TEXTURE INFORMATION BOOSTS VIDEO QUALITY ASSESSMENT

Ao-Xiang Zhang and Yuan-Gen Wang[*]

School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou, China

ABSTRACT

Automatically evaluating the quality of in-the-wild videos is challenging since both the distortion types and reference videos are unknown. In general, humans can make a fast and accurate judgment for video quality. Fortunately, deep neural networks have been developed to effectively model the human visual system (HVS). In this paper, we deeply investigate three elements of HVS, including texture masking, content-dependency, and temporal-memory effects from an experimental perspective. Based on the investigation, we propose to make full use of texture information to boost the performance of video quality assessment (VQA), termed TiVQA in this paper. To be specific, TiVQA first uses the local binary pattern (LBP) operator to detect texture information of each video frame. Then a two-stream ResNet is employed to extract the texture masking and content-dependency embeddings, respectively. Finally, TiVQA integrates both the gated recurrent unit and subjectively-inspired temporal pooling layer to model the temporal-memory effects. Extensive experiments on benchmark datasets including KoNViD-1k, CVD2014, LIVE-Qualcomm, and LSVQ show that the proposed TiVQA obtains state-of-the-art performance in terms of SRCC and PLCC.

***Index Terms*—** Video quality assessment, human visual system, texture masking, content-dependency, temporal-memory effects.

1. INTRODUCTION

Nowadays, we are almost surrounded by digital videos due to the rapid advance of social media [1]. Digital videos have also produced a huge impact on our daily life. With the increasing demand on video quality from end users, video quality assessment (VQA) has become a major concern for video service providers. It is known that the mean opinion score (MOS) of a video is based on the human visual effect. With the success of deep learning in a broad range of application domains, using deep neural networks to simulate the visual characteristic of human eyes has been an important research topic in the VQA field.

In literature, Tu *et al.* [2] proposed to use binary and sequential classification for no-reference VQA (NR-VQA) at coarser levels. Wu *et al.* [3] developed an NR-VQA based on semantic information. Mitra *et al.* [4] used convolutional neural networks (CNN) to predict spatial and temporal entropic differences that can efficiently capture video distortions in space and time. Zhang *et al.* [5] found that spatio-temporal inconsistencies among observers have a significant impact on the reliability of subjective quality evaluation results. Galkandage *et al.* [6] developed a new full reference stereo video quality metric by simulating the HVS response. Researches have shown that HVS has many unique characteristics, such as texture masking, content-dependency, and temporal-memory effects. Particularly, the texture masking refers to the fact that people are more likely to ignore the distortion in the texture area of an image. This effect has been exploited in existing VQA work. For instance, Ma [7] *et al.* used texture masking to build a VQA model by extracting spatial information change. Gunawan and Ghanbari [8] developed a reduced-reference VQA using texture masking. For temporal-memory effects, the study [9] shows that individuals are more largely influenced by the previous frames with lower quality when judging the video quality. How to effectively model the characteristic of HVS is very helpful for the VQA problem.

In this paper, we propose a new NR-VQA method based on modeling of three salient characteristics of HVS, including texture masking, content-dependency, and temporal-memory effects. For the texture masking, we first employ the local binary pattern (LBP) operator to detect the texture image of each video frame, then both the detected texture image and the original video frame are fed to a two-stream ResNet for the content-dependency feature extraction. Finally, the gated recurrent unit (GRU) network and subjectively-inspired temporal pooling layer are adopted to compute the temporal-memory effects. Various experiments are conducted on benchmark datasets, and the experimental results show the superiority of the proposed method over the state-of-the-art.

2. THE PROPOSED METHOD

The framework of our method is shown in Fig. 1, which includes three major modules. In the texture information detection, the LBP operator [10] is used to detect the texture

[*]Email address of the corresponding author: wangyg@gzhu.edu.cn. This work was supported in part by the National Natural Science Foundation of China under Grant 61872099 and in part by the Scientific Research Project of Guangzhou University under Grant YJ2021004.

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

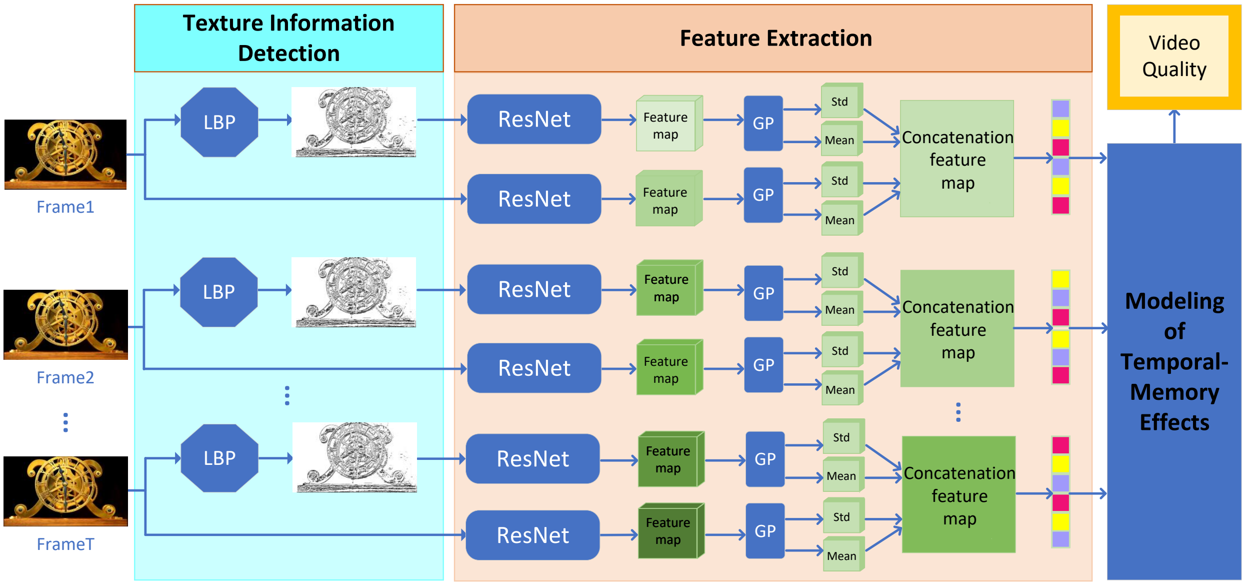

Fig. 1. The overall framework of the proposed method. It includes the texture information detection, the feature extraction, and the modeling of temporal-memory effects modules.

information of each video frame. Then in the feature extraction, the video frame and its texture frame are fed into the two pre-trained ResNet-50 networks on ImageNet [11] to extract the content-aware feature and the texture feature, respectively. In the modeling of temporal-memory effects, the feature vectors were put into the gated recurrent unit (GRU) network and subjectively-inspired temporal pooling layer to obtain the quality score of each video frame. Finally, we calculate the video quality score by linearly combining the min pooling and softmin-weighted average pooling of all the quality scores. In what follows we describe each module in detail.

Table 1. Experimental setup for different datasets.

Parameters	KoNViD-1K	CVD2014	LIVE-Qualcomm	LSVQ
Reduce size	248	192	248	208
Hidden size	64	52	64	56
Batch size	16	38	16	20
τ	44	68	54	20
β	0.52	0.5	0.4	0.5

2.1. Texture Information Detection

Human is more likely to ignore the distortion in the texture region than that in other regions, such as smooth and edge. While the current VQA models are only trained by a unique MOS, thus often cannot weight the different types of the distortions. That is to say, the distortions in different regions are identically captured, resulting in the inconsistency between the VQA model and HVS. This negatively affects the performance of existing VQA methods. Based on this observation, we propose to use the LBP operator to detect the texture image of each video frame for the extra training of deep networks so that the distortion in texture region can be moderately suppressed. According to our experiment, it is not optimal for the networks to capture too much or too little amount of texture information. To obtain good performance, the number of pixels (P) and the radius (R) of the LBP operator are respectively set to 10 and 4 in our method. Assuming that a video sequence has T frames, $\mathbf{I}_t$ denotes the tth frame of the video sequence, and $\mathbf{C}_t$ denotes the texture image of the tth frame. Then LBP operator can compute $\mathbf{C}_t$ by

$$\mathbf{C}_t = \mathrm{LBP}_{P,R}(\mathbf{I}_t), t = 1, 2, ..., T. \tag{1}$$

2.2. Spatial Feature Extraction

In general, images contain a wealth of spatial information (such as thing's features), which, to a large extent, represents the quality of a video. In this module, T video frames and the corresponding T texture images are respectively fed into two pre-trained CNN (i.e. ResNet) to obtain the deep semantic features (respectively denote $\mathbf{M}_{t1}$ and $\mathbf{M}_{t2}$ where $t1, t2 \in [T]$).

$$\mathbf{M}_{t1} = \mathrm{CNN}(\mathbf{I}_t), \tag{2}$$

$$\mathbf{M}_{t2} = \mathrm{CNN}(\mathbf{C}_t). \tag{3}$$

Next, we perform the widely acknowledged global average pooling $(\mathrm{GP_{mean}})$ and global standard deviation pooling $(\mathrm{GP_{std}})$ operations on the feature graphs $\mathbf{M}_{t1}$ and $\mathbf{M}_{t2}$ from the previous convolution layer to obtain the feature vectors $\mathbf{V}_{mean}^{t1}$,$\mathbf{V}_{std}^{t1}$, $\mathbf{V}_{mean}^{t2}$, and $\mathbf{V}_{std}^{t2}$:

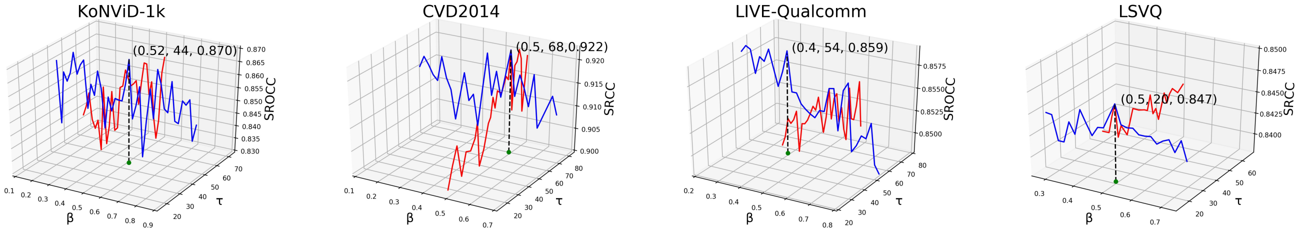

Fig. 2. SRCC performance change of the four datasets under different parameters.

$$\mathbf{V}_{mean}^{t1} = \mathrm{GP}_{\mathrm{mean}}(\mathbf{M}_{\mathrm{t1}}), \tag{4}$$

$$\mathbf{V}_{std}^{t1} = \mathrm{GP}_{\mathrm{std}}(\mathbf{M}_{\mathrm{t1}}), \tag{5}$$

$$\mathbf{V}_{mean}^{t2} = \mathrm{GP}_{\mathrm{mean}}(\mathbf{M}_{\mathrm{t2}}), \tag{6}$$

$$\mathbf{V}_{std}^{t2} = \mathrm{GP}_{\mathrm{std}}(\mathbf{M}_{\mathrm{t2}}). \tag{7}$$

Finally, we concatenate $\mathbf{V}_{mean}^{t1}$, $\mathbf{V}_{std}^{t1}$, $\mathbf{V}_{mean}^{t2}$, and $\mathbf{V}_{std}^{t2}$ as a feature map $\mathbf{V}_t$, which can be written by

$$\mathbf{V}_t = \mathrm{Concat}(\mathbf{V}_{\mathrm{mean}}^{\mathrm{t1}}, \mathbf{V}_{\mathrm{std}}^{\mathrm{t1}}, \mathbf{V}_{\mathrm{mean}}^{\mathrm{t2}}, \mathbf{V}_{\mathrm{std}}^{\mathrm{t2}}). \tag{8}$$

2.3. Temporal-Memory Modeling

As we know, MOS is completely based on human's empirical observation, which largely depends on the temporal-memory effects. Thus, modeling temporal memory plays a vital role on the VQA problem.

Extracting the temporal dependent frame quality. The time domain information can be extracted by a GRU operation [13]. Fully connected layer (FC) can be used to reduce the redundant features of $\mathbf{V}_t$. Thus we combine these two operations to obtain the temporal dependent frame level quality $\mathbf{q}_t$ by

$$\mathbf{q}_t = \mathrm{FC}\left(\mathrm{GRU}\{\mathrm{FC}(\mathbf{V}_t), \mathrm{GRU}\{\mathrm{FC}(\mathbf{V}_{t-1})\}\}\right), \tag{9}$$

where FC represents the fully connected layer.

Temporal pooling strategy. In general, the subjects have a deeper memory for the frames with poor quality than the high quality frames, resulting in lower expectations for the overall quality of the video. To this end, we adopt the temporal hysteresis [9] to the temporal quality pooling. First, we define the memory quality element $\mathbf{x}_n$ as the minimum value of τ previous frames:

$$\mathbf{x}_n = \begin{cases} q_n, & n = 1 \\ \min\limits_{k \in V_{pre}} \mathrm{q_k}, & n > 1 \end{cases} \tag{10}$$

where $V_{pre} = \{\max(1, \mathrm{t} - \tau), \cdots, \mathrm{t} - 2, \mathrm{t} - 1\}$ is the index set of the considered frames, and τ is a hyperparameter that determines how many previous frames are associated with. Then we define the current quality element $\mathbf{y}_t$ using a weighted quality fraction to assign larger weight to these frames with bad quality in τ next frames. A softmin function is used to determine the weight W_t^k as follows:

$$\mathbf{y}_t = \sum_{k \in V_{next}} q_k W_t^k, \tag{11}$$

$$W_t^k = \frac{e^{-q_k}}{\sum_{j \in V_{next}} e^{-q_j}}, k \in V_{next}, \tag{12}$$

where $V_{next} = \{t, t+1, \cdots, \min(\mathrm{t} + \tau, \mathrm{T})\}$, and τ is a hyperparameter that determines how many subsequent frames are associated with.

After considering the temporal hysteresis effect, we can compute the subjective frame scores q_t' by linearly combining the memory quality element with the current quality element. The final video quality Q can be obtained by calculating the average of the subjective frame quality scores as follows:

$$q_t' = \beta \mathbf{x}_t + (1 - \beta) \mathbf{y}_t, \tag{13}$$

$$Q = \frac{1}{T} \sum_{t=1}^{T} q_t', \tag{14}$$

where β is a balance factor between the weights of memory quality elements and the current quality elements.

3. EXPERIMENTAL RESULTS

Source code is available at https://github.com/GZHU-DVL /TiVQA. Our network is built on the Pytorch framework and trained on a machine equipped with four Tesla P100 GPUs. In order to verify the effectiveness of the proposed TiVQA, five mainstream VQA methods including BRISQUE [14], V-BLIINDS [15], TLVQM [16], VSFA [12], and PVQ [17] are selected for the experimental comparison. Four typical in-the-wild video datasets, namely KoNViD-1K [18], CVD2014 [19], LIVE-Qualcomm [20], and LSVQ [17] are used for testing. By convention, 80% of the dataset is for training, and the remaining 20% is for the test. L_1 loss function and Adam [21] optimizer with an initial learning rate of 10^{-5} are used in our method. The results are averaged over 10 runs in the same setup. Note that our method is closely related to the two factors (i.e. τ and β). When changing the values of the two parameters, the performance on different datasets will change. The changing curves are shown in Fig. 2. To obtain good performance, our method takes the optimal values of the two parameters, which are shown in Table 1.

Table 2. SRCC and PLCC performance comparison. The best performance is highlighted in bold. "–" indicates not applicable

Methods	KoNViD-1K		CVD2014		LIVE-Qualcomm		LSVQ	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
BRISQUE [14]	0.654	0.626	0.709	0.715	0.504	0.516	0.579	0.576
V-BLIINDS [15]	0.695	0.658	0.746	0.753	0.566	0.568	–	–
TLVQM [16]	0.773	0.769	0.830	0.850	0.780	0.810	0.772	0.774
VSFA [12]	0.755	0.744	0.880	0.885	0.737	0.732	0.801	0.796
PVQ [17]	0.791	0.786	–	–	–	–	0.827	0.828
TiVQA (Our)	**0.870**	**0.859**	**0.922**	**0.906**	**0.859**	**0.887**	**0.847**	**0.845**

Table 3. Performance comparison on cross datasets.

Training	KoNViD-1K						CVD2014					
Testing	CVD2014		LIVE-Qualcomm		LSVQ		KoNViD-1K		LIVE-Qualcomm		LSVQ	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
VSFA [12]	0.628	0.621	**0.557**	**0.577**	0.442	0.481	0.575	0.564	0.326	0.372	0.290	0.301
TiVQA (Our)	**0.752**	**0.748**	0.520	0.566	**0.513**	**0.548**	**0.666**	**0.652**	**0.357**	**0.401**	**0.347**	**0.369**
Training	LIVE-Qualcomm						LSVQ					
Testing	KoNViD-1K		CVD2014		LSVQ		KoNViD-1K		CVD2014		LIVE-Qualcomm	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
VSFA [12]	**0.664**	**0.612**	0.535	0.561	0.386	0.408	0.784	0.794	0.744	0.742	0.616	0.646
TiVQA (Our)	0.592	0.587	**0.564**	**0.586**	**0.391**	**0.416**	**0.796**	**0.805**	**0.764**	**0.746**	**0.623**	**0.653**

Table 4. Ablation study on texture information.

Texture feature	KoNViD-1K		CVD2014		LIVE-Qualcomm		LSVQ	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
×	0.755	0.744	0.880	0.885	0.737	0.732	0.801	0.796
✓	**0.870**	**0.859**	**0.922**	**0.906**	**0.859**	**0.887**	**0.847**	**0.845**

The overall performance comparison on four datasets are shown in Table 2. We can see from Table 2 that compared with five mainstream methods, our proposed TiVQA performs the best on all the four benchmark datasets. For instance, on the KoNViD-1K dataset, TiVQA obtains 10.0% SRCC and 9.3% PLCC improvement over its best competitor (i.e. PVQ [17]). To demonstrate the generalization of the proposed method, we conduct a cross-dataset experiment. The results are shown in Table 3. It can be seen from Table 3 that in 12 cross-dataset tests, the proposed TiVQA achieves better performance on 10 cross-dataset tests than its competitor. Only for the cross KoNViD-1K and LIVE-Qualcomm test, our method does not perform the best. This might be interpreted as follows: 1) The texture feature leveraged in our TiVQA is not suitable for these two cross datasets. 2) LIVE-Qualcomm includes only 208 videos with 1920×1080 resolution, while KoNViD-1K contains 1,200 videos with 960×540 resolution. The major difference in the resolution and number of samples may negatively affect the function of the texture information. Furthermore, to further demonstrate the effectiveness of texture information, we conduct an ablation study. The experimental results are shown in Table 4. We can see from Table 4 that compared with the method without using the texture masking, our TiVQA performs much better. This is due to the fact that TiVQA can be well trained to match the characteristic of texture masking, thereby greatly boosting the performance. It is worth noting that VSFA [12] is considered as a representative benchmark method and provides a comprehensive experimental analysis on various datasets. According to various experiments, TiVQA shows a significant improvement over VSFA.

4. CONCLUSION

In this paper, we have presented a novel network termed TiVQA for the NR-VQA problem. TiVQA makes full use of the characteristics of HVS. A major novelty is that the texture masking is uniquely leveraged to model the HVS. To detect the texture information of video frame, we utilize the LBP operator and find its optimal setup. Moreover, to adapt different video datasets, TiVQA achieves optimal parameters for the training according to the experiment. Extensive experimental results show that the texture information can greatly boost the performance of VQA methods. Our work sheds some light on the modeling of HVS. In the future, we will investigate more characteristics of HVS for the NR-VQA problem.

5. REFERENCES

[1] K. Naser, V. Ricordel, and P. Le Callet, "Modeling the perceptual distortion of dynamic textures and its application in HEVC," in *IEEE Int. Conf. Image Process. (ICIP)*, pp. 3787-3791, 2016.

[2] Z. Tu, C.-J. Chen, L.-H. Chen, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "Regression or classification? New methods to evaluate no-reference picture and video quality models," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 2085-2089, 2021.

[3] W. Wu, Q. Li, Z. Chen, and S. Liu, "Semantic information oriented no-reference video quality assessment," *IEEE Signal Process. Letters*, vol. 28, pp. 204-208, 2021.

[4] S. Mitra, R Soundararajan, and S. S. Channappayya, "Predicting spatio-temporal entropic differences for robust no reference video quality assessment," *IEEE Signal Process. Letters*, vol. 28, pp. 170-174, 2021.

[5] W. Zhang, W. Zou, F. Yang, L. Lévêque, and H. Liu, "The effect of spatio-temporal inconsistency on the subjective quality evaluation of omnidirectional videos," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 4055-4059, 2019.

[6] C. Galkandage, J. Calic, S. Dogan, and J.-Y. Guillemaut, "Full-reference stereoscopic video quality assessment using a motion sensitive HVS model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 2, pp. 452-466, 2021.

[7] L. Ma, K. N. Ngan, and L. Xu, "Reduced reference video quality assessment based on spatial HVS mutual masking and temporal motion estimation," in *IEEE Int. Conf. on Multimedia and Expo.*, pp. 1-6, 2013.

[8] I. P. Gunawan and M. Ghanbari, "Reduced-reference video quality assessment using discriminative local harmonic strength with motion consideration," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 1, pp. 71-83, 2008.

[9] K. Seshadrinathan and A. C. Bovik, "Temporal hysteresis model of time varying subjective video quality," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 1153-1156, 2011.

[10] T. Ojala, M. Pietikainen, and D. Harwood, "Performance evaluation of texture measures with classification based on Kullback discrimination of distributions," in *IEEE Int. Conf. on Pattern Recognit. (ICPR)*, pp. 582-585, 1994.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 770-778, 2016.

[12] D. Li, T. Jiang, and M. Jiang, "Quality assessment of in-the wild videos," in *ACM Multimedia Conf. (MM)*, pp. 2351-2359, 2019.

[13] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:*1406.1078, 2014.

[14] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695-4708, 2012.

[15] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind prediction of natural video quality," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1352-1365, 2014.

[16] J. Korhonen, "Two-level approach for no-reference consumer video quality assessment," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5923-5938, 2019.

[17] Z. Ying, M. Mandal, D. Ghadiyaram, and A.C. Bovik, "Patch-VQ: 'Patching up' the video quality problem," in *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR),* pp. 14017-14029, 2021.

[18] V. Hosu, F. Hahn, M. Jenadeleh, H. Lin, H. Men, T. Szirányi, S. Li, and D. Saupe, "The konstanz natural video database (KoNViD-1k)," in *Int. Workshop Qual. Multimedia Exper. (QoMEX)*, pp. 1-6, 2017.

[19] M. Nuutinen, T. Virtanen, M. Vaahteranoksa, T. Vuori, P. Oittinen, and J. Häkkinen, "CVD2014—a database for evaluating no-reference video quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3073-3086, 2016.

[20] D. Ghadiyaram, J. Pan, A. C. Bovik, A. K. Moorthy, P. Panda, and K.-C. Yang, "In-capture mobile video distortions: A study of subjective behavior and objective algorithms," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2061-2077, 2018.

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