

A STRONG BASELINE FOR IMAGE AND VIDEO QUALITY ASSESSMENT

Shaoguo Wen¹, Junle Wang¹

¹Turing Lab, Tencent

ABSTRACT

In this work, we present a simple yet effective unified model for perceptual quality assessment of image and video. In contrast to existing models which usually consist of complex network architecture, or rely on the concatenation of multiple branches of features, our model achieves a comparable performance by applying only one global feature derived from a backbone network (i.e. resnet18 in the presented work). Combined with some training tricks, the proposed model surpasses the current baselines of SOTA models on public and private datasets. Based on the architecture proposed, we release the models well trained for three common real-world scenarios: UGC videos in the wild, PGC videos with compression, Game videos with compression. These three pre-trained models can be directly applied for quality assessment, or be further fine-tuned for more customized usages. All the code, SDK, and the pre-trained weights of the proposed models are publicly available at <https://github.com/Tencent/CenseoQoE>.

Index Terms— Image quality assessment, Video quality assessment, Quality of experience, Perceptual quality

1. INTRODUCTION

Image/Video quality assessment(I/VQA) have been a long-standing problem in image/video processing and computer vision, always used as a measurement or optimization target in the fields of video compression, quality monitoring, video recommendation systems and etc. Nowadays, user-generated content (UGC) and video streaming has exploded on the Internet, the enormous amount of video storage and transmission poses new challenges to the size of the video, I/VQA can provide measurement for encoders to reduce the bit rate of the video or compression algorithms to compress the video with little or no perceptual impact on the video quality. Another novel application is used in the recommendation system to provide users with higher quality videos. Taking advantage of such optimizations allows for better user experience at lower cost for the provider which shows great value.

Generally, quality assessment can be categorized into subjective assessment and objective assessment. Subjective assessment usually requires a certain number(15 at least according to ITU-R BT.500 [1]) of people to evaluate the quality

of image or video, then mean of opinions(MoS) is regarded as the final quality score. Subjective assessment always obtain reliable and accurate results for quality assessment, however, it is too expensive and time-consuming to be used in the quality evaluation of visual systems that requires frequent and real-time feedback. The objective assessment predicts a quality score by algorithms that aims to correlate well with human perception which is significantly more practical for real-time image/video quality evaluation.

Many efforts have been made on developing objective algorithms of image/video assessment. The Peak Signal to Noise Ratio (PSNR) [2], the Structural Similarity Index (SSIM) [3] and the Multi-Scale Structural Similarity (MS-SSIM) [4] are usually used as traditional methods for image assessment, but they are not correlate well with human perceptual quality sometimes. The Video Multi-Method Assessment Fusion (VMAF) [5] take use of hand-crafted features and machine learning to generate model to predict quality of videos, but which is limited to when the reference video is available. Deep learning have achieved great success on computer vision in recent years, many works[6][7][8][9] applied CNN and RNN to tackle the problem of I/VQA and achieve high performance. However, we found that many previous works were expanded on poor baselines, besides, the comparison between methods is unfair because some of them obtained the improvement by training with tricks rather than proposed methods themselves. In addition, many state-of-the-arts (SOTA) models designed complicated network architecture which is not suitable for industrial deployment. In this paper, we proposed a simple and effective unified model for image/video quality assessment which acquires a strong baseline by training with some common tricks.

To demonstrate the performance of our method, we conduct experiments on three publicly available databases, i.e., LIVE-VQC [10], KoNViD-1K[11] and YouTube-UGC[12], and three private datasets in different commercial scenarios, i.e., UGC videos in the wild, PGC videos with compression and Game videos with compression.

The main contributions of this work are as follows:

- For the academia, we hope the strong baseline provided by our proposed method help researchers to design more excellent models and achieve higher performance in the I/VQA community.

- For the industry, the model we proposed is simple but high in performance without extra inference consumption, which is useful for industrial deployment of IVQA models to achieve the goal of real-time feedback.
- We release three model weights of our proposed method that have been trained on three carefully designed datasets of different real-world commercial scenarios, which can be directly used for quality assessment or fine tuned on own dataset.

2. RELATED WORK

2.1. Image Quality Assessment

Image Quality Assessment(IQA) can be classified into distortion-specific methods and general-purpose methods according to [13]. The distortion-specific methods[14][15] evaluated the image quality by extracting features of known distortion types, but their application scope is limited because the distortion types are always unknown or mixture. The general-purpose methods are further divided into Natural Scene Statistics (NSS) methods and learning-based methods. The NSS methods extracted features in different sub-bands and estimate the distributional parameters for predicting quality. In learning-based methods[16][17][18], features are extracted and mapped to the MOS by Support Machine Regression or Neural Networks. Deep learning based methods have been developed by many works in recent years which resulted in significant improvements and showed great potential. Kang et al.[19] applied CNN to train IQA model with small image patches rather than images, which improved the performance of the model by augmenting training examples. Liu et al.[6][20] combined CNN with ranking learning to further improve the performance of models. Hossein Talebi and Peyman Milanfar[21] proposed a novel approach to predict both technical and aesthetic qualities of image. Zhu et al.[13] proposed a no-reference IQA metric based on deep meta-learning which tried to learn the meta-knowledge shared by human when evaluating the quality of images with various distortions.

2.2. Video Quality Assessment

Most early VQA models were distortion specific [22][23][24] and focused mostly on transmission and compression related artifacts. Li et al. proposed a learning-based method for FR-VQA named Video Multi-Method Assessment Fusion (VMAF) [5] which extracts features from videos and trains a Support Vector Machine(SVM) model to predict quality of videos. Similiar with IQA, deep learning-based methods obtained promising results in VQA in recent years, Kim et al. [25] utilize CNN models to learn the spatial-temporal sensitivity maps. Liu et al. [26] exploit a 3D-CNN model for codec classification and quality assessment of compressed

videos. Wang et al.[9] create a large scale UGC video dataset and propose a DNN-based framework to thoroughly analyze importance of content, technical quality and compression level in perceptual quality. Tu et al. [7] proposed an efficient model for predicting the subjective quality of UGC videos which leverages a composite of spatio-temporal scene statistics features and deep CNN-based high-level features. Ying et al. [8] created a largest(by far) in the wild UGC video quality dataset and proposed two unique NR-VQA models: a local-to-global region-based NR VQA architecture and a first-of-a-kind space-time video quality mapping engine.

3. THE PROPOSED OBJECTIVE MODEL

3.1. Network Architecture

Conventionally, image/video quality assessment(IVQA) can be divided into three main categories:full-reference (FR), reduced-reference (RR), and no-reference (NR) models. FR model predicts quality score against pristine image/video, while no-reference (NR) model involve no such comparison.

A simple but effective network architecture is proposed for FR and NR models in this paper as depicted in Fig. 1 and Fig. 2 respectively. A light-weight network is applied as the backbone of proposed model for efficient inference, such as Mobilenet [27], Shufflenet [28], ResNet-18[29] etc. The output feature of last convolution layer in the backbone is fed into the Global Averaged Pooling(GAP) [30] module, two Fully Connected (FC) layers with 1024 hidden nodes take the flatten feature obtained by GAP as inputs and predict the final quality score. For image quality assessment, image to be evaluated is fed into the network directly and obtain the quality score. For video quality assessment, video is first extracted and each frame is fed into the model, and then the scores of all frames are frames-wise averaged to obtain the final quality score. The main difference between our proposed NR and FR models is in the input of the network, the NR model take the distortion image as input directly, however, in the FR model, the reference image is first subtracted by the distortion image, the result of subtraction is concatenated with the distortion image and then fed into the FR network. Note that the dimension of first convolution layer should modified from 3 to 6 due to the concatenation in the FR model. In our experiment, the backbone of our model is pre-trained on Imagenet and fine tuned on quality data.

3.2. Loss function

We denote \hat{y} as the predicted score by the objective model, and let y be the ground truth quality score collected from the subjective experiment. n is the batch size of input images in the training phase. 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:

$$L = L_{mae} + \lambda \cdot L_{rank}. \quad (1)$$

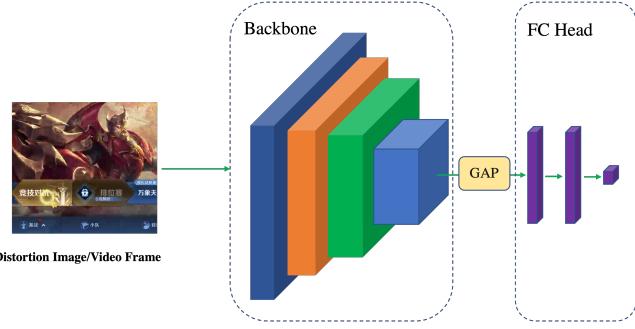


Fig. 1. Network Architecture of NR model.

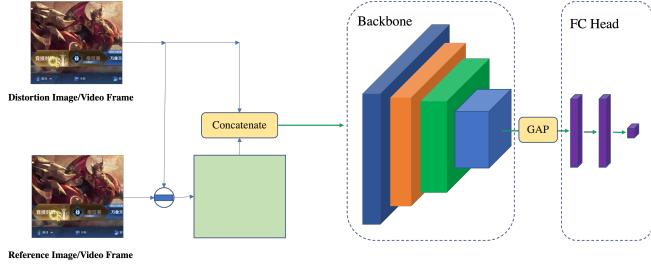


Fig. 2. Network Architecture of FR model.

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

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

The second part is a pair-wise ranking loss L_{rank}^{ij} which is inspired by the metric learning of image quality assessment proposed in [6]. Different with [6], instead of synthetically generating deformations of images over a range of distortion intensities, we apply rank learning in the training data. Specifically, given arbitrary pair of images in the batch inputs, the proposed L_{hard}^{ij} is designed as:

$$L_{rank}^{ij} = \max(0, |y_i - y_j| - e(y_i, y_j) \cdot (\hat{y}_i - \hat{y}_j)), \quad (3)$$

where $e(y_i, y_j)$ is defined as:

$$e(y_i, y_j) = \begin{cases} 1, & y_i \geq y_j \\ -1, & \text{otherwise} \end{cases} \quad (4)$$

Finally, L_{rank} is calculated by:

$$L_{rank} = \frac{1}{n \cdot n} \sum_{i=1}^n \sum_{j=1}^n L_{rank}^{ij} \quad (5)$$

L_{rank} help model capture more detailed information over different distortion or different degrees under same distortion, besides, which can also speed up the convergence of the model.

3.3. Training Tricks

Cosine annealing learning rate decay proposed in SGDR[31] is applied as learning rate schedule in our training phase, only the cosine annealing part is implemented without the restarts part. Supposed that T is the max training epochs, l_{init} is the initial learning rate at the beginning of training, and l_{min} is the minimum learning rate at the end of training, then the decaying of learning rate over training is denoted as:

$$L(t) = l_{min} + \frac{1}{2}(l_{init} - l_{min})(1 + \cos(\frac{\pi * t}{T})) \quad (6)$$

Where t is the current epoch.

Unlike common computer vision tasks, whose inputs can be resized as any shape for fitting the input shape of model, but in the image/video quality assessment tasks, images should be resized with same ratio in avoid to introduce unnecessary distortion and mislead the training of model. According to the setting of Quality of Experiment, we resize short size of image to the max resolution of all quality data, e.g. 1080p, then random cropping is employed on the resized image to match the input shape of model. Resizing and random cropping is a kind of data augmentation which can improve the performance of model. Note that center cropping is applied in the test phase.

The Adam and SGD with momentum optimizer are both implemented in our experiment. Empirically, Adam optimizer is suitable for training from scratch or pre-trained on Imagenet, which can accelerate the convergence of model. SGD optimizer with momentum is suitable for training with pre-trained on quality data, which can improve the robustness of model, especially when evaluating across datasets.

Stochastic Weight Averaging (SWA) [32] is optionally implemented in our experiment, the key idea of SWA is to average multiple model weights produced by SGD with a modified learning rate schedule, which can reach a wider optima for better generalization. Note that SWA is not plugged in when compared with other methods. SWA usually obtains a better generalization without increasing the complexity of the model.

4. EXPERIMENT

4.1. Experimental setup

During training, in addition to the resizing and random cropping introduced in section 3.3, randomly flipped left to right is further implemented for data augmentation. The initial learning rate l_{init} and minimum learning rate l_{min} are set as 10^{-4} and 10^{-7} respectively in the Cosine Annealing learning rate decay strategy. To avoid over-fitting, weight decay was set as 5^{-4} and the momentum is set to 0.9 if SGD optimizer is applied. λ in equation (1) was set to 1.

We conduct evaluation experiment of our proposed model on the three public UGC-VQA databases: LIVE-VQC [10], KoNViD-1K[11] and YouTube-UGC[12]. All the datasets

Table 1. Performances comparison on public datasets.

	LIVE-VQC		KoNViD-1K		YouTube-UGC	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
VGG-19[33]	0.7160	0.6568	0.7845	0.7741	0.6997	0.7025
ResNet-50[33]	0.7205	0.6636	0.8104	0.8018	0.7097	0.7183
RAPIQUE[7]	0.7863	0.7548	0.8175	0.8031	0.7684	<u>0.7591</u>
PatchVQ[8]	0.7205	0.6636	0.837	0.827	-	-
CoINVQ[9]	-	-	0.767	0.764	0.802	0.816
Ours	<u>0.7575</u>	0.7390	<u>0.8245</u>	0.8185	<u>0.7691</u>	0.7554

are randomly split into non-overlapping training and test sets (80%/20%), this process of random split was repeat 20 times and the overall median performance was recorded. It's also worth noting that different videos belonging to the same reference video should be contained in the same train set or test set. Besides, we also evaluate our model on three private datasets in different commercial scenarios: UGC videos in the wild, PGC videos with compression and Game videos with compression, the results demonstrates the effective and high performances of our model. To evaluate the performances of model, the Pearson Linear Correlation Coefficient (PLCC) and Spearman's rank order correlation coefficient (SRCC) are considered as evaluation metrics.

4.2. Experimental results

The results on public datasets are shown in Table 1, where the best and second-best results are respectively marked in bold and underlined fonts. As shown in the table, our model obtain second-best performance of PLCC and SRCC in LIVE-VQC and KoNViD-1K datasets and second-best performance of PLCC in YouTube-UGC. The gap of performance between SOTA models and our model is not very large but our model has fewer parameters and faster inference speed. Besides, the performance of our proposed model far exceeds those models that are often used as baselines, e.g. VGG-19 and ResNet-50. Our strong baseline can achieve 0.7575 in PLCC and 0.7390 in SRCC in LIVE-VQC dataset, which beats standard baseline VGG-19[33] by more than 0.04 in PLCC and 0.08 in SRCC.

To evaluate the performance of our model in real commercial scenario, we conduct subjective experiment under the ITU-R BT.500 [1] standard and build three datasets for training model according to our commercial scenarios, i.e. UGC videos in the wild, PGC videos with compression and Games videos with compression. We first collected and processed a certain number of videos in three commercial scenario data, specifically, around 3000 videos with compression for PGC and Games videos respectively and more than 20000 videos in the wild for UGC. During the experiment, 20-25 were asked to score the videos. 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. The performance of our

model is shown on Table 2, our model achieves high PLCC and SRCC in all three private datasets. Unfortunately, these datasets are not yet publicly available due to some reasons, but models trained on these datasets already available at <https://github.com/Tencent/CenseoQoE>.

Table 2. Performances on private datasets.

	PLCC	SRCC
Games videos(compression)	0.971	0.968
PGC videos(compression)	0.961	0.959
UGC videos(in the wild)	0.902	0.880

5. CONCLUSION

In this study, we propose an efficient and high-performance unified model for image/video quality assessment, which achieved comparable or even surpassing performance in some datasets compared with state-of-arts models and obtained a strong baseline in many public I/VQA datasets. Our proposed model achieves a high trade-off between performance and complexity, so it is very suitable for industrial deployment. Besides, three trained model based on our proposed method are released, which can be directly used for quality assessment in the real-world commercial scenario or fine tuned on own dataset.

6. REFERENCES

- [1] RECOMMENDATION ITU-R BT, "Methodology for the subjective assessment of the quality of television pictures," *International Telecommunication Union*, 2002.
- [2] 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.
- [3] 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.
- [4] Zhou Wang, Eero P Simoncelli, and Alan C Bovik, "Multiscale structural similarity for image quality assessment," in *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. Ieee, 2003, vol. 2, pp. 1398–1402.
- [5] Zhi Li, Christos Bampis, Julie Novak, Anne Aaron, Kyle Swanson, Anush Moorthy, and JD Cock, "Vmaf: The journey continues," *Netflix Technology Blog*, 2018.
- [6] Xialei Liu, Joost van de Weijer, and Andrew D Bagdanov, "Rankiq: Learning from rankings for no-reference image quality assessment," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1040–1049.
- [7] Zhengzhong Tu, Xiangxu Yu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik, "Rapique: Rapid and accurate video quality prediction of user generated content," *arXiv preprint arXiv:2101.10955*, 2021.

- [8] Zhenqiang Ying, Maniratnam Mandal, Deepti Ghadiyaram, and Alan Bovik, “Patch-vq: patching up the video quality problem,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14019–14029.
- [9] Yilin Wang, Junjie Ke, Hossein Talebi, Joong Gon Yim, Neil Birkbeck, Balu Adsumilli, Peyman Milanfar, and Feng Yang, “Rich features for perceptual quality assessment of ugc videos,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13435–13444.
- [10] Zeina Sinno and Alan Conrad Bovik, “Large-scale study of perceptual video quality,” *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 612–627, 2018.
- [11] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe, “The konstanz natural video database (konvid-1k),” in *2017 Ninth international conference on quality of multimedia experience (QoMEX)*. IEEE, 2017, pp. 1–6.
- [12] Yilin Wang, Sasi Inguva, and Balu Adsumilli, “Youtube ugc dataset for video compression research,” in *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2019, pp. 1–5.
- [13] Hancheng Zhu, Leida Li, Jinjian Wu, Weisheng Dong, and Guangming Shi, “Metaiqa: Deep meta-learning for no-reference image quality assessment,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14143–14152.
- [14] Leida Li, Weisi Lin, Xuesong Wang, Gaobo Yang, Khosro Bahrami, and Alex C Kot, “No-reference image blur assessment based on discrete orthogonal moments,” *IEEE transactions on cybernetics*, vol. 46, no. 1, pp. 39–50, 2015.
- [15] Leida Li, Hancheng Zhu, Gaobo Yang, and Jiansheng Qian, “Referenceless measure of blocking artifacts by tchebichef kernel analysis,” *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 122–125, 2013.
- [16] Aladine Chetouani, Azeddine Beghdadi, Shaohua Chen, and Ghilés Mostafaoui, “A novel free reference image quality metric using neural network approach,” in *Proc. Int. Workshop Video Process. Qual. Metrics Cons. Electrn*, 2010, pp. 1–4.
- [17] Peng Ye and David Doermann, “No-reference image quality assessment using visual codebooks,” *IEEE Transactions on Image Processing*, vol. 21, no. 7, pp. 3129–3138, 2012.
- [18] Peng Ye, Jayant Kumar, Le Kang, and David Doermann, “Unsupervised feature learning framework for no-reference image quality assessment,” in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 1098–1105.
- [19] Le Kang, Peng Ye, Yi Li, and David Doermann, “Convolutional neural networks for no-reference image quality assessment,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1733–1740.
- [20] Xiaolei Liu, Joost Van De Weijer, and Andrew D Bagdanov, “Exploiting unlabeled data in cnns by self-supervised learning to rank,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1862–1878, 2019.
- [21] Hossein Talebi and Peyman Milanfar, “Nima: Neural image assessment,” *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3998–4011, 2018.
- [22] Savvas Argyropoulos, Alexander Raake, Marie-Neige Garcia, and Peter List, “No-reference video quality assessment for sd and hd h.264/avc sequences based on continuous estimates of packet loss visibility,” in *2011 Third International Workshop on Quality of Multimedia Experience*. IEEE, 2011, pp. 31–36.
- [23] Katerina Pandremmenou, Muhammad Shahid, Lisisimachos P Kondi, and Benny Lövström, “A no-reference bitstream-based perceptual model for video quality estimation of videos affected by coding artifacts and packet losses,” in *Human Vision and Electronic Imaging XX*. International Society for Optics and Photonics, 2015, vol. 9394, p. 93941F.
- [24] Maria Torres Vega, Decebal Constantin Mocanu, Stavros Stavrou, and Antonio Liotta, “Predictive no-reference assessment of video quality,” *Signal Processing: Image Communication*, vol. 52, pp. 20–32, 2017.
- [25] Woojae Kim, Jongyoo Kim, Sewoong Ahn, Jinwoo Kim, and Sanghoon Lee, “Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 219–234.
- [26] Wentao Liu, Zhengfang Duanmu, and Zhou Wang, “End-to-end blind quality assessment of compressed videos using deep neural networks,” in *ACM Multimedia*, 2018, pp. 546–554.
- [27] 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.
- [28] 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.
- [29] 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.
- [30] 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.
- [31] Ilya Loshchilov and Frank Hutter, “Sgdr: Stochastic gradient descent with warm restarts,” *arXiv preprint arXiv:1608.03983*, 2016.
- [32] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson, “Averaging weights leads to wider optima and better generalization,” *arXiv preprint arXiv:1803.05407*, 2018.
- [33] Zhengzhong Tu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik, “Ugc-vqa: Benchmarking blind video quality assessment for user generated content,” *IEEE Transactions on Image Processing*, vol. 30, pp. 4449–4464, 2021.