Subjective And Objective Quality Assessment Of Mobile Gaming Video

Shaoguo Wen, Suiyi Ling, Junle Wang, Ximing Chen, Yanqing Jing, Patrick Patrick Le Callet

SUBJECTIVE AND OBJECTIVE QUALITY ASSESSMENT OF MOBILE GAMING VIDEO

Shaoguo Wen[2], Suiyi Ling[1] , Junle Wang[2], Ximing Chen[2], Lizhi Fang[2], Yanqing Jing[2], Patrick Le Callet[1]

[1] LS2N, University of Nantes [2]Turing Lab, Tencent

ABSTRACT

Nowadays, with the vigorous expansion and development of gaming video streaming techniques and services, the expectation of users, especially the mobile phone users, for higher quality of experience is also growing swiftly. As most of the existing research focuses on traditional video streaming, there is a clear lack of both subjective study and objective quality models that are tailored for quality assessment of mobile gaming content. To this end, in this study, we first present a brand new Tencent Gaming Video dataset containing 1293 mobile gaming sequences encoded with three different codecs. Second, we propose an objective quality framework, namely Efficient hard-RAnk Quality Estimator (ERAQUE), that is equipped with (1) a novel hard pairwise ranking loss, which forces the model to put more emphasis on differentiating similar pairs; (2) an adapted model distillation strategy, which could be utilized to compress the proposed model efficiently without causing significant performance drop. Extensive experiments demonstrate the efficiency and robustness of our model.

Index Terms— Subjective quality assessment, objective quality metric, gaming video, model distillation

1. INTRODUCTION

Gaming video streaming is composed of a wide-range of online services including gaming video streaming, esports broadcasting and cloud gaming services. In the past decade, the booming international popular gaming video streaming platforms including Twitch, YouTube, Smashcast, Afreeca TV, Gosu Gamers, *etc.*, and cloud gaming platforms like Google Stadia and Nvidia Geforce Now, are gradually taking a large share of video streaming [1]. Most of these services have also mobile app to meet the growing prevalence of mobile usage in modern life. According to [2], gaming content also makes up a significant proportion of of User Generated Contents (UGC) on social media platforms, where a certain amount of users browse/view the contents using their mobile phone by a daily base. As reported in [3], mobile gaming accounted for 58.8% of the digital games market in 2019. This percentage is growing even more dramatically due to the global COVID-19 pandemic in 2020. Users' higher requirements for better Quality of Experience (QoE) necessitate robust quality control of these gaming contents. Different from the contents on common video streaming platforms [2], quality assessment of gaming videos raise new challenges to the community as: their temporal complexity could be significantly higher; unlike natural videos, they are graphical rendered/generated contents; most of the gaming streaming platform requires real-time quality evaluation; and cloud gaming is sensitive to delay. Notwithstanding the fact that subjective study is time-consuming and expensive to conduct, it is still the foundation for the development of rigorous objective quality metrics. As most of the existing subjective studies focus mainly on video on traditional Video on Demand streaming services [1], studies conducted for gaming contents, especially for mobile applications, are still scarce.

With the burgeon of video encoding techniques and hardware based acceleration frameworks, dedicated fast codecs have been developed to alleviate relevant stresses. The development of the video objective quality metrics should also keep pace with the one of the codecs. Among No-Reference (NR), Reduced-Reference (RR), and Full-Reference (FR) metrics, NR metrics are of greater value, and significantly more piratical for real-time gaming content quality evaluation [4] since pristine reference is not always existing or accessible. Advanced deep learning based NR quality metric enjoyed a big leap in performance in recent years. The development of Deep Learning (DL) has brought along a new wave of NR quality assessment models that perform markedly better than traditional metrics. However, very few of them target the gaming application, typically for mobile users. Recall that most of those DL based models are complex, cumbersome and thus barely could be deployed on mobile directly for real-time prediction.

Based on the discussions above, in this study, we present a novel large-scale mobile gaming video data set, and a brand new Efficient hard-RAnk Quality Estimator (ERAQUE) to remedy the lack of existing subjective and objective studies.

2. RELATED WORK

Subjective study: Recently, an image dataset for gaming content was released [5], where each individual image was labeled with four dimensions including the 'overall quality' dimension. However, in this study the 'overall quality' was subjectively assessed from an aesthetic perspective of view.

Suiyi Ling and Shaoguan Wen make equal contribution, Junle Wang is the corresponding author.

The 'GamingVideoSET' [1], consists of 576 sequences encoded using H.264 from 24 gaming content. KUGVD [6] is another public available gaming video dataset, but only 6 source contents were collected. Cloud Gaming Video Dataset (CGVDS) [7] is one of the largest existing public gaming video dataset in the domain. Although the number of considered contents in this subjective study is slightly larger than the one in GamingVideoSET, more training data is still required for mainstream deep learning models. Most importantly, none of the above subjective studies were conducted for mobile games, with mobile phones.

Objective models: The demand of No-Reference quality metrics is growing significantly, especially for real-time quality control. Several dedicated blind quality metrics were developed to meet the rapidly growing need. In [8], Utke *et al.* explored different Convolutional Neural Network (CNN) architectures, *e.g.,* DenseNet$_{121}$, without the help of reference video, to predict VMAF score for gaming videos. A blind quality metric was presented in [9], where Support Vector Regression (SVR) was applied to regress quality scores using frame-level index. Recently, the G.1072 planning model [10], which evalutes the quality of gaming videos based on the videos' information, including metadata, codec information, etc., was summarized in the ITU-T recommendation standard (ITU-T.G1072). Similarly, a parametric video quality metric was introudced in [11] by Zadtootaghaj *et al.* Among existing NR video metrics, DEMI [4] is one of the most recent state-of-the-art video quality gaming metric. It was developed by training a CNN to capture videos distortions guided by full reference quality metric, finetuning the obtained network on a smaller quality datasets, and finally predicting the quality scores utilizing Random Forest. Following a similar recipe, a three-step framework NDNetGaming [12] was designed using also CNN, with a novel temporal pooling methodology that considers the temporal masking effect. However, none of them were designed and tested on mobile gaming contents.

3. THE SUBJECTIVE STUDY

In this section, details of our novel Tencent Gaming Video (TGV) dataset and the subjective experiment are provided. Overall information of the dataset is summarized in Table 1.

Table 1. Summary of out TGV dataset.

Duration	5 sec
Frame rate	30 f/s
Resolution	480P, 720P, 1080P
Bitrate	$100, 200, ..., 10^4$

Content: Our gaming video dataset contains 1293 sequences. In total 150 source gaming videos were collected from 17 mobile games. These games are available from the Tencent internal gaming center. Thumbnails of selected mobile gaming contents are shown in Fig. 1. There different codecs, including the H264, H265, and one internal Tencent codec, were employed to encode each of the content with a variety of different setting/configurations, *e.g.,* CRF values were selected randomly from $[1, ..., 50]$.

Fig. 1. Thumbnails of selected mobile gaming contents.

Environment & Observers: The Absolute Category Rating (ACR) subjective protocol was used for the test. During the experiment, 19 participants with normal or corrected-to-normal acuity were asked to score the gaming videos. Among them, there were 13 males, and 6 females, most of them are between 23-26 years old. The subjective experiment was conducted in an regularized office room with, where illuminance, *etc.* was controlled according to [13]. During the subjective experiment, the observers was asked to launch a internally developed platform to start the test, using their own mobile phone. In another word, different models of mobile phone were utilized by different participants to conduct the subjective test, which is consistent with real application scenario. After collecting the subjective data, two outliers were removed using the screening tools recommended by [13].

4. THE PROPOSED OBJECTIVE MODEL

4.1. Network Architecture

The architecture of the proposed ERAQUE for gaming video quality assessment is depicted in Fig. 2. In quality domain, due to limited training data, deep complex models are prone to overfitting [14]. Therefore, the first part of the network is a light-weight backbone network. Subsequently, the Global Averaged Pooling [15] is plugged in to obtain more streamlined latent representation following with two Fully Connected (FC) layers to output the final quality score. When predicting quality score for gaming videos, frames are first extracted and fed into the model, and then the frame-wise scores are averaged to obtain the final quality score of the video. Based on different use case under different scenarios, network like Mobilenet [16], Shufflenet [17], ResNet-18 *etc.*, could be employed accordingly. In this study, we start with the classical ResNet 18 network that was pre-trained on Imagenet as backbone, and finetune it using quality data. This model is further compressed our re-adapted model distillation strategy. Details are shown in Section 4.3.

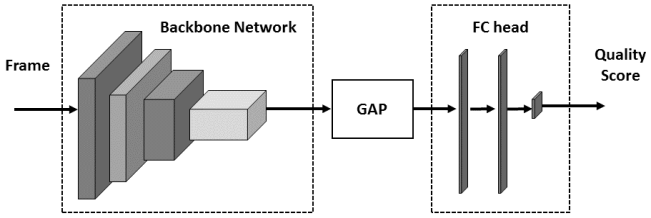

Fig. 2. Architecture of the proposed quality metric.

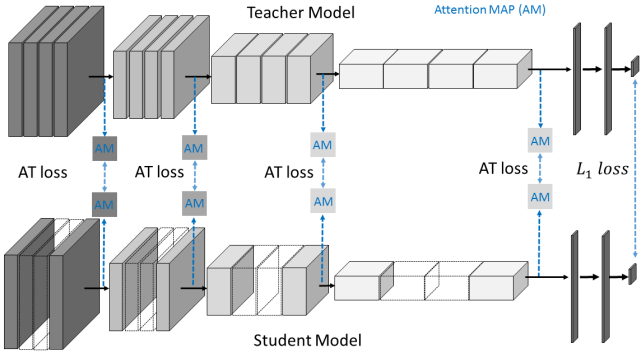

Fig. 3. Teacher-student network for model distillation.

4.2. Loss function

Let $\hat{y}$ and y be the quality score predicted by the objective model, and the ground truth quality score collected from the subjective experiment respectively. n is the total number of considered videos. The loss function of our proposed model is composed of two parts as defined below, where λ is a parameter that balances the two losses:

$$L = L_{mae} + \lambda \cdot L_{hard-rank}. \tag{1}$$

The first part is the Mean Absolute Error (MAE) loss L_{mae} between the ground truth and predicted scores:

$$L_{mae} = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{n}. \tag{2}$$

The second part is a hard pair-wise ranking loss $L^{ij}_{hard-rank}$ that is inspired by the metric learning based framework proposed in [18]. Given a pair of stimuli (y_i, y_j), the proposed $L^{ij}_{hard-rank}$ is designed as:

$$L^{ij}_{hard-rank} = l_2(y_i, y_j) \cdot max(0, |y_i - y_j| - l_1(y_i, y_j) \cdot (\hat{y_i} - \hat{y_j})), \tag{3}$$

where l_1 and l_2 are further defined in equation (4) and (5) correspondingly, and τ is a margin (similar to the ones defined in [18]):

$$l_1 = \begin{cases} 1, & y_i \geq y_j \\ -1, & otherwise \end{cases} \tag{4}$$

$$l_2 = \begin{cases} 1, & 0 < |y_i - y_j| \leq \tau \\ 0, & otherwise \end{cases} \tag{5}$$

In the quality domain, it is easier for most of the objective models to predict the quality scores for pairs that are of significantly different quality [19, 20]. If the visual difference of two considered stimuli is high enough for observers to differentiate them, the obtained subjective data would be of less bias and errors [21], and thus cause less uncertainty for the prediction of objective quality models [22]. Therefore, following a similar concept of hard sample defined in [23], for quality assessment of gaming video, we consider the pairs that have similar quality scores as hard samples. Intuitively, l_2 is a gating function, where only hard sample pairs that have close quality scores (*i.e.*,$0 < |y_i - y_j| \leq \tau$) are taken into account, and $L^{ij}_{hard-rank}$ is thus named as the 'hard pair-wise ranking loss'. In this study, we first normalize the ground truth quality scores into a range of $[0, 1]$, and set $\tau = 0.1$ empirically.

4.3. Model Distillation

In general, when using ResNet-like models following a de-facto standard, the network is commonly constructed by stacking a certain number of blocks. These models could be divided into 4 stages, where each stage contains a certain number of repeated residual blocks following by a sequence of convolution layers that downsample the input feature. For ResNet18, there are 2 residual block at each stage. Obviously, the more blocks there is, the more sophisticated the model is, *i.e.*, more parameters. Knowledge distillation [24, 25] can transfer knowledge from a complex model, *i.e.*, the teacher model, to a lighter version of itself, *i.e.*, the student model. This 'teacher-student' model compression technique, is of great potential to obtain lighter network with more stationary performance, as training a small model from the scratch often ends up encountering issues like underfitting.

The concise overview of our re-adapted teacher-student training framework for model distillation is depicted in Fig. 3. In the framework of the proposed objective model, the backbone ResNet18 is utilized as the teacher model within the knowledge distillation pipeline. In concrete words, the backbone ResNet18 was first trained until the performance no longer increases, and was freezed during the distillation procedure, as shown in the upper part of Fig. 3.

Afterwards, ResNet18-tiny was used as the student model, where there was only 1 block at each stage and the remaining architecture is the same as original ResNet18, as presented in the lower part of Fig. 3. Unlike the default setting considered in common classification task, essentially, our model is a regression model. Therefore, the knowledge distillation loss proposed in [24] is not applicable in our use case. Hence, the Knowledge distillation (KD) loss l_{kd} is re-adapted as:

$$l_{kd} = |\hat{y}_s - \hat{y}_t|, \tag{6}$$

where $\hat{y}_s$ and $\hat{y}_t$ are the predicted scores from student model and teacher model respectively.

It is worth noting that, throughout the knowledge distillation process, the student network does not only learn how to

predict better the quality score, but also the latent representation using the intermediate convolution layers. Inspired by the attention mechanism proposed in [25], their attention loss is re-adapted to distillate intermediate feature maps between the student model and teacher model. Let $\{\mathbf{A}\}$ be the set of feature map output from the aforementioned 4 stages in the ResNet network, and A^i denotes feature map output from the i_{th} stage. Then, we have:

$$\widetilde{A}^i = \sum_{i=1}^{C^i} |A^i|, \tag{7}$$

where C^i is the i_{th} channel of A^i. The attention loss used for transfer attention could be then defined as:

$$l_{at}^i = \frac{1}{2n^i} \cdot \left\| \frac{vec(\widetilde{A}_{stu}^i)}{||vec(\widetilde{A}_{stu}^i)||_2} - \frac{vec(\widetilde{A}_{tea}^i)}{||vec(\widetilde{A}_{tea}^i)||_2} \right\|_2, \tag{8}$$

where $vec(\cdot)$ is the vectorization operation, n^i represents the total number of nodes in A^i. Finally, the total loss L_{T-S} for training the student model could be written as:

$$L_{T-S} = L_{org} + l_{kd} + \frac{1}{4}\sum_{i=1}^{4} l_{at}^i \tag{9}$$

where L_{org} is the distillation-free loss given in equation (1).

5. EXPERIMENT

The proposed TGV dataset was employed to benchmark the NR metrics designed for quality assessment of gaming videos as summarized in section 2. The entire dataset was divided into 80%, 20% as training and hold-out testing sets respectively. It is worth mentioning that there is no overlap in terms of gaming contents between the training and testing set to ensure the generality of the model. Apart from those NR metrics, 4 commonly used, especially in treaming industry [2], full reference metrics were also tested, including the Peak Signal to Noise Ratio (PSNR) [26], the Structural Similarity Index (SSIM) [27], the Multi-Scale Structural Similarity (MS-SSIM) [28], and the Video Multi-Method Assessment Fusion (VMAF) [29]. To evaluate the performances of the considered objective quality metrics, commonly used performance evacuation methodology including the Pearson correlation coefficient (PCC), the Spearman's rank order correlation coefficient (SCC), the Kendall correlation coefficient (KCC) and the Root Mean Squared Error (RMSE) are computed between the mean opinion scores and the predicted quality scores.

During fine-tuning (training), all the frames extracted from videos were first rescaled to 1080P with zero paddings. To augment the data, the input frames were further randomly cropped into 540×960, and flipped left to right or vice versa. The initial learning rate was set as 10^{-04}, and the Cosine Annealing learning rate decay strategy was applied. We used the Adam optimizer, and the training was stopped after 50 epochs. To avoid overfitting, weight decay was set as 5^{-04}. λ in equation (1) was set to 1.

Overall performance: The results are shown in Table 5, where the best performances are highlighted in bold. ERAQUE$_{teacher}$, ERAQUE$_{student}$ and ResNet18 tiny, are the original teacher backbone model, the student model after distillation and the ResNet18 tiny that is trained from scratch correspondingly. In general, the proposed ResNet18$_{teacher}$ achieves the best performance among all the compared metrics, including the full reference metrics. It even far surpasses the other CNN based state-of-the-art metric NDNetgaming, with PCC values of 0.9646 $vs.$ 0.8579. The distilled version of our model also outperforms the other metrics. These observations demonstrate the effectiveness of our approach. The performances of the two ResNet18 tiny models also elucidate the advantage of training light-weight models via knowledge distillation.

Table 2. Performances comparison of considered Metrics.

	PCC	**SCC**	**KCC**	**RMSE**
Full Reference Metric (FR)				
PSNR [26]	0.7467	0.7348	0.5445	0.7077
SSIM [27]	0.8709	0.8654	0.6763	0.5229
MS-SSIM [28]	0.8492	0.8457	0.6536	0.5618
VMAF [29]	0.9130	0.9102	0.7436	0.4341
No Reference Metric (NR)				
GamingPara [11]	0.5568	0.4402	0.3268	0.9084
ITU-T G.1072 [10]	0.0633	0.0534	0.0412	1.0628
DEMI [4]	0.7536	0.7455	0.5382	0.7479
NDNetgaming [12]	0.8579	0.8477	0.6578	0.5845
ResNet18 tiny	0.8298	0.8219	0.6398	2.6031
ERAQUE$_{student}$	0.9646	0.9641	0.8436	2.6387
ERAQUE$_{teacher}$	**0.9714**	**0.9712**	**0.8635**	**0.2697**

Advantage of knowledge distillation: Information of the teacher, student networks are summarized in Table 5, where $\#para$ indicates the number of network parameters. It is obvious that the complexity of the student network is around half of the teacher model in any respect. To further verify whether the difference of performance between the teacher and student is significant, *i.e.,* whether the predicted scores are significantly different, the F-test analysis based on the residual difference between the predicted objective scores and the subjective Mean Opinion Score (MOS) values as described in [30] was employed. The obtained results indicate no significant different between the performances of the proposed model and the distilled model, which demonstrates the efficiency of the teacher-student model distillation strategy.

Table 3. Comparison of teacher, student models.

	$\#para$	**FLOPs**	**storage size of the model**
ERAQUE$_{teacher}$	11 M	18.91 G	45 M
ERAQUE$_{student}$	5.4 M	9.28 G	21 M

6. CONCLUSION

In this study, a large-scale subjective study for the quality assessment of mobile gaming videos. According to our best knowledge, our TGV is one of the largest existing relevant

dataset. To better quantify the quality of mobile gaming contents, a novel quality assessment framework has also been presented. According to the experimental results, the proposed metrics outperform state-of-the-art quality metrics, and our model distillation strategy could achieve high trade-off between the model complexity and the performance.

7. REFERENCES

[1] Nabajeet Barman, Saman Zadtootaghaj, Steven Schmidt, Maria G Martini, and Sebastian Möller, "Gamingvideoset: a dataset for gaming video streaming applications," in *2018 16th Annual Workshop on Network and Systems Support for Games (NetGames)*. IEEE, 2018, pp. 1–6.

[2] Suiyi Ling, Yoann Baveye, Deepthi Nandakumar, Sriram Sethuraman, and Patrick Le Callet, "Towards better quality assessment of high-quality videos," in *Proceedings of the 1st Workshop on Quality of Experience (QoE) in Visual Multimedia Applications*, 2020, pp. 3–9.

[3] Arthur Zuckerman, "13 CURRENT GAMING TRENDS: 2020/2021 DATA, STATISTICS & PREDICTIONS," `https://comparecamp.com/gaming-trends/#2`, 2020, [Online; accessed 19-Jan-2020].

[4] Saman Zadtootaghaj, Nabajeet Barman, Rakesh Rao Ramachandra Rao, Steve Göring, Maria G Martini, Alexander Raake, and Sebastian Möller, "Demi: deep video quality estimation model using perceptual video quality dimensions," in *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2020, pp. 1–6.

[5] Suiyi Ling, Junle Wang, Wenming Huang, Yundi Guo, Like Zhang, Yanqing Jing, and Patrick Le Callet, "A subjective study of multi-dimensional aesthetic assessment for mobile game image," in *Proceedings of the 1st Workshop on Quality of Experience (QoE) in Visual Multimedia Applications*, 2020, pp. 47–53.

[6] Nabajeet Barman, Emmanuel Jammeh, Seyed Ali Ghorashi, and Maria G Martini, "No-reference video quality estimation based on machine learning for passive gaming video streaming applications," *IEEE Access*, vol. 7, pp. 74511–74527, 2019.

[7] Saman Zadtootaghaj, Steven Schmidt, Saeed Shafiee Sabet, Sebastian Möller, and Carsten Griwodz, "Quality estimation models for gaming video streaming services using perceptual video quality dimensions," in *Proceedings of the 11th ACM Multimedia Systems Conference*, 2020, pp. 213–224.

[8] Markus Utke, Saman Zadtootaghaj, Steven Schmidt, and Sebastian Möller, "Towards deep learning methods for quality assessment of computer-generated imagery," *arXiv preprint arXiv:2005.00836*, 2020.

[9] Saman Zadtootaghaj, Nabajeet Barman, Steven Schmidt, Maria G Martini, and Sebastian Möller, "Nr-gvqm: A no reference gaming video quality metric," in *2018 IEEE International Symposium on Multimedia (ISM)*. IEEE, 2018, pp. 131–134.

[10] "Itu-t recommendation g.1072: Opinion model predicting gaming quality of experience for cloud gaming services," in *International Telecommunication Union*. ITU, 2020.

[11] Saman zadtootaghaj, Steven Schmidt, Saeed Shafiee Sabet, Sebastian Moeller, and Carsten Griwodz, "Quality estimation models for gaming video streaming services using perceptual video quality dimensions," in *Proceedings of the 11th International Conference on Multimedia Systems*. ACM, 2020.

[12] Markus Utke, Saman Zadtootaghaj, Steven Schmidt, Sebastian Bosse, and Sebastian Moeller, "NDNetGaming - Development of a No-Reference Deep CNN for Gaming Video Quality Prediction," in *Multimedia Tools and Applications*. Springer, 2020.

[13] ITU, "Methodology for the subjective assessment of the quality of television pictures," *Recommendation ITU-R BT.500*, 2012.

[14] Suiyi Ling, Jing Li, Zhaohui Che, Junle Wang, Wei Zhou, and Patrick Le Callet, "Re-visiting discriminator for blind free-viewpoint image quality assessment," *IEEE Transactions on Multimedia*, 2020.

[15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[16] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[17] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6848–6856.

[18] Xialei Liu, Joost van de Weijer, and Andrew D Bagdanov, "Rankiqa: Learning from rankings for no-reference image quality assessment," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1040–1049.

[19] Suiyi Ling, Jing Li, Anne Flore Perrin, Zhi Li, Lukáš Krasula, and Patrick Le Callet, "Strategy for boosting pair comparison and improving quality assessment accuracy," *arXiv preprint arXiv:2010.00370*, 2020.

[20] Jing Li, Rafal Mantiuk, Junle Wang, Suiyi Ling, and Patrick Le Callet, "Hybrid-mst: A hybrid active sampling strategy for pairwise preference aggregation," in *Advances in neural information processing systems*, 2018, pp. 3475–3485.

[21] Jing Li, Suiyi Ling, Junle Wang, Zhi Li, and Patrick Le Callet, "Gpm: A generic probabilistic model to recover annotator's behavior and ground truth labeling," *arXiv preprint arXiv:2003.00475*, 2020.

[22] Suiyi Ling, Yoann Baveye, and Patrick Le Callet, "Rate-distortion video coding and uncertainties: to be blindly chasing marginal improvement or to be greener," in *Applications of Digital Image Processing XLIII*. International Society for Optics and Photonics, 2020, vol. 11510, p. 115100F.

[23] Suiyi Ling, Andreas Pastor, Jing Li, Zhaohui Che, Junle Wang, Jieun Kim, and Patrick Le Callet, "Few-shot pill recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9789–9798.

[24] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[25] Sergey Zagoruyko and Nikos Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," *arXiv preprint arXiv:1612.03928*, 2016.

[26] Zhou Wang and Alan C Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *IEEE signal processing magazine*, vol. 26, no. 1, pp. 98–117, 2009.

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

[28] Zhou Wang, Eero P Simoncelli, and Alan C Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. Ieee, 2003, vol. 2, pp. 1398–1402.

[29] Zhi Li, Christos Bampis, Julie Novak, Anne Aaron, Kyle Swanson, Anush Moorthy, and JD Cock, "Vmaf: The journey continues," *Netflix Technology Blog*, 2018.

[30] Suiyi Ling, Jesús Gutiérrez, Ke Gu, and Patrick Le Callet, "Prediction of the influence of navigation scan-path on perceived quality of free-viewpoint videos," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 1, pp. 204–216, 2019.