neural networks introduced in 2015 [15]. The ResNet utilizes skip connections to jump over some layers. This helps training deeper networks without falling into the problem of vanishing gradients. ResNet employs residual learning to further deepen the CNN network, which can be interpreted by a number of deeper bottleneck architectures. Each bottleneck has three convolutional layers with kernel dimensions of 1×1, 3×3 and, 1×1 respectively. A shortcut connection is then added from the input of the bottleneck to its output. Several versions of this model were developed with the main difference lying in the number of layers. We use ResNet-18/34/50 in this study.

VGG : it is a convolutional neural network architecture proposed in [16]. This network is characterized by its simplicity and use only 3×3 convolutional layers stacked on top of each other in an increasing depth. It also includes 1×1 convolution filters acting as a linear transformation of the input, followed by ReLU [36] activation. The convolution stride is fixed to 1 pixel, so to preserve the spatial resolutions. Different versions

of this network exists, but we only use widely used ones *i.e.* Vgg-16/19 [19], [20], [25], [31].

DenseNet : it is a neural network composed of dense blocks introduced in [17]. In each block, the layers are densely connected, with $L(L+1)/2$ direct connections, where L is the number of layers. Each layer in DenseNet receives additional input from all preceding layers and concatenates them with its own feature-maps before feeding them to the subsequent layers. This allows the model to reuse low-level features. The DenseNet-121 is considered in this study with the configuration used in [39].

Inception : this network architecture introduced in [18] is composed of convolutional blocks known as Inception modules. The latter contains 1×1, 3×3 and 5×5 convolutions as well as a pooling layer. The introduction of such module aims to allow for more efficient computation and deeper networks through a dimensionality reduction as well as the use of various convolutional filter sizes instead of using a single one. Several versions of this network also exist. The Inception-V3 model introduced factorized and smaller convolutions, helping to reduce the computational cost by decreasing the number of parameters involved in the network. This version of Inception is used in our comparative study.

TABLE I: Number of parameters (in million) in each selected model without their top layers, and the dimension of the output vector for feature representation (fv).

Model	#Params (M)≈	Size Output fv
ResNet-50	23	2048
ResNet-34	21	512
ResNet-18	11	512
Vgg-16	14	512
Vgg-19	20	512
DenseNet-121	7	1024
Inception-V3	21	2048

To effectively conduct transfer learning from another domain (*i.e.* image classification) using the abovementioned pre-trained models, fine-tuning is required. The latter allows removing the constraints on the label spaces of the source and target domains, *i.e.* from object classes to MOSs. Following the formulation suggested in [40], transfer learning can be expressed as follows:

$$f_s^* = \underset{f_s \in \mathcal{H}}{\operatorname{argmin}} \frac{1}{N_s} \sum_{i=1}^{N_s} l_s(f_s(x_s, i, q_s, i)) + \alpha R(D_s, f_s), \quad (1)$$

where (x_s, i, q_s, i) is the i-th tuple of the data sample and label of the source domain, N_s represents the number of samples in the source domain, $R(.)$ is a regularization term controlled by the weight α, and f_s is a function that lies in a Hilbert space $\mathcal{H}$. f_s is optimized by means of the loss function l_s using the data from the source domain D_s.

B. Omnidirectional IQA databases

To date, IQA of omnidirectional content is suffering from the unavailability of large, reliable and comprehensive databases. It is mostly due to the complexity and difficulty of the construction task. Indeed, building an IQA database requires subjective experiments to gather human opinions represented as MOS, in addition to an appropriate environment and test conditions. As a special case, for omnidirectional subjective tests, observers view the images using HMDs. These devices are far from offering a perfect representation of the omnidirectional content. They may introduce some defects like the screen door effect (SDE) with an impact on the quality ratings. Neglecting such phenomena may result in an unreliable evaluation. Another common issue with subjective scores in general is the non-linear nature of the obtained scores requiring a non-linear regression using a five parameter logistic function, as recommended in the ITU-R recommendations [41], prior to the performance evaluation. However, it is to be recalled that such a regression cannot be performed if the native correlation is below 0.7. Otherwise, the correlation value cannot be considered as reliable because the quality of regression would be very low. To date, five databases for omnidirectional IQA are proposed in the literature including MVAQD [42], Kim. *et al.* [32], CVIQ [13], OIQA [43] and Huang. *et al.* [14]. Unfortunately, only CVIQ and OIQA are publicly available and used for IQA models training. A previous comparative study in [6] showed a very low correlation between IQA metrics and MOS provided by Huang. *et al.* compared to CVIQ. This opens questions regarding the reliability of existing databases. Questions are still under investigation, especially regarding the use of HMDs for subjective experiments. It is a very delicate context as there are no recommendations nor guidelines on how to perform such experiments for omnidirectional applications. A study on this matter can be found in [44].

Our study is carried out using the CVIQ [13] and OIQA [43] databases containing ERP images. Details about each database are provided below, and samples are shown in Fig. 2. Table II summarizes these details in terms of number of reference and distorted images, distortion types, number of involved subjects, rating scale and the used HMD. There are some similarities across both databases in terms of number of participants, rating scales, and HMD. Because most of the important conditions are common, this will ease the analysis of the results.

CVIQ: it includes 16 pristine omnidirectional images and 528 impaired ones. Distortions used to create this database are only compression related, *i.e.* JPEG compression (JPEG) with quality factors ranging from 50 to 0 in addition to H.264/AVC (AVC) and H.265/HEVC (HEVC) with quantization parameter (QP) from 30 to 50. Eleven levels are used for each distortion. The authors used a single stimulus paradigm with a rating scale of 10 levels from the lowest to the highest quality to gather the MOS. 20 subjects participated to the construction of this database.

OIQA: it includes 320 distorted omnidirectional images obtained from 16 reference ones using four distortion types with five levels each, including JPEG compression (JPEG), JPEG 2000 compression (JP2K), Gaussian blur (BLUR) and Gaussian white noise (WN). Subjective scores are given in the range of 1 (bad) to 10 (excellent). 20 subjects were involved

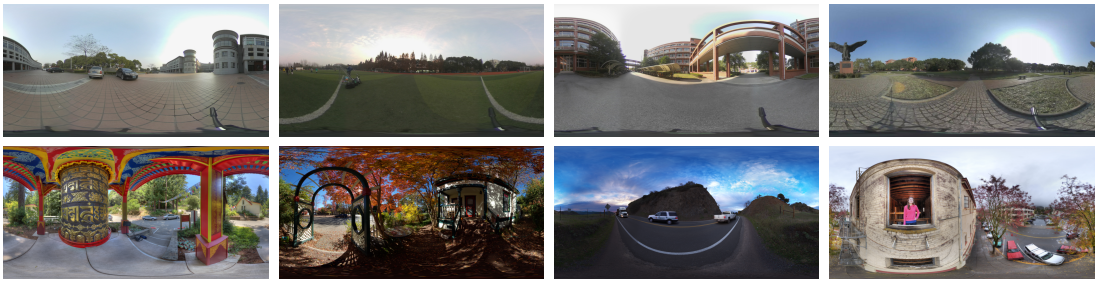

Fig. 2: Samples from the used databases: (top) CVIQ and (bottom) OIQA.

TABLE II: Summary of the used state-of-the-art omnidirectional image databases.

Database	Ref Images	Distorted Images	Distortion Types	Resolution	Subjects	Rating scale	HMD
CVIQ	16	528	JPEG / AVC / HEVC	4096×2048	20	10	HTC Vive
OIQA	16	320	JPEG / JPEG2k / GB / WGN	11332×5666 / 13322×6661	20	10	HTC Vive

in the test.

C. Content-based splitting strategy

Machine learning-based IQA tasks are typically learning a predictive model from quality assessment databases. When training data-driven models, one must ensure the accuracy, representativity, and reliability of the databases. Data biases are a major issue for learning-based IQA that is often overlooked. The consequences of such an issue are significant. It implies that, regardless of the used model, any computational prediction would have the same biases as the training data. Furthermore, the performance of a trained model is reported only on the testing set in which the selection may induce biases related to the content. A popular and straightforward approach is to split the training and testing sets based on pristine images. This means that the model is evaluated on unseen content independently of the existing distortions in the database. However, the obtained sets may lack diversity in terms of spatial complexity and colourfulness and may induce representativity biases, resulting in a test set that is not illustrative of the used database. Biases are mostly present, whether the data is split arbitrarily or based on more qualified criteria. However, minimizing those biases guarantees a validation on representative sets of the trained model.

For this benchmark, we first tackle the issue of content induced bias. To minimize such a bias, we use spatial information (SI) and colorfulness information (CFI) as criteria for the splitting strategy to make sure that, the performance of the models are reported on a limited-bias set of images. SI is an indicator of edge energy and therefore used to account for spatial complexity. The CFI is a perceptual indicator of the variety and intensity of colors in images. SI and CFI are computed according to the ITU-T P.910 recommendations [45] and the metric described in [46], respectively. Fig. 3 shows SI/CFI plots of pristine images on CVIQ and OIQA databases. As it can be seen, the variability of SI is higher in OIQA than in CVIQ, indicating that the latter database lacks diversity of content in terms of spatial complexity in comparison to OIQA. A similar conclusion holds in the case of CFI.

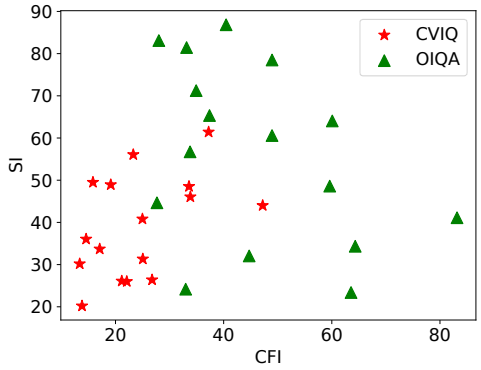

Fig. 3: Spatial information (SI) / colorfulness information (CFI) plot of pristine images in CVIQ and OIQA databases.

To select the training/testing sets, we used the Euclidean distance. For a couple of pristine images I_1 and I_2 characterized by (CFI_{I_1}, SI_{I_1}) and (CFI_{I_2}, SI_{I_2}) respectively, the distance $D(I_1, I_2)$ is expressed as follows:

$$D(I_1, I_2) = \sqrt{(CFI_{I_1} - CFI_{I_2})^2 + (SI_{I_1} - SI_{I_2})^2} \quad (2)$$

Based on the previous observation, we intend to demonstrate the existence of data biases when testing the effectiveness of a trained deep learning model on a given set from a database. The selection of the testing set influences the prediction correlation independently of the used database. Three splitting strategy are compared in this work. The first one is a random splitting by taking 20% on each iteration. The second one splits the databases in such a way that the images in the testing set are clustered in terms of SI/CFI (will be referred to as SI/CFI (a)). The third strategy takes the images that are the most spread-out in terms of SI/CFI (will be referred to as SI/CFI (b)). For all strategies, we ensure a complete separation of the training and testing sets, *i.e.*, the distorted images linked to the same pristine image are allocated to the same set.

D. Projection-based training

Within this configuration, we first investigate the use of ERP images as inputs to the selected models. It is rather straightforward and aims at evaluating CNN models on high-resolution ERP images. The input ERP are down-sampled into a resolution of 1024×512. This implies an adaptation of the model in order to match the shape of the input images. The output feature maps are provided to the quality regression block described in Sec. III-A. The use of ERPs as direct input may be thought of as estimating global quality rather than local to specific regions on the scene [31]. Despite the geometric distortions occurring on this type of projection, investigating the effect of using high resolution content with CNN models seems appropriate. Also, the models will learn from additionally distorted content (*i.e.* distortion from the databases as well as the projection-induced ones). Providing an analysis regarding the impact of the latter is within the scope of this benchmark. We will refer to this configuration as C_{ERP}.

In addition to the use of ERPs, we intend to provide a performance analysis on the use of cube-map projection (CMP). The CMP introduces less distortions compared to ERP. However, it provides separate content in form of cube faces. In fact, this projection requires a re-projection from ERP to CMP. It uses the six faces of a cube as the projection shape. The CMP is generated by first rendering the scene six times from a viewpoint. So, from each ERP image I, six faces are obtained $\{Left_I, Front_I, Right_I, Back_I, Top_I, Bottom_I\}$. An illustration of the re-projection is provided in Fig. 4.

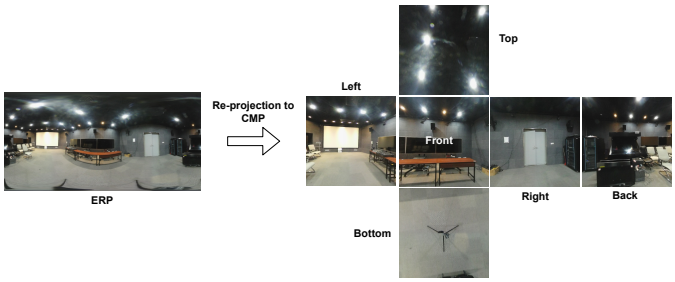

Fig. 4: ERP to CMP re-projection resulting in six faces: left, front, right, back, top and bottom.

One way to deal with the CMP as input to CNN models is to build a multichannel CNN as introduced in [30]. This way implies multiple CNNs in parallel where each is fed with one of the six obtained faces $\{Left, Front, Right, Back, Top, Bottom\}$. The output feature maps from these channels are concatenated, regressed and used to derive a quality score. The optimization of the model as well as the prediction is made on the six channels simultaneously and not individually. The use of CMP under a multichannel paradigm will be referred to as C_{CMP} and the adopted architecture is depicted in Fig. 6. A different way consists of taking each face as a separate content which involves a patch-based training scheme. Details on this approach are provided in Sec. III-F.

E. Radial-based training

As mentioned previously, the viewing experience of omnidirectional images is quite different from traditional ones. A user can only see the actual rendered FoV from the spherical representation. The next rendered FoV (viewport) is determined by his head movement around the x, y, and z axes. A slight head rotation will change the rendered viewport. The most important part of the actual viewed viewport is the content surrounding its center. Therefore, we only consider this latter to predict the quality on. To avoid any confusions, we will not call it viewports as it only represents a portion of it and, most of the time, this region is extracted as a square shape as in [30], [31], [33]. Indeed, a viewport is not square and using this term to describe square patches or regions could be misleading. As a result, we will refer to it as region.

By focusing on possible regions to predict the quality of omnidirectional images, we seek an agreement with the viewing experience of this kind of images. Also, in this case geometric distortions, created by the previously described sphere to plane projection, will be avoided. Another avoided type of distortion, is content discontinuity, artificial borders and oversampling created by the CMP projection [47]. This can lead to a loss of the semantic content. One solution to avoid such unwanted results, is the use of the radial content rather than the projected one. It can be done by mapping the ERP content to the sphere (*i.e.* from plane to 3D space). Then projecting back the viewed content, which consists of important regions from possible viewports, to 2D representation (*see* Fig. 5).

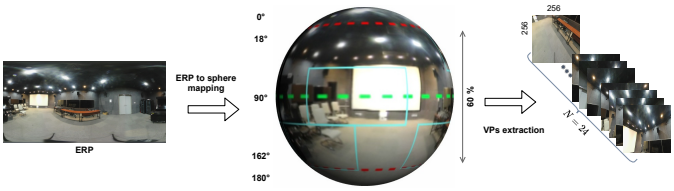

Fig. 5: Viewport selection for the spherical content configuration. Blue areas represent the selected viewports. In total, 24 regions surrounding the equatorial line (From $18°$ to $162°$) are extracted from the spherical content.

In addition to the used exploration behavior described above, it is now admitted that the human gaze is biased towards the equatorial line when viewing omnidirectional images [29]. Inspired by this and the fact that more than 30% of the content is often not viewed [48], we generate viewports surrounding the equatorial line representing more than 60% of the input content. Each center of a possible viewport R_i from the possible candidates k (up to $k = 24$) is extracted and projected from the spherical representation to the 2D plane. Then, the extracted contents are used as an input of a pre-trained model (among the seven selected networks). Similarly to C_{CMP}, this configuration implies a multichannel paradigm. The number of parallel channels depends on the number of extracted content. Accordingly, the complexity at this stage is proportionally increased. The output feature maps generated by the different channels are concatenated before feeding them to the quality regression block described in Sec. III-A. The training and

prediction flow is depicted in Fig. 6. We will refer to this as C_{Radial} in the remaining of the paper.

For this configuration, we first train the models with eights inputs before increasing their number by 8 until 24. This involves expanding the architecture of the models by adding more channels to fit the additional inputs. Such a strategy is motivated by the intent to analyze the impact of increasing inputs for a multichannel paradigm by finding the trade-off between accuracy of the models and the induced complexity.

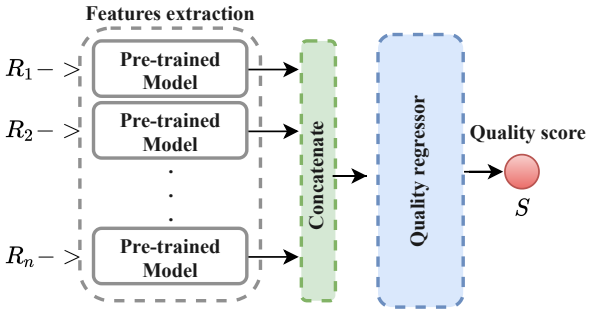

Fig. 6: Architecture of the multichannel CNN. Ri with $i \in \{1, 2, .., n\}$ stand for the extracted regions. Architecture adopted for C_{CMP} ($n = 6$) and C_{Radial} ($n \in \{8, 16, 24\}$).

F. Patch-based training

Differently from C_{CMP} and C_{Radial}, this configuration adopts a patch-based learning scheme. This means all considered regions from the omnidirectional images are seen as individual content, necessitating distinct labelling. Unfortunately, ground truth label (MOS) for individual patches are unavailable, since only the omnidirectional image-level ground truth MOS is provided. This heavily increases the challenge of IQA when adopting a patch-based training. A straightforward solution is to assign the same MOS of the omnidirectional image to the derived patches. This was first introduced in [12] and adopted by other researchers in [27], [34].

Within this configuration, two different approaches to extract patches from omnidirectional images are used. First, the regions extracted for C_{Radial} are considered as individual patches, 24 from each image. Second, the six faces from C_{CMP} where each face is taken as a separate patch. This configuration involves the use of a single channel CNN rather than a multichannel one. By using different approaches to extract patches, we aim to provide a better understanding on how the extraction method can influence the training and the prediction accuracy of the selected models. The quality score of the entire omnidirectional images is obtained by an average pooling of patches scores belonging to the same image. We will refer to this configuration as $C_{Patches}$.

G. Training on 2D IQA databases

The lack of databases for IQA of omnidirectional images hinders the promotion and development of CNN-based IQA models. In fact, designing a deep neural network and training it requires large-scale and representative databases. This is the main reason for adopting fine-tuning pre-trained models. Yet, pre-trained models have their limits, specially when used for a different task that may require specific type of ground truth. IQA is one of the most sensitive image processing tasks. The state-of-the-art for 2D IQA is well-developed compared to the omnidirectional one. Exploiting what exists may benefit to omnidirectional IQA. One of the aspects that we can exploit is 2D benchmark databases to train CNN models for IQA. Hence, the models weights will be optimized according to this specific task, from earlier layers to the top layers. A similar approach was used in [31] where they trained ResNet-18 [15] on the LIVE [49] database to further improve its accuracy. Other databases can be exploited to compensate for the lack of available data, such as the categorical image quality (CSIQ) database [50] and Tampere Image Database (TID2013) [51]. Other databases may also be useful.

In this study, two strategies of training on 2D-IQA databases are investigated. The first consists of training the selected models separately on each database. Each model is then trained from scratch using the ground truth provided by LIVE, CSIQ, and TID2013. The obtained knowledge is then transfered to omnidirectional-IQA by fine-tuning the obtained weights on CVIQ and OIQA. This way, all models are trained and fine-tuned for the same exact task. The second strategy consists of combining 2D-IQA databases in a large training dataset. This strategy is inspired by the work proposed in [52]. As combining IQA databases is rather difficult, requiring additional subjective experiments to ascertain the homogeneity of ratings according to the levels of degradation, the authors proposed a smart approach by using image pairing based on the Thurstone model. The ground truth labels are computed as the probability $P_{(x,y)}$ of the quality of x being higher than y, *i.e.* quality ranking task rather than visual quality prediction. By doing so, a large scale training dataset can be obtained. However, it requires a Siamese network with x and y as inputs and $P_{(x,y)}$ as outputs. The reader should refer to the original paper [52] for more details. After training the model on the combined dataset, the weights are saved and used to perform transfer learning on CVIQ/OIQA.

For training on 2D databases, we unfreeze the trainable layers of the used models. The new weights are optimized according to the regression of extracted features to visual quality scores for the first strategy, and quality rankings for the second one. The $C_{Patches}$ under CMP is used to perform the fine-tuning. For LIVE and TID2013, we cropped the regions surrounding the center with a resolution of 256×256 for all images, as they contain heterogeneous resolutions or rectangular shapes. This way, we avoid altering the content due to inappropriate resampling. Additionally, the input images are not normalized, which enables the proposed method to also cope with distortions introduced by luminance and contrast changes [27]. For both training starategies, all models are trained for 300 epochs and early stopping by monitoring the validation loss. We will refer to this configuration as C_{2D} in the following.

For the end-to-end training and transfer learning of all configurations, the error between predicted and target scores is computed using the L_2 loss function.

TABLE III: Characteristics of the used 2D IQA databases.

Databases	# of pristine images	# of distorted images	Distortion types	Levels of distortions
LIVE [49]	29	779	JP2K, JPEG, WN, GB, FF.	5
CSIQ [50]	30	866	JP2K, JPEG, WN, GB, FF, Contrast.	6
TID2013 [51]	25	3000	JP2K, JPEG, WN, GB and others.	5

IV. Results and discussion

A. Experimental setup

The proposed benchmark is implemented using TensorFlow [53]. The training of the considered configurations was performed on a server equipped with an Intel Xeon Silver 4208 2.1GHz CPU, 192GB of RAM, and an Nvidia Telsa V100S 32GB GPU. We use the RMSProp [54] optimizer for training the models. The learning rate is set to 0.001 with exponential decay. All models are trained with a batch size of 8 according to [55] for 25 epochs. We set the input dimension of all models to $(256, 256, 3)$ for the C_{CMP}, C_{Radial}, $C_{Patches}$ and C_{2D}. As for the C_{ERP}, we set it to $(1024, 512, 3)$.

The databases are split using the well-known Pareto principle and the criterion discussed in Sec. III-C, 80% is dedicated for training, and the remaining 20% for testing. For the sake of a fair comparison, all configurations were trained/tested using the same splitting scheme. Five-fold cross-validation is used for a complete evaluation within the selected database.

B. Data biases evaluation

To analyze the performance of splitting strategies and demonstrate the influence of content-induced biases, we compared three schemes as discussed in Sec. III-C. We trained the selected CNN models on ERP images for this assessment, since both databases come with this format. The performance results in terms of correlation accuracy (PLCC) and monotonicity (SRCC) are summarized in Table IV for both databases. The mean values of five-folds are given to provide a complete and fair assessment.

As can be observed, each splitting scheme resulted in a different performance, regardless of the database. This actively demonstrates the existence of biases when splitting databases for training and testing. Besides, it shows the impact of the used strategy on the reported validation performances. Since we are dealing with IQA which is a delicate task compared to classification or object detection for instance, one should consider the selection of a representative set of data for testing the efficiency of CNN models. Content representativeness for IQA may be expressed by a variety of attributes such as those we used, *i.e.* spatial complexity and colorfulness.

From Table IV, the random splitting scheme resulted in the best performance in terms of PLCC and SRCC for CVIQ. However, the observation is different for OIQA where the random splitting performance is outperformed by both SI/CFI based splitting schemes for all models. A possible reason could be related to the content composing each database, *i.e.* diversity of the images in terms of visual content and spatial complexity of the scenes. Based on this assumption and the previously discussed observation about the distribution

TABLE IV: Performance evaluation of the splitting strategies on CVIQ/OIQA databases. The best performing models are highlighted in **bold** for rows and underlined for columns. (a) and (b) stands for the SI/CFI-based schemes

		CVIQ						
		RNet-50	RNet-34	RNet-18	Vgg-16	Vgg-19	DNet-121	Incep-V3
Rand	PLCC	0.900	0.733	0.799	0.813	0.772	**0.903**	0.809
	SRCC	0.831	0.672	0.729	0.734	0.667	**0.832**	0.740
(a)	PLCC	0.844	0.727	0.787	0.789	0.791	**0.862**	0.810
	SRCC	0.770	0.639	0.728	0.734	0.696	**0.782**	0.707
(b)	PLCC	0.837	0.725	0.800	0.803	0.735	**0.858**	0.719
	SRCC	0.774	0.634	0.705	0.691	0.639	**0.767**	0.622
		OIQA						
Rand	PLCC	**0.749**	0.590	0.565	0.586	0.534	0.725	0.732
	SRCC	**0.710**	0.512	0.505	0.568	0.505	0.686	0.696
(a)	PLCC	**0.837**	0.641	0.803	0.662	0.608	0.818	0.803
	SRCC	**0.826**	0.624	0.771	0.610	0.592	0.783	0.775
(b)	PLCC	**0.899**	0.695	0.779	0.613	0.694	0.860	0.790
	SRCC	**0.877**	0.666	0.762	0.576	0.665	0.829	0.764

of the characteristics regarding CVIQ images (*see* Fig. 3), one can conclude that OIQA is more diverse than CVIQ. For the SI/CFI based splitting strategies, the (b) resulted in a more representative set since it selects diverse content and, intuitively represents approximately the used database in terms of content. In addition, it allows a more reliable evaluation of the performance accuracy within databases.

As observed for CVIQ, the random splitting scheme resulted in the best performance overall, except for ResNet-18 and Vgg-19. We believe that it is strongly related to the nature and diversity of the content. Additionally, between SI/CFI-based strategies, a very slight difference can be observed for CVIQ as they appear to be competing with each other. The opposite is observed on OIQA, where a noticeable difference can be reported in terms of correlation and monotonicity. On the same database, the SI/CFI (b) resulted in a better performance compared to the (a) strategy. Based on the above observations, we adopt the SI/CFI (b) strategy to train/test the considered configurations. By doing so, the performances will be reported on the most representative sets of the selected databases.

C. Projection-based evaluation

1) C_{ERP}: To assess the performances of selected pre-trained models on high-resolution ERP images, we provide in Table IV (SI/CF based splitting (b)) the PLCC and SRCC scores obtained for both databases. Knowing that no omnidirectional peculiarities have been considered with this configuration, its performances are still satisfactory for almost all

models. On average, the best performing model within this strategy is ResNet-50 followed by DenseNet-121, while the least performing one is Vgg-19. This is valid for both accuracy (PLCC) and monotonicity (SRCC) of the predictions. In fact, ResNet-50 obtained a PLCC (resp. SRCC) value of 0.844 (resp. 0.770) on CVIQ and 0.899 (resp. 0.877) on OIQA. DenseNet-121 achieved 0.862 (resp. 0.782) on CVIQ and 0.860 (resp. 0.829) on OIQA. These two models outperformed the other CNN models, regardless the used database. ResNet-50 is more popular compared to DenseNet-121, especially within the IQA community. DenseNet-121 model is under-represented for IQA tasks, and most of the recent works adopted either ResNet or Vgg [19], [25], [32] as backbones. These choices are often made based on previous conclusions derived from other image processing tasks.

Comparing the results on CVIQ and OIQA, one can notice better correlations on OIQA, supporting the previous assumption on the nature of content in this database. The diversity of content helps models to better train and generalize. However, with the C_{ERP}, only 422 images are used for fine-tuning on CVIQ and 256 on OIQA. The more diverse the training data, the more examples to train on are required. Therefore, one can conclude that achieving a siginificant generalization ability on diversified databases requires larger training sets.

TABLE V: Performance evaluation of cross-database validation under the C_{ERP}. Best performing models in **bold**.

Train/Test Dist.		Metric	RN-50	RN-34	RN-18	Vgg-16	Vgg-19	DN-121	Inc-V3
OIQA / CVIQ	Overall	PLCC	**0.820**	0.403	0.410	0.309	0.485	0.750	0.687
		SRCC	**0.751**	0.305	0.437	0.254	0.485	0.716	0.651
	JPEG	PLCC	**0.903**	0.339	0.360	0.436	0.556	0.813	0.761
		SRCC	**0.751**	0.250	0.325	0.331	0.540	0.717	0.644
	AVC	PLCC	**0.811**	0.434	0.447	0.320	0.521	0.695	0.681
		SRCC	**0.769**	0.396	0.423	0.314	0.497	0.672	0.653
	HEVC	PLCC	**0.741**	0.432	0.604	0.215	0.385	0.731	0.633
		SRCC	0.700	0.401	0.588	0.200	0.374	**0.713**	0.618
CVIQ / OIQA	Overall	PLCC	**0.476**	0.256	0.268	0.295	0.320	0.472	0.474
		SRCC	**0.433**	0.279	0.256	0.304	0.302	0.386	0.431
	JPEG	PLCC	**0.768**	0.264	0.404	0.324	0.281	0.754	0.285
		SRCC	**0.762**	0.326	0.351	0.345	0.178	0.732	0.278

We conducted a cross-database validation to provide a better understanding of the use of ERP images with CNN models. We first trained the models on OIQA before testing their performance on CVIQ and *vice versa*. The performance results are summarized in Table V in terms of PLCC and SRCC. We provide results for both the overall databases and on individual distortions. For training on CVIQ and testing on OIQA, we only provide results on the JPEG distortion according to [30].

Despite the satisfactory results obtained by the selected models on each database separately, one can observe very low performances on the cross-database validation. This depicts the limitation of C_{ERP} when used with different CNN architectures, except for ResNet-50 and DenseNet-121. The latter models achieved good results in both cases. Training on OIQA and testing on CVIQ gave better results compared to the reverse case. One can observe a PLCC (resp. SRCC) value of 0.820 (resp. 0.751) on the overall database obtained by ResNet-50 when trained/tested on OIQA/CVIQ compared to 0.476 (resp. 0.433) when trained/tested on CVIQ/OIQA. A similar behavior is noticed with DenseNet-121 and the other models. This could be explained by the heterogeneousness of the distortions in OIQA, combining compression artifacts with Gaussian blur and white noise. Testing the performances of fine-tuned CNN models, primarily trained for classification on unseen distortions, resulted in poor performances. Among the models, the performances of ResNet-50 show a significant difference compared to ResNet-18/34 and Vgg-16-19. A possible explanation could be in the fine-tuning strategy [56]. It is known that the hyperparameters are key factors in achieving the best performance. These parameters are usually tuned according to the model, its architecture and depth, and the training datasets. However, as the focus of the study is rather benchmarking omnidirectional related configurations, the used hyperparameters are fixed for all models.

When comparing the performance on individual distortions, it can be seen that training on OIQA yields better results. Even though JPEG is present in both databases, training on OIQA resulted in significantly higher PLCC and SRCC scores. This finding holds for all seven models. A possible explanation is that the levels of JPEG distortion applied in both databases are different (five in OIQA and eleven in CVIQ). Still, the same class of artifacts should not result in such a significant difference. Perhaps compressing with eleven levels is not the best option because it results in less discernible differences between some stimulus (impaired images). When tested on CVIQ, Resnet-50 and DenseNet-121 also performed well regarding AVC and HEVC. These distortions are not available in OIQA, demonstrating the efficacy of these models in generalizing to comparable distortions.

2) C_{CMP}: With the intent to provide a comparison of pre-trained models' performances when used on CMP projection format and assess the influence of this type of projection, we provide in Table VI results in terms of PLCC and SRCC. Overall, the prediction performances are more correlated on CVIQ compared to OIQA. This is because CVIQ contains only compression artifacts, while OIQA contains various ones. The diversity of distortions may lead to a less generalized correlation across the entire database. This observation is applicable irrespective of the architecture of the model.

TABLE VI: Performance evaluation in terms of PLCC and SRCC of pre-trained models using C_{CMP}. Best performances are highlighted in **bold** for each database.

Database	Metric	RNet-50	RNet-34	RNet-18	Vgg-16	Vgg-19	DNet-121	Inc-V3
CVIQ	PLCC	**0.835**	0.751	0.786	0.743	0.776	0.825	0.739
	SRCC	**0.814**	0.657	0.760	0.726	0.714	0.730	0.653
OIQA	PLCC	**0.775**	0.562	0.583	0.493	0.493	0.673	0.607
	SRCC	**0.722**	0.532	0.561	0.498	0.498	0.596	0.548

From Table VI, it can be noticed that ResNet-50 outperforms the other models in terms of prediction accuracy and monotonicity on both databases. DenseNet-121 ranked second, but as it has fewer parameters compared to ResNet-50 (*see*

Table I). Its performance can be considered as a trade-off between accuracy and complexity. One can also observe that Vgg-16 and Vgg-19 performed the worst among the seven models on OIQA. It is also the case of Inception-V3 on CVIQ.

TABLE VII: Performance evaluation of cross database validation under the C_{CMP}. Best performing models in **bold**.

Train/Test Dist.		Metric	RN-50	RN-34	RN-18	Vgg-16	Vgg-19	DN-121	Inc-V3
OIQA / CVIQ	All	PLCC	**0.804**	0.429	0.598	0.400	0.106	0.697	0.374
		SRCC	**0.738**	0.308	0.566	0.389	0.080	0.625	0.394
	JPEG	PLCC	**0.914**	0.525	0.649	0.381	0.206	0.867	0.617
		SRCC	**0.819**	0.318	0.469	0.283	0.208	0.752	0.472
	AVC	PLCC	**0.743**	0.464	0.680	0.552	0.188	0.598	0.349
		SRCC	**0.705**	0.371	0.641	0.539	0.127	0.551	0.407
	HEVC	PLCC	**0.703**	0.382	0.690	0.442	0.314	0.506	0.385
		SRCC	**0.647**	0.247	0.673	0.448	0.256	0.498	0.376
CVIQ / OIQA	All	PLCC	0.304	0.252	0.308	0.211	0.172	**0.487**	0.227
		SRCC	0.287	0.261	0.306	0.149	0.158	**0.431**	0.228
	JPEG	PLCC	0.506	0.227	0.405	0.409	0.254	**0.687**	0.484
		SRCC	0.470	0.233	0.346	0.407	0.250	**0.649**	0.388

A cross-database assessment was performed using CVIQ and OIQA to demonstrate the generalization ability of selected pre-trained models under the CMP configuration. Firstly, we trained the models on OIQA and tested them on CVIQ. The performance results on the overall database as well as per distortion types are provided in Table VII. As it can be seen, the performances on the overall database are below 0.7, except for ResNet-50 and DenseNet-121 when tested on JPEG, achieving second-best performance. Is it worth mentioning that, none of the models were dedicated to quality assessment as they were trained on ImageNet. Only the regression block is trained for the IQA task. Besides, the only common distortion between OIQA and CVIQ is JPEG. This is reflected in the same table, where an improvement of PLCC and SRCC for JPEG could be observed compared to the overall performance. The correlation performances shifted from 0.80 to 0.91 for ResNet-50, and from 0.69 to 0.86 for DenseNet-121. The performances of the other models improved as well, but remains below the 0.7 threshold. Regarding AVC and HEVC distortions, the performances dropped compared to JPEG and even to the overall scores, yet still acceptable.

Then, we trained on CVIQ and tested on OIQA. The correlation results are summarized in the lower part of Table VII. One can observe low performances compared to previous results. The training on CVIQ seems to lead to less generalize models. The overall performances are very low as the models are trying to predict on unlearned distortions (*i.e.* WGN and GB). Besides, the performances on JPEG are low too compared to those obtained when trained on OIQA. Despite the low performances, the contrast to training on OIQA regarding the best-performing model can be noticed. The DenseNet-121 outperformed the other models, even ResNet-50.

D. Radial-based evaluation

In this section, we discuss the performance evaluation of the radial content-based configuration C_{Radial}. Table. VIII gathers the scores for CVIQ and OIQA. The previous observation regarding the best-performing models is still valid for this configuration. Overall, ResNet-50 and DenseNet-121 performed the best, with DenseNet-121 ranking first on CVIQ and ResNet-50 on OIQA.

Overall, one can notice that the performances obtained on CVIQ are mostly better compared to OIQA. A minimum PLCC (resp. SRCC) value of 0.72 (resp. 0.69) is obtained on CVIQ, while 0.39 (resp. 0.36) on OIQA. The reason might be the distortions contained in OIQA. With this configuration, additional distortions due to projection can be avoided. Therefore, the reported results are more representative as they were obtained based on the actual viewed content.

The fact of increasing the number of inputs improved the performances of the selected models. One can notice that, in average, the performance increases with increased inputs for all models, and declines with $R = 24$ for Resnet-34 and Inception-V3. This behavior does not apply to ResNet-18, as we notice (the best score with $R = 8$) and then a decrease with additional inputs. This actively demonstrates an overfitting behavior. A similar behavior is shown by Vgg-16. Increasing the number of inputs leads to a higher number of channels where CNN models become more prone to overfitting. It is worth mentioning that the variation of the number of regions results in a variation of the quality prediction accuracy as well as the prediction monotonicity.

E. Patch-based evaluation

To compare the performance of the multichannel paradigm versus the patch-wise training scheme, we trained the selected models using the output of the CMP as patches in addition to the regions generated for C_{Radial} (*see* Sec. III-F). Table IX summarizes the obtained results.

In average, the radial-based method performed better. A PLCC (resp. SRCC) value of 0.821 (resp. 0.760) on CVIQ and 0.778 (resp. 0.756) on OIQA compared to 0.800 (resp. 0.751) and 0.708 (resp. 0.668) with CMP on CVIQ and OIQA respectively. A PLCC difference of approximately 2.6% on CVIQ and 9.4% on OIQA is observed when using patches obtained on the sphere. This illustrates the usefulness of using radial content against the projected one. Another possible reason is the number of extracted patches, providing the models with more training examples. Looking into individual performances, DenseNet-121 ranked the best for radial patches and ResNet-50 for CMP patches. Despite the heterogeneity of the distortions on OIQA compared to CVIQ, training the DenseNet-121 using a patch-wise lead to a better accuracy. Another noteworthy observation is related to the Vgg-16/19 performances. They achieve comparable performance to ResNet-50 for radial configuration on CVIQ. Knowing that the pre-trained version of Vgg-16/19 scored among the worst in the previous configurations, their performances under $C_{Patches}$ prove to be satisfactory.

An in-depth analysis shows a significant difference in terms of performances among different models. This could be related to the pre-trained version of the models. Performing transfer learning with various amount of training examples

TABLE VIII: Performance evaluation of pre-trained models with the C_{Radial} on CVIQ/OIQA databases in terms of PLCC/SRCC. Best performing model is highlighted in **bold** for CVIQ and underlined for OIQA.

# inputs	Metric	ResNet-50		ResNet-34		ResNet-18		Vgg-16		Vgg-19		DenseNet-121		Inception-V3	
		CVIQ	OIQA	CVIQ	OIQA	CVIQ	OIQA	CVIQ	OIQA	CVIQ	OIQA	CVIQ	OIQA	CVIQ	OIQA
R = 8	PLCC	0.788	0.795	0.720	0.606	0.829	0.688	0.801	0.387	0.801	0.567	**0.841**	0.772	0.757	0.704
	SRCC	0.713	0.758	0.698	0.559	**0.794**	0.661	0.707	0.361	0.707	0.517	0.747	0.727	0.740	0.644
R = 16	PLCC	0.807	0.842	0.770	0.544	0.751	0.725	0.772	0.482	0.809	0.482	**0.857**	0.822	0.793	0.791
	SRCC	0.747	0.816	0.689	0.530	0.722	0.705	0.700	0.508	0.716	0.508	**0.769**	0.780	0.755	0.752
R = 24	PLCC	0.830	0.851	0.726	0.663	0.764	0.769	0.802	0.394	0.821	0.608	**0.859**	0.890	0.775	0.749
	SRCC	0.781	0.809	0.687	0.629	0.740	0.740	0.747	0.389	0.743	0.623	**0.782**	0.876	0.722	0.723

TABLE IX: Performance evaluation of pre-trained models with the $C_{Patches}$ on CVIQ/OIQA database in terms of PLCC, SRCC. Best performances are highlighted in **bold** for rows and underlined for columns.

Input type	Metric	ResNet-50		ResNet-34		ResNet-18		Vgg-16		Vgg-19		DenseNet-121		Inception-V3		Average	
		CVIQ	OIQA	CVIQ	OIQA	CVIQ	OIQA	CVIQ	OIQA	CVIQ	OIQA	CVIQ	OIQA	CVIQ	OIQA	CVIQ	OIQA
Radial	PLCC	0.861	0.836	0.604	0.671	0.792	0.787	0.861	0.816	0.859	0.771	**0.907**	**0.925**	0.864	0.641	0.821	0.778
	SRCC	0.820	0.810	0.533	0.626	0.716	0.774	0.816	0.787	0.791	0.734	**0.851**	**0.917**	0.792	0.640	0.760	0.756
CMP	PLCC	**0.857**	**0.867**	0.756	0.678	0.827	0.755	0.795	0.557	0.785	0.552	0.807	0.845	0.772	0.701	0.800	0.708
	SRCC	**0.833**	**0.848**	0.725	0.667	0.783	0.713	0.711	0.465	0.709	0.518	0.768	0.818	0.726	0.645	0.751	0.668

is affecting the deeper and shallower models in different ways. For instance, deeper models such as DensNet-121 and Vgg-16/19 achieved good performances with radial compared to CMP. Whereas, with ResNet-18/34/50 the reverse can be observed. The models are fine-tuned using augmented databases of varying sizes. The radial configuration generates 10128 patches on CVIQ (resp. 6144 on OIQA) while the CMP configuration generates 2532 (resp. 1536 patches). Four times less the amount of training data is generated with CMP compared to the radial starategy with an impact on the model's achievable performances. This configuration and the training strategy appear to influence different backbones in different ways.

We performed a cross-database validation with the patch-wise configuration to verify the generalization ability of the selected models when trained using a patch-wise scheme. The performance results are gathered in Table X. The same observation regarding the best performing model still valid, ResNet-50 and DenseNet-121 achieved the best performance overall and per-distortion. Good results are obtained on the JPEG distortion with a PLCC (resp. SRCC) values of 0.94 (resp. 0.86) using radial patches, and 0.92 (resp. 0.83) using CMP patches by ResNet-50 when trained on OIQA and tested on CVIQ. Parallelly, satisfying results are obtained on AVC and HEVC distortions. Vgg-16/19 and Inception-V3 achieved satisfactory results with radial when trained on OIQA. PLCC/SRCC values above 0.90/0.80 on JPEG are obtained. One can also observe that the models trained on OIQA demonstrate a stronger generalization ability when tested on CVIQ compared to the opposite. This supports the previous observation concerning the richness of OIQA versus CVIQ in terms of content and distortions. Besides, the radial-based method resulted in the best performance compared to CMP regardless of the used database. This depicts the importance of using radial rather than projected content on the one hand. On the other hand, generating more examples for the models to train on, improved the accuracy, demonstrating the impact of having a large amount of data.

F. Training on 2D IQA databases evaluation

With the intent to evaluate whether the use of 2D IQA databases improves the performance of CNN models compared to performing transfer learning, we trained all selected models on LIVE, CSIQ, TID2013, and combined databases (All). As it is known, deep neural networks require large-scale databases in order to achieve better accuracy as well as avoiding over-fitting. To analyze the learning behavior, we provide in Fig. 7 the contrast $(val\_loss - loss/val\_loss + loss)$ between training and validation losses for the five folds (F-1 to F-5). A contrast equal to 0 depicts an equal loss between training and validation, whereas a contrast equal or close to 1 suggests an important gap between both losses, with val_loss being higher and the opposite if equal or close to -1. We can see that training on LIVE is leading to a better generalization to the prediction of MOS, but with a non-smooth behavior. Training on the combined datasets (All) led to the best behavior of the training losses, as the contrast is stable and close to 0. This is a generalization to predict the probability ranking, as discussed in Sec. III-G, suggesting a robust performance during training. Training on TID2013 has a higher contrast, meaning that the models have difficulty to generalize. The gap between training and validation losses is much higher than those of LIVE and CSIQ. This is also demonstrated by the provided curves (*see* Fig. 7). A possible reason is that TID2013 contains many diverse distortions, a total of 24 types, and it may need more examples to learn from in order to demonstrate a better generalization ability. From the provided curve, we can notice that the progress of training/validation loss is more stable on TID2013, despite the previous observation. This led to a quicker convergence for all models. Indeed, training on TID2013 required fewer epochs when compared to training on CSIQ, LIVE, and the combined datasets. Among the models,

TABLE X: Performance evaluation of pre-trained models with the $C_{Patches}$ on CVIQ/OIQA database in terms of PLCC, SRCC. Best performances are highlighted in **bold** for rows and underlined for columns.

Training / Testing	Distortion	Metric	ResNet-50		ResNet-34		ResNet-18		Vgg-16		Vgg-19		DenseNet-121		Inception-V3	
			Radial	CMP	Radial	CMP	Radial	CMP	Radial	CMP	Radial	CMP	Radial	CMP	Radial	CMP
Train: OIQA & Test: CVIQ	Overall	PLCC	**0.886**	**0.843**	0.705	0.593	0.637	0.607	0.789	0.710	0.767	0.665	0.859	0.841	0.764	0.684
		SRCC	**0.846**	0.790	0.672	0.552	0.560	0.547	0.734	0.682	0.711	0.652	0.810	**0.791**	0.721	0.631
	JPEG	PLCC	**0.948**	0.928	0.834	0.709	0.735	0.755	0.929	0.854	0.905	0.786	0.945	**0.929**	0.901	0.846
		SRCC	**0.865**	**0.835**	0.721	0.567	0.569	0.535	0.816	0.761	0.795	0.719	0.845	0.817	0.819	0.737
	AVC	PLCC	**0.846**	**0.782**	0.722	0.656	0.620	0.678	0.726	0.715	0.645	0.657	0.800	0.775	0.676	0.625
		SRCC	**0.842**	**0.768**	0.705	0.622	0.599	0.616	0.704	0.691	0.642	0.645	0.780	0.754	0.651	0.586
	HEVC	PLCC	**0.814**	0.721	0.645	0.570	0.546	0.640	0.632	0.599	0.649	0.648	0.749	**0.733**	0.648	0.524
		SRCC	**0.804**	0.716	0.636	0.521	0.522	0.575	0.627	0.579	0.646	0.659	0.730	**0.728**	0.625	0.509
Train: CVIQ & Test: OIQA	Overall	PLCC	0.558	0.441	0.336	0.453	0.469	0.455	0.318	0.330	0.436	0.309	**0.570**	**0.598**	0.469	0.497
		SRCC	**0.534**	0.430	0.309	0.453	0.468	0.439	0.287	0.320	0.370	0.270	0.523	**0.564**	0.438	0.419
	JPEG	PLCC	**0.858**	0.631	0.484	0.363	0.720	0.492	0.637	0.587	0.709	0.631	0.855	**0.825**	0.759	0.700
		SRCC	0.790	0.554	0.414	0.297	0.679	0.423	0.579	0.533	0.664	0.554	**0.801**	**0.723**	0.724	0.607

TABLE XI: Performance evaluation of C_{2D} in terms of PLCC, SRCC. The best performing models are highlighted in **bold** for rows and underlined for columns. 'All' stands for combined datasets.

2D database	Metric	ResNet-50		ResNet-34		ResNet-18		Vgg-16		Vgg-19		DenseNet-121		Inception-V3	
		CVIQ	OIQA	CVIQ	OIQA	CVIQ	OIQA	CVIQ	OIQA	CVIQ	OIQA	CVIQ	OIQA	CVIQ	OIQA
LIVE	PLCC	0.915	0.884	0.910	0.920	0.898	**0.923**	0.877	0.610	0.861	0.805	**0.948**	0.837	0.897	0.905
	SRCC	0.852	0.873	0.847	0.907	0.836	**0.912**	0.804	0.587	0.791	0.782	**0.918**	0.827	0.832	0.893
CSIQ	PLCC	0.912	0.876	0.934	0.923	0.903	0.919	0.873	0.832	0.892	0.847	**0.943**	**0.931**	0.908	0.911
	SRCC	0.849	0.865	0.887	0.914	0.839	0.910	0.810	0.813	0.823	0.833	**0.906**	**0.921**	0.886	0.898
TID2013	PLCC	0.909	0.879	0.918	**0.905**	0.898	0.894	0.849	0.816	0.813	0.721	0.906	0.880	**0.923**	0.853
	SRCC	0.847	0.863	0.859	**0.889**	0.829	0.878	0.768	0.799	0.737	0.704	0.842	0.858	**0.875**	0.833
ALL	PLCC	**0.938**	0.882	0.728	0.381	0.821	0.342	0.889	0.804	0.886	0.692	0.869	**0.908**	0.751	0.696
	SRCC	**0.895**	0.862	0.669	0.371	0.774	0.322	0.832	0.773	0.841	0.679	0.838	**0.917**	0.735	0.834

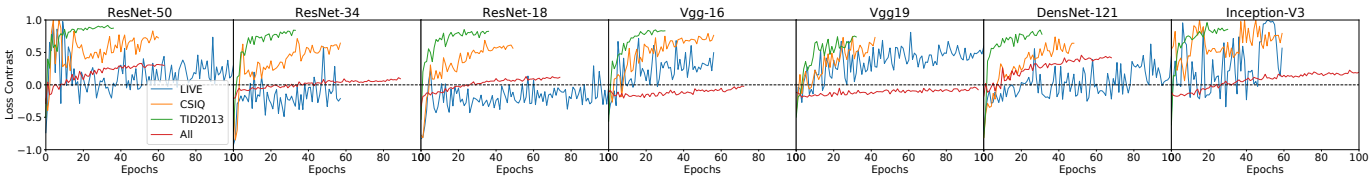

Fig. 7: Contrast ($val\_loss - loss / val\_loss + loss$) between training and validation losses for all models trained on 2D-IQA databases ($0 \rightarrow$ equal loss between training and validation losses). 'All' stands for combined datasets.

ResNet-34 converges quicker on each database, followed by Vgg-16.

In addition to the training behavior shown above, we provide the performance accuracy of the weights obtained from training on 2D databases, *i.e.* LIVE, CSIQ, and TID2013 individually and combined together. Table XI gathers the performance results on CVIQ/OIQA in terms of PLCC/SRCC. On average, the performances are quite satisfactory on both databases. A maximum PLCC (resp. SRCC) value of 0.948 (resp. 0.921) is achieved by DenseNet-121. Overall, the latter scored the best among the selected models when trained on each database separately and ResNet-50 when trained on the combined datasets. Training the models on 2D IQA databases appears to improve their performances in both correlation accuracy and monotonicity. The achieved efficiency is competitive except for Vgg-16 on OIQA when trained on LIVE. Despite the small size of IQA databases, the obtained performances actively demonstrate the usefulness of training CNN models on them. Acquiring knowledge about quality after being pre-trained to predict it is increasing the performances.

Among the selected 2D databases, training on TID2013 results in a poor performance compared to LIVE and CSIQ. It could be explained by the lack of generalization due to the limited number of instances per distortion. Indeed, insufficient amount of data may lead to overfitting, especially when training from scratch. When trained on LIVE and TID2013, ResNet-50 is outperformed by its smaller variants, ResNet-18/34. The difference is greater on OIQA than on CVIQ, which solely contains compression artifacts. It is known that deeper models require large databases in order to reach a generalization ability and sufficient accuracy, especially on databases with diverse content.

Regarding pre-training on the combined datasets, some improvement can be observed when compared to the use of imageNet weights reported in Table IX. For example, ResNet-50 performances in terms of PLCC/SRCC shifted from

0.857/0.833 on CVIQ to 0.938/0.895, representing 9%/7.2% of improvement. On the same database, similar improvements can be seen with other models such as Vgg-16/19, as well as slight improvements with DenseNet-121, Inception-V3, and ResNet-18. Analyzing the performance on OIQA, ResNet-18/34 performed poorly, with accuracy and monotonicity scores below 0.5. As demonstrated by the lower performances, fine-tuning on databases with diverse content and degradation is less efficient for different models with varying depths. This indicates less generalization compared to training on individual databases.

G. Computational complexity

With the aim to compare the computational complexity of the selected models under different configurations, we measure the required prediction time per input image. Since the inference analysis is independent from the training, we used a different hardware configuration. A computer equipped with an Intel® Core™ i9-9880H @ 2.30GHz, 32GB of RAM, and an Nvidia Quadro T2000 MAX-Q 4GB GPU is selected to measure the computational complexity. Fig. 8 represents the average of the computational time required over ten images. Overall, DenseNet-121 requires the longest time, followed by Vgg-16/19. Considering this, one can conclude that DenseNet-121 and the Vgg-based models are heavier in terms of computational complexity, followed by Inception-V3, and finally ResNets. The training time is definitely not proportional to the complexity of the used model in terms of number of parameters (*see* Table I). DenseNets concatenations require high GPU memory and therefore more training time [57], while ResNet models implement skip-connections, allowing to jump over some layers and reducing the computational time [15], [58]. Despite the number of parameters of ResNet-50, the latter spent less time than VGG-16/19. DenseNet-121 has fewer parameters among the selected models, and yet it requires more time than ResNets.

Among the configurations, the multichannel appears to demand more computational time, except with Vgg-16/19. The computational time required by Vggs is highly impacted by the input shape. As it can be seen, Vgg-16/19 required the longest time when used with ERP images, suggesting that the architecture of the model plays a major role in the computational complexity.

In addition to the computational time, we measured the number of floating-point operations (FLOPs) with regard to the input shape. The latter determines the number of FLOPs providing insight on the computations required by the model. A large FLOPs number implies a higher complexity, suggesting a longer calculation time. The FLOPs are reported in Table XII. One can observe that having an input shape of 1024*512*3 resulted in larger FLOPs. However, according to the computational time associated with each configuration, the first observation that emerges is that the FLOPs is not proportional with the required computational time. This could be explained by the fact that other operations are involved, especially memory-based ones, as discussed previously regarding DeseNet-121. In addition, some architectures implement

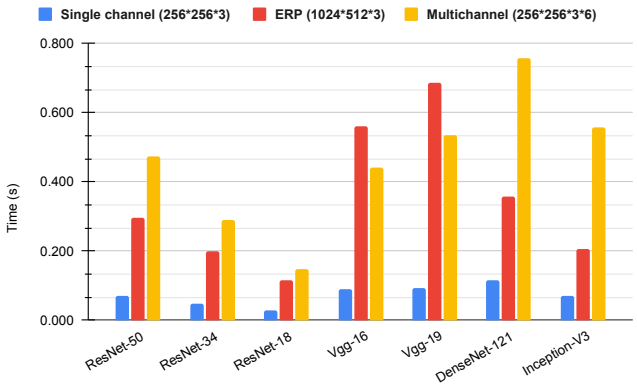

Fig. 8: Computational complexity in terms of required prediction time per image.The average over ten samples is provided.

TABLE XII: The number of FLOPs (in billion) with regard to the input shapes.

Input shape	256*256*3	1024*512*3	256*256*3*6
ResNet-50	5.04	40.35	30.27
ResNet-34	4.80	38.42	28.82
ResNet-18	2.39	19.09	14.32
Vgg-16	20.1	160.5	120.3
Vgg-19	25.5	203.9	152.9
DenseNet-121	3.70	29.61	22.21
Inception-V3	3.97	36.23	23.14

skip-connections such as ResNets, allowing a more optimized utilization of the computational resources. Vgg-16/19 have the largest FLOPs independently of the used configuration. In contrast to the other models, both Vggs required significantly higher computational time. This confimrs that the computational time is strongly affected by the architecture.

H. Overall performance evaluation

The training and validation of a CNN model are often made on randomly selected sets. The adopted splitting scheme may result in biased sets. Limiting these biases helps to improve the reliability of the models as well as the reported performance. With this idea in mind, we first conducted a comparison of three splitting strategies (*see* Sec. III-C). The results demonstrated the existence of content-induced biases, as the performances were different for each splitting strategy. In addition, it showed the difference in terms of content diversity in the available omnidirectional IQA databases. CVIQ appears to contain less diversity compared to OIQA. In our case, in order to provide accurate and reliable results, we adopted for all configurations, the splitting scheme that uses scene complexity and colorfulness as splitting criteria.

Predicting visual quality on projected content (*i.e.* ERP and CMP images) for omnidirectional IQA is straightforward and does not require an additional pre-processing step for extracting viewports or patches. However, the achieved performances are quite poor, except for ResNet-50 and DenseNet-121. This observation is confirmed when conducting cross-database validation, in which the performances decreased substantially. The limitations of using projected content, as well

as the limitations of CVIQ in offering enhanced generalization ability, were demonstrated. When we trained the models on OIQA and tested on CVIQ, the results were better than when the reverse was performed. Because OIQA comprises a variety of distortions, it benefited the models in achieving higher correlation accuracy and monotonicity when compared to CVIQ, which solely incorporates compression artifacts.

The use of radial-content (*i.e.* spherical content) helps to mimic the exploration behavior of users, predicts visual quality on the actual viewed content, and avoids geometric distortions due to projection. This approach results in a set of regions for quality predictions. The challenge is to determine the number of these regions as well as their locations. Regarding the latter, it is preferable to focus on the equator, as demonstrated in [29]. We utilized 8, 16, and 24 extracted regions to investigate the influence of their number on the performance of the selected models. An improvement was observed with the increase of number of inputs, showing the importance of feeding CNN models with more content to learn from. An overall improvement was also observed regarding all models compared to the use of projected content. Except for Vgg-16/19 on OIQA where a low performance is observed. The performance of DenseNet-121 (resp. ResNet-50) stood out from the rest of the models on CVIQ (resp. OIQA).

Training a CNN model on selected regions from an omnidirectional image either implies the use of a multichannel CNN or a patch-wise training. The multichannel CNN learns from multiple inputs that are linked to a single ground truth (*i.e.* MOS), while patch-wise learning involves labeling each extracted patch independently. The multichannel strategy is investigated with the CMP- and Radial content-based configuration. For the patch-wise, two techniques were evaluated, the use of radial content and faces from the cube-map projection as patches. Overall, the superiority of using the radial content was observed. With this configuration, the DenseNet-121 and ResNet-50 still outperform the other models. The cross-database evaluation supports the idea of using radial content as well as generating larger training sets. Good performances were obtained when we trained the models on OIQA and tested on CVIQ, especially for JPEG distortion. Despite the difference in the used levels of JPEG, five on OIQA and eleven on CVIQ, PLCC/SRCC values above 0.90/0.80 were achieved.

Except for training on 2D databases, ResNet-50 and DenseNet-121 performed the best across all tested configurations. This actively demonstrates the effectiveness of these models for IQA tasks when used with the ImageNet weights. When trained on 2D databases, ResNet-18/34 and Inception-V3 achieved competitive performances noticeably better compared to those obtained with their original weights. This shows that deeper models need large databases, while less deep ones may achieve high accuracy with fewer data. In addition, actual 2D IQA databases are limited in comparison to ImageNet. Building IQA databases is time-consuming, which is why transfer learning is usually adopted; tiny databases would not allow CNN models to reach a substantial degree of accuracy. However, when we trained the selected models on LIVE, CSIQ, and TID2013 databases, we could observe an improvement over the pre-trained versions. Overall, the best performances were achieved when trained on LIVE and CSIQ. These databases share four distortion types with OIQA. In terms of loss contrast and training convergence, we discovered that the broader the database (*i.e.* various distortion type), the faster the model trains and less contrast it obtains (*see* Fig. 7). This may be due to fewer examples to learn from when an important number of distortion is used. When we trained on the combined datasets, some improvement were observed, especially with CVIQ, while less generalization is achieved when fine-tuned on diverse database (OIQA).

I. Summary

The main takeaways of this benchmark are:

- When training CNNs models for IQA, a complete separation of the training and testing sets should be performed. Otherwise, the validation would be biased as the model will have already seen the content. In addition, to avoid the representativeness-bias a content-oriented splitting strategy should be considered.
- IQA datasets for training CNN models may suffer from diversity, either in terms of content or distortions. Consequently, the generalization ability and robustness of the model may be highly affected.
- The use of projected contents limits the achievable performances, especially the generalization ability. The fact that this content present geometric distortions and less fidelity with the viewed content resulted in limited performances. In this case, the use of radial content could be more effective.
- Patch-based training is as efficient as multichannel models featuring several CNNs in parallel. With proper patches sampling and training strategy, the patch-based training should be considered since it drastically reduces the complexity. By doing so, the inference time is improved while maintaining promising accuracy.
- The design of multichannel models should properly consider the number of channels. The latter may influence the prediction performances in addition to being highly complex, leading to training difficulties.
- According to the experimental results, there is no linear relationship between the accuracy nor the monotonicity of the model and its complexity.
- Pre-training on 2D-IQA is helpful for increasing backbone performance over ImageNet weights. However, when databases are combined, some models perform poorly in terms of generalization, failing to account for the difference between pre-training and fine-tuning tasks.

V. Conclusion

In this paper, we explored the usage of well-known CNN models for IQA of omnidirectional images. The reason for this choice is that these models were trained on large-scale databases, and transfer learning techniques may benefit IQA. We conducted an empirical and analytical evaluation by covering different CNN architectures, image representations, and training strategies to provide recommendations on the use of CNNs for omnidirectional IQA. Seven pre-trained

CNN models were fine-tuned and compared based on various configurations, including the use of projected and radial content, multichannel paradigm and patch-wise training, and retraining on well-known 2D IQA databases. The obtained results showed the superiority of retraining CNN models on IQA databases over the use of ImageNet pre-trained versions. The use of radial content led to better performance and generalization ability compared to projected content, especially with the patch-wise training. Among the selected models, ResNet-50 and DenseNet-121 performed the best. We believe that this work sheds light on the usage of pre-trained CNN models for IQA and paves the way for further research. One critical factor is the scarcity of large-scale, accurate, and reliable omnidirectional databases. It can be viewed as the foundation of any quality assessment validation scenario, and such databases are in urgent need in order to promote the development of IQA models for such content.

REFERENCES

[1] D. M. Chandler, "Seven challenges in image quality assessment: past, present, and future research," *International Scholarly Research Notices*, 2013.

[2] D. R. Bull, "Chapter 10 - measuring and managing picture quality," in *Communicating Pictures*. Oxford: Academic Press, 2014, pp. 317–360.

[3] Y. Niu, Y. Zhong, W. Guo, Y. Shi, and P. Chen, "2D and 3D image quality assessment: A survey of metrics and challenges," *IEEE Access*, vol. 7, pp. 782–801, 2018.

[4] G. Zhai and X. Min, "Perceptual image quality assessment: a survey," *Science China Information Sciences*, vol. 63, pp. 1–52, 2020.

[5] M. Yu, H. Lakshman, and B. Girod, "A framework to evaluate omnidirectional video coding schemes," in *IEEE ISMAR*, Fukuoka, Japan, 2015, pp. 31–36.

[6] A. Sendjasni, M. Larabi, and F. Cheikh, "On the Improvement of 2D Quality Assessment Metrics for Omnidirectional Images," in *Electronic Imaging, IQSP XVII*, Burlingame, California USA, 2020, pp. 287–1.

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

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

[9] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *IEEE ISCAS*, Paris, France, 2010, pp. 253–256.

[10] C. M. Schneck, "Visual perception," *Occupational Therapy for Children. sixth ed. Mosby Inc*, pp. 373–403, 2010.

[11] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. Bovik, "Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment," *IEEE Signal processing magazine*, vol. 34, no. 6, pp. 130–141, 2017.

[12] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *IEEE CVPR*, Columbus, OH, 2014, pp. 1733–1740.

[13] 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 MMSP*, Vancouver, 2018, pp. 1–6.

[14] 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 Trans. Image Process*, vol. 27, no. 12, pp. 6039–6050, 2018.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, Las Vegas, NV, 2016, pp. 770–778.

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

[17] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE CVPR*, Honolulu, HI, 2017, pp. 4700–4708.

[18] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *IEEE CVPR*, Las Vegas, NV, 2016, pp. 2818–2826.

[19] A. Chetouani and L. Li, "On the use of a scanpath predictor and convolutional neural network for blind image quality assessment," *Signal Processing: Image Communication*, vol. 89, p. 115963, 08 2020.

[20] N. Ahmed and H. Asif, "Perceptual quality assessment of digital images using deep features," *Computing and Informatics*, vol. 39, no. 3, pp. 385–409, 2020.

[21] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[22] B. Zoph, V. Vasudevan, and J. S. Q. and, "Learning transferable architectures for scalable image recognition," *arXiv:1707.07012*, 2017.

[23] F. Gao, J. Yu, S. Zhu, O. Huang, and Q. Tian, "Blind image quality prediction by exploiting multi-level deep representations," *Pattern Recognition*, vol. 81, pp. 432–442, 2018.

[24] S. Bianco, L. Celona, P. Napoletano, and R. Schettini, "On the use of deep learning for blind image quality assessment," *Signal, Image and Video Processing*, vol. 12, no. 2, pp. 355–362, 2018.

[25] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind Image Quality Assessment Using a Deep Bilinear Convolutional Neural Network," *IEEE Trans. Circ. Syst. Video Tech*, vol. 30, no. 1, pp. 36–47, 2020.

[26] C. Huang and J. Wu, "Multi-task deep cnn model for no-reference image quality assessment on smartphone camera photos," *arXiv preprint arXiv:2008.11961*, 2020.

[27] 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. Image Process*, vol. 27, no. 1, pp. 206–219, 2018.

[28] J. Kim and S. Lee, "Deep blind image quality assessment by employing FR-IQA," in *IEEE ICIP*, 2017, pp. 3180–3184.

[29] 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*, vol. 24, no. 4, pp. 1633–1642, 2018.

[30] 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 ISCAS*, Sapporo, Japan, 2019, pp. 1–5.

[31] J. Xu, W. Zhou, and Z. Chen, "Blind omnidirectional image quality assessment with viewport oriented graph convolutional networks," *IEEE Trans. Circ. Syst. Video Tech*, vol. 31, no. 5, pp. 1724–1737, 2021.

[32] H. G. Kim, H. Lim, and Y. M. Ro, "Deep virtual reality image quality assessment with human perception guider for omnidirectional image," *IEEE Trans. Circ. Syst. Video Tech*, vol. 30, no. 4, pp. 917–928, 2020.

[33] A. Sendjasni, M. Larabi, and F. A. Cheikh, "Perceptually-weighted CNN for 360-degree Image quality assessment using visual scan-path and JND," in *IEEE ICIP*, Anchorage, Alaska, 2021, pp. 1439–1443.

[34] T. Truong, H. Tran, and T. Thang, "Non-reference Quality Assessment Model using Deep learning for Omnidirectional Images," in *IEEE ICAST*. Morioka: IEEE, 2019, pp. 1–5.

[35] O. Russakovsky, J. Deng, H. Su, J. Krause, and S. S. et al., "ImageNet Large Scale Visual Recognition Challenge," *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.

[36] V. Nair and G. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010.

[37] J. Brownlee, *Deep learning for computer vision: image classification, object detection, and face recognition in python*. Machine Learning Mastery, 2019.

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

[39] M. Utke, S. Zadtootaghaj, S. Schmidt, S. Bosse, and S. Möller, "Ndnetgaming-development of a no-reference deep CNN for gaming video quality prediction," *Multimedia Tools and App.*, pp. 1–23, 2020.

[40] L. Bowen, Z. Weixia, T. Meng, Z. Guangtao, and W. Xianpei, "Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception," *arXiv preprint arXiv:2108.08505*, 2021.

[41] ITU-R, *Methodology for the subjective assessment of the quality of television pictures BT Series Broadcasting service*, 2012, vol. 13.

[42] X. Zheng, G. Jiang, M. Yu, and H. Jiang, "Segmented spherical projection-based blind omnidirectional image quality assessment," *IEEE Access*, vol. 8, pp. 31 647–31 659, 2020.

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

[44] A. Sendjasni, M. Larabi, and F. Cheikh, "On the Influence of Head-Mounted Displays on Quality Rating of Omnidirectional images," in *Electronic Imaging. (To appear)*, 2021.

[45] T. B. Pierre Lebreton, "Siti," https://vqeg.github.io/software-tools/quality%20analysis/siti/, online; accessed 30 March 2021.

[46] D. Hasler and S. Süsstrunk, "Measuring colorfulness in natural images," in *Human vision and electronic imaging VIII*, vol. 5007. International Society for Optics and Photonics, 2003, pp. 87–95.

[47] L. Li, Z. Li, X. Ma, H. Yang, and H. Li, "Advanced spherical motion model and local padding for 360 video compression," *IEEE Trans. Image Process*, vol. 28, no. 5, pp. 2342–2356, 2018.

[48] C. Li, M. Xu, X. Du, and Z. Wang, "Bridge the gap between vqa and human behavior on omnidirectional video: A large-scale dataset and a deep learning model," in *ACM MM*, 2018.

[49] H. Sheikh, Z. Wang, L. Cormack, and A. Bovik, "Live image quality assessment database release 2," *https://live.ece.utexas.edu/research/Quality/subjective.htm*, 2005.

[50] E. Larson and D. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of electronic imaging*, vol. 19, no. 1, p. 011006, 2010.

[51] N. Ponomarenko, O. Ieremeiev, V. Lukin, and K. E. et al., "Color image database TID2013: Peculiarities and preliminary results," in *IEEE EUVIP*, 2013, pp. 106–111.

[52] Z. Weixia, M. Kede, Z. Guangtao, and Y. Xiaokang, "Uncertainty-aware blind image quality assessment in the laboratory and wild," *IEEE Transactions on Image Processing*, vol. 30, pp. 3474–3486, 2021.

[53] A. Martín, P. Barham, J. Chen, and Z. C. et al., "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} Symp. on Operating Sys. Design and Impl.*, Savannah, GA, 2016, pp. 265–283.

[54] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[55] D. Masters and C. Luschi, "Revisiting small batch training for deep neural networks," *arXiv preprint arXiv:1804.07612*, 04 2018.

[56] L. Hao, C. Pratik, Y. Hao, L. Michael, R. Avinash, B. Rahul, and S. Stefano, "Rethinking the hyperparameters for fine-tuning," *arXiv preprint arXiv:2002.11770*, 2020.

[57] C. Zhang, P. Benz, D. Argaw, S. Lee, J. Kim, F. Rameau, J. Bazin, and I. Kweon, "Resnet or densenet? introducing dense shortcuts to resnet," in *IEEE/CVF Winter Conference on Applications of Computer Vision*, Waikoloa, Hawaii, 2021, pp. 3550–3559.

[58] J. Yamanaka, S. Kuwashima, and T. Kurita, "Fast and accurate image super resolution by deep cnn with skip connection and network in network," in *ICNIP*. Springer, 2017, pp. 217–225.

Abderrezzaq Sendjasni is currently a PhD candidate in signal and image processing at the University of Poitiers, France and NTNU, Norway. He received his BSc (computer science) from the university of Chlef, Algeria in 2015, his MSc (embedded system engineering) from the computer science department, university of Oran 1, Algeria in 2017. His research interests include image processing, human visual perception, image quality assessment, 360-degree images, immersive application, and deep learning.

Mohamed-Chaker Larabi (M'05-SM'07) received his Ph.D. degree from Université de Poitiers in 2002. He is currently Associate Professor with the same university. He is also Deputy Scientific Director of the GdR-ISIS (French research group on signal and image processing). He participated in several national and international projects. He supervised more than 20 Ph.D. students and published more than 200 papers. His actual scientific interests deal with quality of experience and bio-inspired processing/coding/optimization of images and videos, such as 2D, 3D, HDR, and 360/VR/AR/MR. He has been elected to serve as a member of the IEEE SPS IVMSP, MMSP, and EURASIP TAC-VIP Technical Committees. He is a member of the CIE, IS&T, and the MPEG and JPEG committees. He served as the Chair for the JPEG Advanced Image Coding and the Test & Quality Subgroup, and he acted as the French Head of Delegation for several years. He was the Program Chair of the EUVIP 2011 and 2018, Plenary Chair of the EUVIP 2013, Chair of the EI Image Quality and System Performance (2014–2016, 2021-2023), short courses Co-Chair of the Electronic imaging symposium (2016–2018), Technical Co-Chair of the EUVIP 2018, Special Sessions Co-Chair of ICIP 2016, Publicity Chair of ICIP 2017 and 2021, Workshop Co-Chair of ICME 2022, and Exhibits & Demo Show Chairs of ICIP 2022. He is part of the steering committees of ICME, AVSS and EUVIP. He played several roles in different conferences. He serves as Associate Editor for the IEEE Transactions on Image Processing, the IEEE Transactions on Circuits and Systems for Video Technology, the Springer Signal, Image and Video Processing, the SPIE/IS&T Journal of Electronic Imaging, the IEEE ACCESS, and Elsevier journal of Visual Communication and Image Representation and Signal Processing: Image Communication.

Faouzi Alaya Cheikh received the Ph.D. degree in information technology from Tampere University of Technology, Tampere, Finland, in April 2004. He had worked as a Researcher with the Signal Processing Algorithm Group, Tampere University of Technology, in 1994. Since 2006, he has been affiliated with the Department of Computer Science and Media Technology, Gjøvik University College, Norway, as an Associate Professor. Since January 2016, he has also been with the Norwegian University of Science and Technology (NTNU). He teaches courses on image and video processing and analysis and media security. He is currently the co-supervisor of five Ph.D. students. He has been involved in several European and national projects, such as ESPRIT, NOBLESS, COST 211Quat, HyPerCept, IQ-Med, and H2020 ITN HiPerNav. His research interests include e-Learning, 3-D imaging, image and video processing and analysis, video-based navigation, biometrics, pattern recognition, embedded systems, and content-based image retrieval. In these areas, he has published over 100 peer-reviewed journal articles and conference papers, and supervised four postdoctoral researchers, five Ph.D., and a number of M.Sc. thesis projects. He is a member of NOBIM and Forskerforbundet [The Norwegian Association of Researchers (NAR)]. He is on the Editorial Board of the IET Image Processing Journal and the Journal of Advanced Robotics & Automation, and the technical committees of several international conferences. He is an expert reviewer to a number of scientific journals and conferences related to the field of his research.