A BENCHMARK OF VARIANCE OF OPINION SCORES IN IMAGE QUALITY ASSESSMENT

Jianxun Lou[1,2], Xinbo Wu[2], Yingying Wu[2], Padraig Corcoran[2], Gualtiero Colombo[2],
Roger Whitaker[2], and Hantao Liu[2]

[1]School of Computer Science, Northeast Electric Power University, Jilin, China
[2]School of Computer Science and Informatics, Cardiff University, Cardiff, United Kingdom

ABSTRACT

Mean opinion score (MOS) has been used as the benchmark to measure the perceived quality of digital images. However, the usefulness of MOS diminishes when a substantial variation between individual opinions occurs. It is critical to measure the stimulus-driven variance of opinion scores (VOS) and scrutinise images that evoke a large VOS, and consequently, use VOS to inform our interpretation of MOS. In this paper, we create a VOS benchmark for individual differences in image quality assessment and analyse the importance of VOS classification as a function of distortion intensity, distortion type and scene content. In addition, a simple yet effective deep learning-based model is built, aiming to identify images with a large variation in viewers' quality judgements.

Index Terms— Image quality assessment, individual differences, variance, distortion, deep learning

1. INTRODUCTION

Image quality assessment (IQA) has been extensively studied, encompassing both subjective and objective assessments [1–6]. In subjective IQA, a number of human subjects are asked to participate in a psychovisual experiment and express their opinions on perceived image quality [1–4]. The subjective data provide the ground truth for the development of objective algorithms that can automatically predict image quality as perceived by humans. The research in IQA is of significant benefit to subject areas such as telecommunications [7], medical imaging [8], and computer vision [9].

In the literature, the mean opinion score (MOS) - the average of individual opinions of human subjects - is customarily used as the benchmark to measure the perceived quality of images [5, 6]. Until now, there has been limited focus on the diversity in subjects' opinions in assessing the quality of an image. Generally, MOS is consider most useful when the degree of variability amongst individuals is within an acceptable range. However, the significance of MOS depreciates when a substantial variation between individual opinions frequently occurs. As illustrated in Fig. 1, two images exhibit similar MOS values (i.e., 67.7 and 66.9 in the range of [0, 100]), yet the *variance of opinion scores* (VOS) reported by individuals is notably distinct (i.e., 0.960 and 0.385 in the range of [0, 1]). This observation presents an often overlooked challenge in image quality assessment, i.e., while MOS can provide a general assessment of image quality, its validity could be better interpreted by the associated VOS that reflects the degree of disparity in evaluations amongst subjects. Hence, it is critical to understand individual differences in IQA and use the variance of opinion scores (VOS) to enhance the interpretation of MOS in specific IQA-related applications.

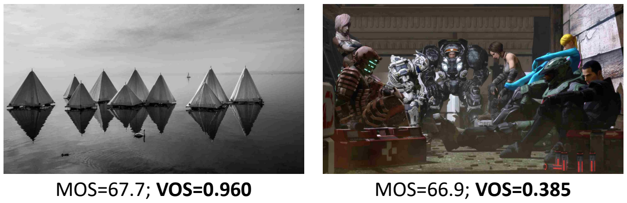

Fig. 1. Illustration of two images exhibiting similar mean opinion score (MOS) values (i.e., 67.7 and 66.9 in the range of [0, 100]). The variance of opinion scores (VOS) reported by individuals is notably distinct (i.e., 0.960 and 0.385 in the range of [0, 1]).

Several studies have been conducted to investigate the variability in subjects' opinions in image quality assessment. For example, the research [10] hypothesised a relationship between the mean opinion score and the standard deviation of individual scores, highlighting the significance of rating variability in the evaluation of the quality of experience. Some methods have been proposed to predict the distribution of subjective opinion scores within a group of participants [11, 12]. These attempts primarily concentrate on reproducing individual ratings provided by viewers in image quality assessment. In some application scenarios, the usefulness of MOS can be optimised by having an additional measure of variance of opinion scores (VOS), which provides a direct and quantifiable indicator of the consistency of subjective ratings for each

Corresponding author: Xinbo Wu (wux37@cardiff.ac.uk)

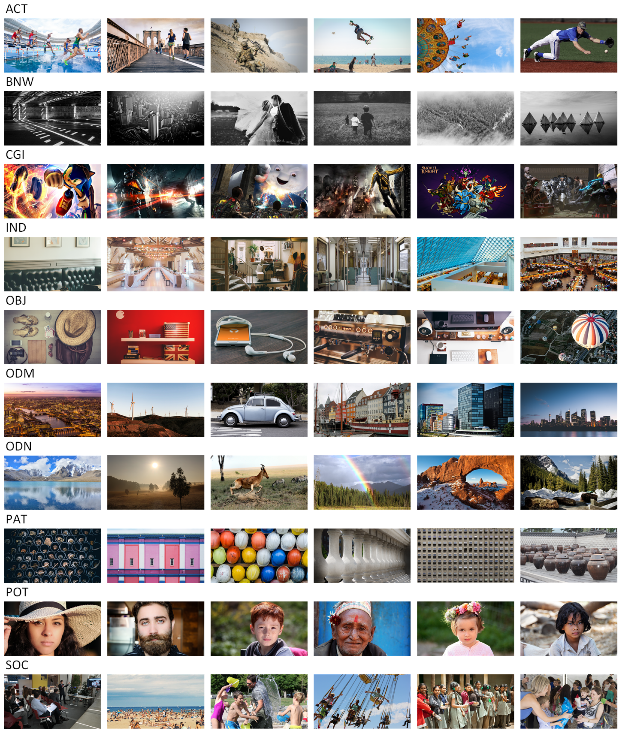

Fig. 2. Reference visual scenes in the CUID dataset [4].

stimulus. To achieve this, research is needed to measure VOS in image quality assessment, and consequently, to develop a computational solution to predict VOS for images of varying perceived quality.

In this paper, we create a benchmark of VOS in IQA. It includes 540 images of diverse content with varying degrees of perceived quality and each containing 19 scores from individual subjects engaged in a fully controlled psychovisual experiment. We performed a thorough analysis to reveal plausible attributes of VOS including distortion intensity, distortion type, and scene content, providing insights into the stimulus-driven influencing factors for the diversity of subjective opinions in IQA. Furthermore, we develop a simple yet effective deep learning model to identify images exhibiting low consistency in subjective quality ratings. This model can be used to provide supplementary measures to existing objective IQA metrics, bridging the gap between the measurement of perceived quality and the awareness of the diversity in viewers' opinions for specific visual stimuli.

2. BENCHMARK OF VARIANCE OF OPINION SCORES

2.1. CUID dataset

The CUID dataset [4] was purposely built to study image quality perception in a fully controlled laboratory environment using a rigorous psychovisual experiment design. In the CUID dataset, reliable image quality ratings were collected using a within-subjects method and a large degree of stimulus variability. It contains 60 diverse reference scenes from 10 distinct visual content categories, including Action (ACT): images depicting high activity; Black and White (BNW): grayscale images; Computer-Generated Imagery (CGI): computer-generated graphic images; Indoor (IND): images captured from indoor scenarios; Object (OBJ): images featuring various objects; Outdoor Man-made (ODM): images from outdoor scenarios with man-made objects; Outdoor Natural (ODN): images from outdoor scenarios with natural scenes; Pattern (PAT): images with repeating objects; Portrait (POT): close-up shots of human faces; and Social (SOC): images depicting interactions between people. Fig. 2 illustrates the reference visual scenes. A total of 540 distorted images were generated from the 60 references, simulating three different distortion types including contrast change (CC), JPEG compression (JPEG), and motion blur (MB); and three distinct distortion intensity levels including low-level (Q1: perceptible but not annoying distortions), medium-level (Q2: noticeable and annoying distortions), and high-level (Q3: very annoying distortions). Details about the CUID dataset can be found in [4].

2.2. Variance of opinion scores (VOS)

To study how the combination of natural content and unnatural distortions can impact the variability of subjective opinions in image quality assessment, we analyse the individual ratings provided by viewers for each of the 540 distorted images in the CUID dataset. First, to account for the differences between subjects in the use of the scoring scale, the raw subjective scores were calibrated towards the same mean and standard deviation, by converting them into z-scores as detailed in [4]. Then the variance of opinion scores (VOS) of each image is calculated:

$$\mathrm{VOS}_j^* = \frac{1}{N}\sum_{i=1}^{N}(s_i - \bar{s})^2, \tag{1}$$

where j denotes the j-th distorted image contained in the CUID dataset; s_i, $\bar{s}$, and N denote the i-th subjective score, the mean of all subjective scores, and the total number of scores provided for the j-th image, respectively. To make the VOS values easy to interpret, we linearly map the VOS values to the range between 0 and 1 as follows:

$$\mathrm{VOS}_j = \frac{\mathrm{VOS}_j^* - \min(\mathrm{VOS}^*)}{\max(\mathrm{VOS}^*) - \min(\mathrm{VOS}^*)}. \tag{2}$$

Fig. 3 shows examples of images of high and low VOS values in the CUID dataset. The histogram of these normalised VOS values is illustrated in Fig. 4, where a higher VOS value indicates significant variability in subjective opinions for an image. As a result, a new benchmark of variance of opinion scores (VOS) is created, namely VOSIQ dataset.

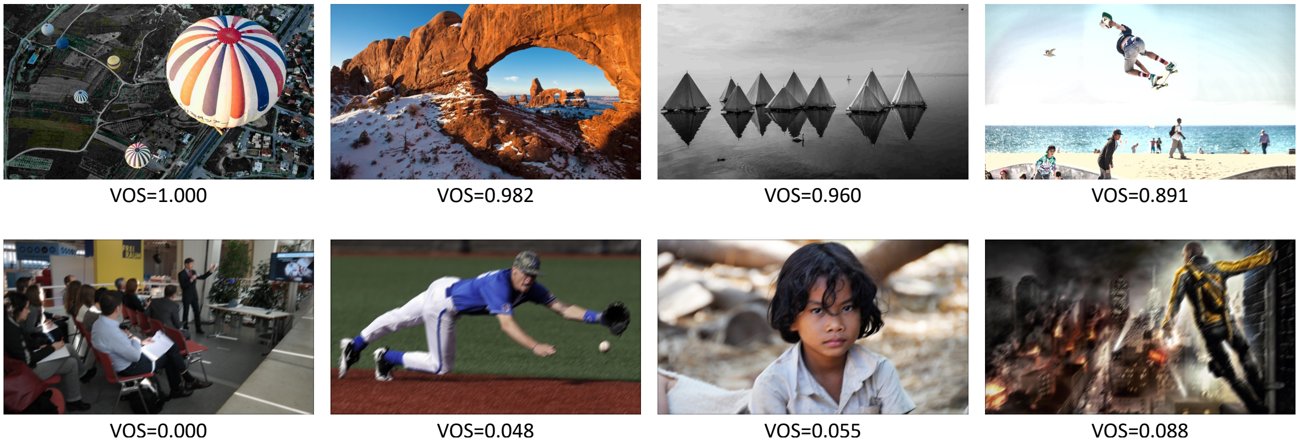

Fig. 3. Examples of images from the VOSIQ dataset. The first row features images with a high VOS value, and the second row displays images with a low VOS value.

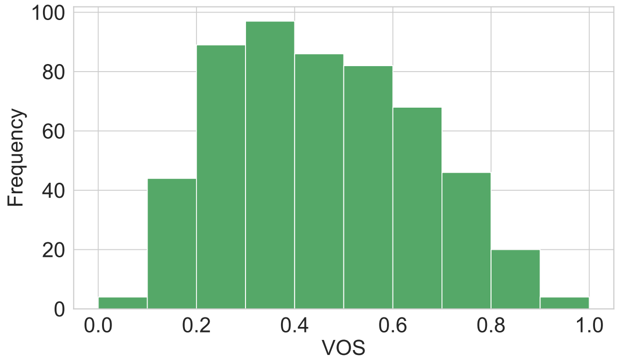

Fig. 4. Histogram of VOS values in the VOSIQ dataset.

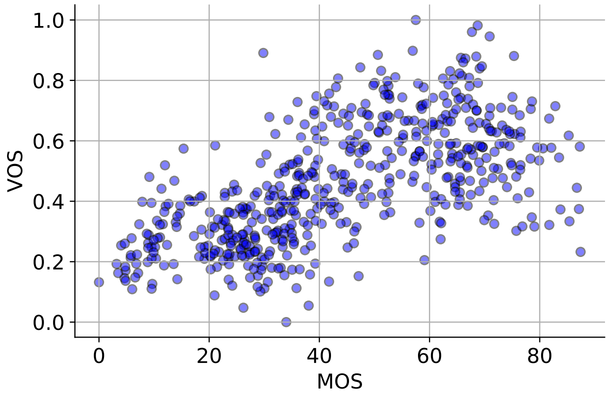

Fig. 5. Scatter plot showing the relationship between VOS and MOS for the VOSIQ dataset.

To investigate the relationship between VOS and MOS, we plot VOS against MOS for the VOSIQ dataset, as shown in Fig. 5. It can be seen from the visualisation that changes in VOS cannot be fully explained by the variations in image quality (i.e., MOS). This suggests a need for further analysis of attributes of VOS to gain a comprehensive understanding of stimulus-driven individual differences in image quality assessment.

3. ANALYSIS OF VARIANCE OF OPINION SCORES

When judging the quality of an image, a high VOS value indicates a substantial divergence in individual opinions in quality perception, whereas a relatively low VOS value suggests a tendency towards consensus in quality assessment amongst viewers. Now, we analyse how this disparity in image quality perception (i.e., the classification of VOS) is affected by plausible stimulus-driven influencing factors including distortion intensity, distortion type, and scene content.

Let μ and σ represent the mean and standard deviation of the VOS values calculated for all images contained in the VOSIQ dataset, respectively. We propose the following VOS classification method: VOS values exceeding $\mu + \sigma$ are categorised as a high VOS (HVOS); and those falling below this threshold are defined as a low VOS (LVOS). After applying this method, the images of the VOSIQ dataset are divided into two groups with one HVOS group having the mean VOS value of 0.7497 and one LVOS group having the mean VOS value of 0.3878. The statistical significant difference between these two groups is verified by a Mann-Whitney U test (p-value$<$0.01), validating the rationale of the proposed classification method. Now, we analyse the attributes of VOS and their impact of the VOS classification.

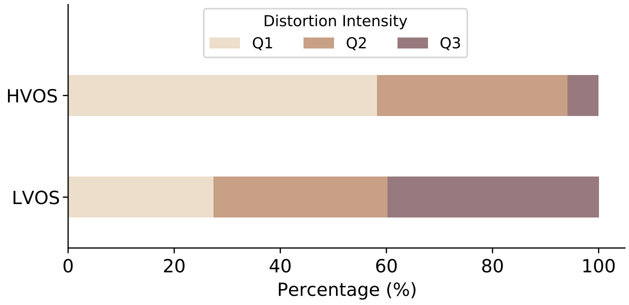

Fig. 6. Distribution of HVOS and LVOS images over different levels of distortion intensity: Q1 (low-level), Q2 (medium-level), and Q3 (high-level).

3.1. Impact of distortion intensity

We investigate the impact of image distortion intensity on the VOS classification. For each VOS class (i.e., HVOS or LVOS), we calculate the distribution in percentage (image count) over the pre-defined three distortion intensity levels i.e., Q1 (low-level), Q2 (medium-level), and Q3 (high-level). The results are presented in Fig. 6, showing some interesting patterns. First, for images classified as HVOS, the majority falls into the category of low-level distortion (Q1), representing 58% of all HVOS images. This tends to suggest a considerable level of subjectivity and perceptual variation amongst viewers when discerning subtle distortions. The images of HVOS rarely (i.e., 6%) present in the category of high-level distortion (Q3), indicating a consensus between viewers for assessing pronounced distortions. Second, for images classified as LVOS, they are evenly spread across all distortion levels, which implies the impact of distortion intensity is not significant. It should be noted that the images of LVOS pose a negligible concern in image quality assessment, as the use of MOS can reliability interpret the subjects' opinions. Overall, these findings suggest that the level of distortion intensity plays a crucial role in determining the distribution of images of HVOS. The combination of specific scene content and subtle distortions is likely to cause a high variation in viewers' opinions on perceived image quality.

3.2. Impact of distortion type

This analysis entails calculating the distribution in percentage (image count) over the three distortion types contained in the VOSIQ dataset, including contrast change (CC), JPEG compression (JPEG), and motion blur (MB), once for the HVOS class and once for the LVOS class. The results are presented in Fig. 7. First, the figure shows that the images of HVOS are predominantly present in the CC and JPEG distortion types, comprising 44% and 41% of all HVOS images, respectively. The MB distortion type exhibits less cases of HVOS images (i.e., 15%). This indicates that CC and JPEG distortions, compared to MB distortions are more likely to elicit a large variation in subjective opinions amongst viewers. This might be attributed to the fact that CC and JPEG often cause some localised distortions and when combining with specific scene content they can induce different levels of impact on viewers' assessments of overall image quality. MB distortions are uniformly distributed in the spatial domain, and the effect on individual reviewers tends to be more consistent. Second, similar to the above findings for distortion intensity, images of LVOS are evenly spread across all distortion types, which means the impact of distortion type is not significant. Again, it is worth noting that the images of LVOS pose a negligible concern, and that MOS can be a reliable measure of image quality in this case.

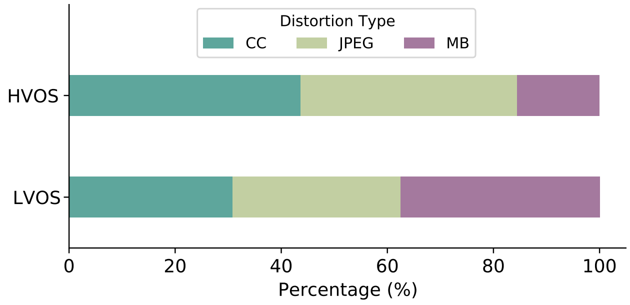

Fig. 7. Distribution of HVOS and LVOS images over different distortion types: CC (contrast change), JPEG (JPEG compression), and MB (motion blur).

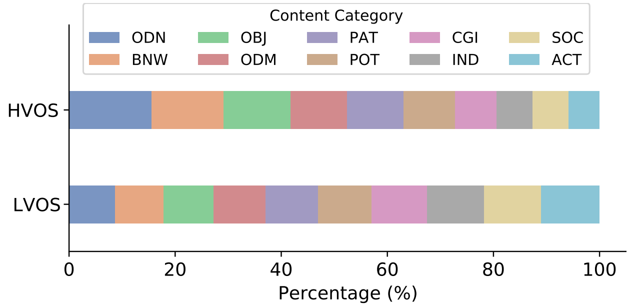

Fig. 8. Distribution of HVOS and LVOS images over 10 different scene categories.

3.3. Impact of scene content

One of the unique features of the VOSIQ dataset is that the stimuli represent 10 distinct natural scene categories. We hereby investigate the effect of scene content on the VOS classification. We calculate the distribution in percentage (image count) over 10 scene categories for HVOS and LVOS, respectively. The results are illustrated in Fig. 8. With regard to the HVOS, certain scene categories are more prone to invoking a large variation of subjective opinions than other categories. It is evident that three scene categories including 'ODN', 'BNW', and 'OBJ' exhibit a notably higher propor-

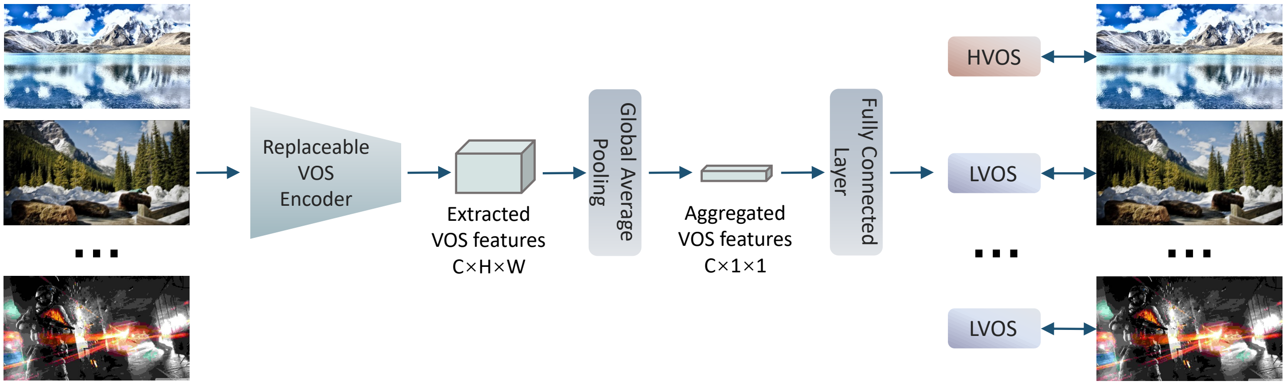

Fig. 9. The schematic overview of the proposed VOSNet for VOS classification.

tion of HVOS images; and a low proportion is found in the 'CGI', 'IND', 'SOC', and 'ACT' categories. A plausible reason could be that visually complex and abstract scenes, such as 'ODN', 'BNW', and 'OBJ' may elicit diverse opinions from viewers, leading to a higher VOS value. Again, similar to above findings, the distribution of LVOS over all scene categories is rather flat (note that in the case of LVOS, MOS alone serves as a reliable measure of image quality). Overall, the analysis suggests that scene category is a significant factor that contributes to the diversity in viewers' image quality ratings.

4. COMPUTATIONAL MODEL OF VARIANCE OF OPINION SCORES

4.1. Proposed VOSNet model

Variance of opinion scores (VOS) can support the validity of the mean opinion score (MOS) that is conventionally used to measure the perceived quality of an image. The VOSIQ dataset provides the ground truth of VOS for image quality assessment, which can facilitate the development of computational models capable of automatically producing the VOS information. It should, however, be noted that depending on the specific application, the use of VOS and its computational model could vary. In this paper, we aim to provide a support tool to better interpret the significance of the customarily used MOS. More specifically, the goal is to identify the images that evoke a large VOS, suggesting the need of further scrutiny of their actual quality in the subsequent stages of processing. To this end, instead of predicting the VOS values with a regression model, a more practical solution is to develop a classifier that can label images as HVOS or LVOS. We propose a simple yet effective model named VOSNet as detailed below.

To construct the VOSNet model, we employ a widely validated deep learning architecture for classifying images into HVOS class or LVOS class as illustrated in Fig. 9. Firstly, input images are processed by a deep VOS encoder, which extracts VOS related features. The VOS encoder is readily replaceable, allowing for the selection of a specific encoder based on practical requirements. In this study, we exploit five commonly used backbone networks, originally designed for ImageNet [13] classification, each in turn serving as the deep VOS encoder. These backbone encoders include MobileNet-V3 [14], EfficientNet-B4 [15], ResNet-50 [16], VGG-16 [17], and ConvNeXt-T [18]. To apply each of these backbones in our VOSNet model, its classification head is removed to generate VOS features. After extracting deep features, they are progressed through the Global Average Pooling (GAP) to produce aggregated VOS features; and subsequently passed to a fully connected layer to obtain the predicted labels.

4.2. Results

We conducted a k-fold cross-validation (k=10) for comprehensive evaluation of the proposed VOSNet model on the VOSIQ dataset. In our implementation, the VOSIQ dataset was partitioned into ten equal, non-overlapping subsets. In each subset, we ensured a consistent number of HVOS images. The cross-validation process was structured in a way that in each run, distinct sets were allocated to serve different purposes: one set for testing, one set for validation, and the remaining eight sets were dedicated to training. This approach rigorously prohibited any overlap or sharing of parameters between runs, ensuring that the models were evaluated on entirely unseen samples. As a result, the average results from all ten runs provided a comprehensive depiction of the model's performance. Also, optimal models were achieved by implementing early stopping, activated after five epochs if no improvement occurs, and utilising the cross entropy loss function with the AdamW optimiser [19]. During training, parameters pre-trained on ImageNet were loaded for the initialisation of the model.

The experimental result are listed in Table 1. It can be seen that different backbone encoders give slightly different results, but all models generally can achieve promising performance. The models based on ResNet-50, VGG-16, and ConvNeXt-T outperform models based on MobileNet-V3 and

Table 1. Performance of proposed VOSNet based on different backbone encoders on the benchmark VOSIQ dataset.

VOSNet-Backbone	Precision	Recall	F1 Score	AUC
MobileNet-V3 [14]	0.7300	0.6278	0.6635	0.5697
EfficientNet-B4 [15]	0.7410	0.5870	0.6311	0.5853
ResNet-50 [16]	0.7908	0.6833	0.7150	0.6782
VGG-16 [17]	0.8003	0.7074	0.7354	0.6968
ConvNeXt-T [18]	0.8069	0.7074	0.7360	0.7079

EfficientNet-B4 in identifying HVOS images. This might be due to certain architectures being potentially better suited for the task of classifying VOS in image quality assessment.

5. CONCLUSION

In this paper, we have built a first-of-its-kind benchmark of variance of opinion scores (VOS) in image quality assessment. VOS measures the degree of variability in image quality ratings given by individual viewers. We have analysed plausible attributes of VOS including distortion intensity levels, distortion types and scene content. The findings signify the importance of identifying images that evoke a high VOS value for practical applications. To provide a computational solution, we have developed a deep learning model, namely VOSNet which can achieve good performance in classifying images of high or low VOS. Future work will focus on the improvement of the VOSNet model.

6. REFERENCES

[1] H.R. Sheikh, M.F. Sabir, and A.C. Bovik, “A statistical evaluation of recent full reference image quality assessment algorithms,” *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440–3451, 2006.

[2] Eric C Larson and Damon M Chandler, “Most apparent distortion: full-reference image quality assessment and the role of strategy,” *Journal of electronic imaging*, vol. 19, no. 1, pp. 011006–011006, 2010.

[3] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe, “Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment,” *IEEE Transactions on Image Processing*, vol. 29, pp. 4041–4056, 2020.

[4] Lucie Lévêque, Ji Yang, Xiaohan Yang, Pengfei Guo, Kenneth Dasalla, Leida Li, Yingying Wu, and Hantao Liu, “Cuid: A new study of perceived image quality and its subjective assessment,” in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 116–120.

[5] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[6] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang, “Maniqa: Multi-dimension attention network for no-reference image quality assessment,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022, pp. 1190–1199.

[7] Anush Krishna Moorthy, Lark Kwon Choi, Alan Conrad Bovik, and Gustavo de Veciana, “Video quality assessment on mobile devices: Subjective, behavioral and objective studies,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 652–671, 2012.

[8] Yueran Ma, Jean-Yves Tanguy, Richard White, Padraig Corcoran, and Hantao Liu, “Impact of radiologist experience on medical image quality perception,” in *2023 15th International Conference on Quality of Multimedia Experience (QoMEX)*, 2023, pp. 177–182.

[9] Huasheng Wang, Jianxun Lou, Xiaochang Liu, Hongchen Tan, Roger Whitaker, and Hantao Liu, “Sspnet: Predicting visual saliency shifts,” *IEEE Transactions on Multimedia*, pp. 1–12, 2023.

[10] Tobias Hoßfeld, Raimund Schatz, and Sebastian Egger, “SOS: The mos is not enough!,” in *2011 Third International Workshop on Quality of Multimedia Experience*, 2011, pp. 131–136.

[11] Yixuan Gao, Xiongkuo Min, Wenhan Zhu, Xiao-Ping Zhang, and Guangtao Zhai, “Image quality score distribution prediction via alpha stable model,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 6, pp. 2656–2671, 2023.

[12] Yixuan Gao, Xiongkuo Min, Yucheng Zhu, Xiao-Ping Zhang, and Guangtao Zhai, “Blind image quality assessment: A fuzzy neural network for opinion score distribution prediction,” *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2023.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[14] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al., “Searching for mobilenetv3,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1314–1324.

[15] Mingxing Tan and Quoc Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

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

[18] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11976–11986.

[19] Ilya Loshchilov, et al., "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.