

The following publication Ngai-Wing Kwong, Sik-Ho Tsang, Yui-Lam Chan, Daniel Pak-Kong Lun, and Tsz-Kwan Lee "No-reference video quality assessment metric using spatiotemporal features through LSTM", Proc. SPIE 11766, International Workshop on Advanced Imaging Technology (IWAIT) 2021, 1176629 (13 March 2021) is available at <https://dx.doi.org/10.1117/12.2590406>

No-Reference Video Quality Assessment Metric Using Spatiotemporal Features Through LSTM

Ngai-Wing Kwong^a, Sik-Ho Tsang^b, Yui-Lam Chan^a, Daniel Pak-Kong Lun^{a,b}, and Tsz-Kwan Lee^c

^a Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong SAR;

^b Centre for Advances in Reliability and Safety Limited (CAiRS), Hong Kong Science Park, New Territories, Hong Kong SAR;

^c School of Information Technology, Deakin University, Australia

ABSTRACT

Nowadays, a precise video quality assessment (VQA) model is essential to maintain the quality of service (QoS). However, most existing VQA metrics are designed for specific purposes and ignore the spatiotemporal features of nature video. This paper proposes a novel general-purpose no-reference (NR) VQA metric adopting Long Short-Term Memory (LSTM) modules with the masking layer and pre-padding strategy, namely VQA-LSTM, to solve the above issues. First, we divide the distorted video into frames and extract some significant but also universal spatial and temporal features that could effectively reflect the quality of frames. Second, the data preprocessing stage and pre-padding strategy are used to process data to ease the training for our VQA-LSTM. Finally, a three-layer LSTM model incorporated with masking layer is designed to learn the sequence of spatial features as spatiotemporal features and learn the sequence of temporal features as the gradient of temporal features to evaluate the quality of videos. Two widely used VQA database, MCL-V and LIVE, are tested to prove the robustness of our VQA-LSTM, and the experimental results show that our VQA-LSTM has a better correlation with human perception than some state-of-the-art approaches.

Keywords: video quality assessment, no reference, long short-term memory, spatiotemporal, pre-padding, masking layer

1. INTRODUCTION

Recently, media platforms and social networks have dramatic growth in video content. However, the video is inevitably distorted since it has been processed, compressed, and transmitted before it finally reaches the end-users. These distortions may affect the human visual experience (HVE). Therefore, a precise VQA metric is highly demanded to provide a satisfactory end-user experience and maintain the QoS. The objective VQA method becomes an attractive and challenging topic in recent years since it could allow automatic quality estimation without any labor force and more suitable for real-time application than Subjective VQA. According to the availability of information on the reference video¹, there are three types of objective VQA methods. They are full-reference (FR)²⁻⁵, reduced-reference (RR)⁶, and no-reference (NR)⁷⁻¹³ VQA methods. Comparing with the FR/RR-VQA method, the NR-VQA method does not require any information from the reference video to assess the distorted video. Since the original video is not always available, the NR-VQA method is a more natural and preferable way to evaluate the perceived video quality in real applications.

Some existing NR-VQA approaches have shown promising results. Wang⁷ proposed a VQA metric that focused on the influence of blockiness and blur artifact to predict the video quality with weighing and linear regression strategy. Zhu⁸ proposed an NR-VQA model for measuring the distortion of compressed video by using intra-subband features and inter-subband features. DeepBVQA⁹ targets on low-resolution video dataset that uses Convolutional Neural Network (CNN) to extract the spatial cues and extract sharpness variation as temporal features to evaluate the video quality. Zhang¹⁰ also trained a CNN by the deformations of 3D discrete cosine transform of video blocks to extract the significant features of the distorted videos and predict the perceptual quality by mapping with frequency histogram. Although the aforementioned NR-VQA methods have explored the spatial and temporal features of distorted videos with different strategies, those NR-VQA methods may ignore the spatiotemporal features of videos, which affect the performance of NR-VQA metrics.

With the development of deep learning, it is potent for learning data representation, and it can also automatically learn abstract features. LSTM is one of the deep learning methods that could process the whole sequences of data and is suitable for making predictions based on time series data. However, the LSTM model is rarely used in VQA metrics for two reasons.

First, the general LSTM model is not suitable for processing videos of various lengths since it requires fixed-length input. Second, various length videos as inputs affect the performance of LSTM. Therefore, we propose the VQA-LSTM to solve the above problems. First, for a general use purpose, we extract some significant but also universal spatial and temporal features that could adequately reflect the quality of frames. Second, the data preprocessing stage is used to normalize the data. Also, due to the fixed-length data input requirement of the LSTM model, the pre-padding strategy is used to handle the variable-length input to be a fixed-length input to ease the training and improve the performance. Finally, a three-layer LSTM model is designed to learn the sequence of spatial features as spatiotemporal features and the sequence of temporal features as the gradient of temporal features to evaluate the quality of videos comprehensively. Also, a masking layer is incorporated with the LSTM model to reduce the impact of variable length input to improve the performance. Therefore, the significant contributions of this paper are threefold: 1) A general LSTM model is proposed to learn the spatiotemporal features and the gradient of temporal features of videos to gauge their quality comprehensively. 2) The pre-padding strategy and masking layer are used to ease the training and improve the performance when adopting the LSTM model in VQA metrics. 3) Our proposed general-purpose VQA-LSTM could be universal for various distortions and have a strong correlation with human perception.

The rest of this paper is organized as follows. In Section 2, the proposed framework and the details are described. In Section 3, the experimental result is presented. Finally, the conclusion and future works are stated in Section 4.

2. PROPOSED MODEL AND METHODOLOGY

The overview of our VQA-LSTM structure is shown in Fig. 1(a). First, the video is divided into frames, and their universal spatial and temporal features are extracted frame by frame to generate feature vectors. In the data preprocessing stage, all feature vectors are normalized to the same scale. With the pre-padding strategy, variable-length feature vectors are promoted to a fixed-length sequence data input for the LSTM model to ease the training. Also, we incorporate the masking layer with the LSTM model to reduce the influence of the pre-padding strategy. Finally, the LSTM model could learn the temporal change of spatial features as spatiotemporal features and the variation of temporal features as the gradient of temporal features to assess the quality of videos.

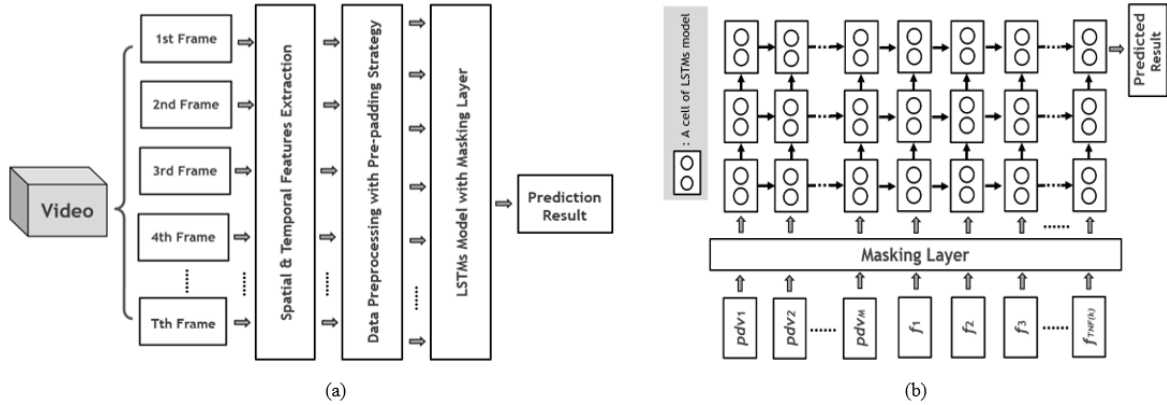


Figure 1. (a) The framework of our proposed VQA-LSTM to predict the quality of the input video. (b) Network Architecture of LSTM incorporated with masking layer

2.1 Spatial Feature Extraction

Sharpness, blocking, and Gaussian noise artifacts are the most common and universal spatial features found in frames¹. Extracting and quantifying those spatial features could adequately reflect the quality of frames. Therefore, to develop a general use purpose NR-VQA model, we extract all the above significant spatial features to evaluate the quality of frames.

For the *Gaussian noise*, in this paper, the block-based noise estimation approach with zero-mean operator¹⁴ is employed to estimate the standard deviation of noise to represent the Gaussian noise of frames. First, to compute the variance of Gaussian noise, we define a 3×3 zero-mean noise estimation operator L as below:

$$L = \begin{bmatrix} 1 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 1 \end{bmatrix} \quad (1)$$

Thus, we could compute the variance of Gaussian noise by applying L to the noisy frame I and average over the frame with width W and height H . Therefore, the standard deviation of noise δ_n could be computed as below:

$$\delta_n = \sqrt{\frac{\pi}{2}} \frac{1}{6(W-2)(H-2)} \sum_{x=1}^{W-2} \sum_{y=1}^{H-2} |I(x, y) * L| \quad (2)$$

For detecting the *blocking artifact*, the block-based blocking artifact detection approach¹⁵ is used. First, we divide the frame into non-overlapping 8×8 blocks and evaluate each region block's blocking artifact regarding its corresponding four neighboring regions. Since the blocking artifact represents incoherence or discontinuities between blocks, the gradient slope across block boundaries could be used to detect blocking artifact. Assume $d_{i,j}$ is the boundary slope of the block (i, j) , and $\overline{d'_{i,j}}$ represents the average slopes next to the boundary. For the two neighboring blocks in the horizontal direction, $d_{i,j}$ and $\overline{d'_{i,j}}$ are defined as below:

$$d_{i,j}(a) = x_{i,j}(a,0) - x_{i,j-1}(a, N-1) \quad (3)$$

$$d'_{i,j}(a, b) = x_{i,j}(a, b) - x_{i,j}(a, b-1) \quad (4)$$

$$\overline{d'_{i,j}}(a) = [d'_{i,j-1}(a, N-1) + d'_{i,j}(a, 1)]/2 \quad (5)$$

where $a = 0, 1, \dots, 7, N = 8$ and $x_{i,j}$ is the pixel value in block (i, j) . Then, the formula of the Mean Absolute Difference of Slope (MADS) is defined as follows:

$$MADS_{i,j} = \frac{1}{N} \sum_{a=0}^{N-1} |d_{i,j}(a) - \overline{d'_{i,j}}(a)| \quad (6)$$

Avoiding the misdetection of real object edges, a threshold of $MADS$, four, is used in this paper. If the $MADS$ of edges is larger than the threshold, it is identified as the blocking edge. After that, we take the average of $MADS$, \overline{MADS} , for all blocking edges to represent the blocking artifact of the frame.

To estimate the *sharpness*, we measure the pixel gradient to identify the sharp edge. The pixel gradient is the difference of the pixel value in both horizontal and vertical directions, represented by Δx and Δy , respectively. We calculate the squared difference along in both horizontal and vertical directions row by row and column by column. If the squared difference of two adjacent pixels is significant, it could be identified as a sharp edge. Therefore, the mean square root value of Δx^2 and Δy^2 could represent the sharpness of frames. The equation is shown as follows:

$$Sharpness = \frac{1}{(H-1)(W-1)} \sqrt{\Delta x^2 + \Delta y^2} \quad (7)$$

2.2 Temporal Feature Extraction

For the temporal features, we measure the similarity percentage of inter-frame (SPIF) to indicate three different levels of Frame Freeze (FF) artifacts to reflect the temporal features of frames. First, we compute the pixel difference PD of each location (x, y) of t^{th} frame and $t-1^{\text{th}}$ frame by the following equation:

$$PD(x, y) = \begin{cases} 1 & F(x, y) - F'(x, y) = 0 \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

where $x = 0, 1, \dots, W-1, y = 0, 1, \dots, H-1, F(x, y)$ is the pixel information of t^{th} frame, and $F'(x, y)$ is the pixel information of $t-1^{\text{th}}$ frame. Thus, PD values in (8) can be used to estimate the $SPIF(t)$ of t^{th} frame as below:

$$SPIF(t) = \frac{1}{W \times H} \sum_{x=0}^{W-1} \sum_{y=0}^{H-1} PD(x, y) \quad (9)$$

For the first frame of video, its $SPIF(t)$ value is set to 0. Based on the $SPIF(t)$, this paper proposes three types of frame freeze artifacts which are analyzed and classified as follows:

$$AFF(t) = \begin{cases} 1 & SPIF(t) = 1 \\ 0 & \text{otherwise} \end{cases} \quad VFF(t) = \begin{cases} 1 & SPIF(t) \geq 0.9 \\ 0 & \text{otherwise} \end{cases} \quad CFF(t) = \begin{cases} 1 & SPIF(t) \geq 0.75 \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

Absolutely Frame Freeze (AFF) is the typical type of the FF that the $SPIF(t)$ is equal to one, which means t^{th} frame is the same as $t-1^{\text{th}}$ frame. *Visually Frame Freeze (VFF)* is a virtual frame freeze derived by HVE. A high $SPIF$ value implies

that the two adjacent frames are similar, and the difference is tiny. Due to the resolving power of human eyes, the video observer could misidentify it as FF. Besides, *Content-related Frame Freeze (CFF)* is another virtual frame freeze caused by continuous high-density frames with similar content in inter-frames. Videos may contain at least 20 to 50 frames within a second. Due to the resolving power of human eyes, if the continuous frames contain high-density and similar content, the observer is hard to identify the difference of frames and misidentifies it as FF.

2.3 Data Preprocessing & Pre-padding Strategy

After the feature extraction stage, $TNF(k)$ features vectors are generated for the k^{th} distorted video, where $k = 1, 2, 3, \dots, K$, and K is the total number of videos in the dataset, and $TNF(k)$ is the total number of frames of k^{th} distorted video. Each feature vector, \mathbf{f}_v , contains six feature values, including δ_n , \overline{MADS} , *Sharpness*, *AFF*, *VFF*, and *CFF*. However, since each feature has its own data range, which is harmful to training, normalization is performed to re-scale the data on a common scale without distorting the differences in data or loss information. Specifically, after normalization, all features are re-scaled into the range [0,1] that benefit the LSTM model to optimize the gradient descent while learning the data representation and improving the performance.

Besides, since the length of the TNF values are different for each video sequence, and the LSTM model requires fixed-length input, the pre-padding method is used to handle the various length input to be a fixed-length input for the LSTM model. Specifically, first, we set the fixed-length $FM = \max_{k=0}^{k=K} \{TNF(k)\}$. Then, the padding data \mathbf{pdv}_M are padded in front of the feature vectors to increase the length of feature vectors to be FM -length feature vectors. Therefore, the input sequence of k^{th} video for the LSTM model is defined as below:

$$seq(k) = \begin{cases} [\mathbf{pdv}_1, \mathbf{pdv}_2, \dots, \mathbf{pdv}_M, f_1, f_2, \dots, f_{TNF(k)}] & TNF(k) < FM \\ [f_1, f_2, f_3, f_4, f_5, \dots, f_{TNF(k)-1}, f_{TNF(k)}] & TNF(k) = FM \end{cases} \quad (11)$$

where $\mathbf{pdv}_M = [-1, -1, -1, -1, -1, -1]$, and $M = FM - TNF(k)$. Those padding data, \mathbf{pdv} , are meaningless and do not affect the original data. Therefore, no matter how long are the distorted videos which make the extracted feature vectors, $TNF(k)$, with different lengths, after the data preprocessing stage, a set of fixed FM -length feature vectors, $seq(k)$, is generated and fed into the LSTM model. It could provide an equal-length sequences data to the LSTM model to ease the training. The padding in (11) is the pre-padding method. We also test the performance¹⁶ of our VQA-LSTM when the post-padding method in (12) is used.

$$Pseq(k) = \begin{cases} [f_1, f_2, \dots, f_{TNF(k)}, \mathbf{pdv}_1, \mathbf{pdv}_2, \dots, \mathbf{pdv}_M] & TNF(k) < FM \\ [f_1, f_2, f_3, f_4, f_5, \dots, f_{TNF(k)-1}, f_{TNF(k)}] & TNF(k) = FM \end{cases} \quad (12)$$

Since the LSTM model has memory to process the time series data by its forget, input, and output gates, if \mathbf{pdv} is placed in front of \mathbf{f}_v , the forget gate could easily reduce the influence of \mathbf{pdv} . Also, with \mathbf{f}_v placed at the back, it can benefit the gradient descent to achieve better performance for our VQA-LSTM. The experimental results of the pre-padding and post-padding methods are shown in section 3.

2.4 Network Architecture & Masking Layer

LSTM is one of the deep learning methods that could process the whole sequences of data and make predictions based on time series data. Therefore, we make good use of the LSTM model to learn the sequence of spatial features as spatiotemporal features, and the gradient of temporal features to evaluate the quality of videos.

In this paper, we designed a many-input one-output LSTM model incorporated with a masking layer, as shown in Fig. 1(b). There are three LSTM layers, and each cell of LSTM has 60 neurons. After the data preprocessing stage, a set of fixed FM -length feature vectors is fed into the LSTM model and processed sequentially. For those spatial features, δ_n , \overline{MADS} , and *Sharpness*, the LSTM model could be used to learn the temporal variation of spatial features along with time series that could represent the spatiotemporal features of the video. For the temporal features, *AFF*, *VFF*, and *CFF* indicate the variation level of content between inter-frame. Thus, the LSTM model is used to process the entire sequences of temporal data to disclose the gradient of temporal features, which could better reflect the whole quality of video.

Although the pre-padding method, Eq. (11), takes care of the various length input issue for the sake of proper training, it also slightly affects the LSTM model's performance. To further resolve this problem, we incorporate a masking layer between the input data and LSTM layers. The masking layer could mask the padding data, \mathbf{pdv} , in the input sequence and skip its timestep. Thus, the LSTM model could only focus on meaningful data, \mathbf{f}_v , to improve performance. The experimental results of the LSTM model with and without the masking layer are shown in section 3.

3. EXPERIMENTAL RESULTS

3.1 Data Sets and Evaluation

1) *MCL-V Video Quality Database*: MCL-V¹⁷ is a high-definition VQA database that contains 96 distorted videos. These distorted videos are generated from 12 reference videos with four different distortion levels and two H.264/AVC compression distortions with and without downsampled. The video sequences contain diversified content that can better reflect the HVE. Each distorted video is a 6-second video with a frame rate of either 20 fps, 22 fps, 25 fps, or 30 fps containing 120-180 various frame lengths. Besides, the MOS of each distorted video is provided to be a ground truth of HVE for testing the VQA-LSTM.

2) *LIVE Video Quality Database*^e: LIVE¹⁸ contains 10 reference videos and 150 distorted videos. 15 distorted videos from each reference video are generated with four different distortion types: wireless distortions, IP distortions, H.264 compression, and MPEG-2 compression. Each distorted video is a 10-second video, except for one group of distorted videos, with a frame rate of 25 fps or 50 fps containing 217-500 various frame lengths. Besides, the ground truth, DMOS scores, are provided for all distorted videos and tested with the VQA-LSTM.

When conducting experiments, each database was divided into two non-overlapping data set. 80% of the distorted videos were used for training, and the remaining 20% were used for testing, out of which 20% of the training set was used for validation. When training the LSTM model, 10,000 epochs are used for training with Mean Square Error loss function, Adam optimizer, and an initial learning rate of 0.0001. We also used full batch learning, in which the gradient calculation can better represent the sample population. Moreover, in our model, *FM* is set to 180 for MCL-V and 500 for LIVE since those are the largest frame number of videos in each database. To evaluate the correlation between MOS/DMOS scores and our proposed model, we used the Pearson Linear Correlation Coefficient (PLCC), and the Spearman Rank Order Correlation Coefficient (SROCC).

3.2 Performance Evaluation on MCL-V and LIVE Video Database

The ablation experiment is performed on MCL-V, which is shown in Table 1. It clearly shows that the pre-padding strategy and the masking layer could improve the performance and solve the restriction of LSTM when adopted in VQA metrics with various length videos. Besides, in Table 2(a), we compared the performances of VQA-LSTM against those of the other NR-VQA and FR-VQA models. The proposed method surpasses other NR-VQA metrics in terms of PLCC and SROCC. As compared to the FR-VQA, our NR-VQA metrics also outperforms some universal FR-VQA metrics such as SSIM¹⁹, VIF²⁰, and STMAD². Although our performance is on par with Zhang's FR-VQA metric using CNN⁵, our VQA-LSTM has a satisfying result and is more suitable to implement in the real application without requiring the reference video. Similarly, some NR/FR-VQA metrics are compared with the VQA-LSTM on the LIVE video database, as shown in Table 2(b). Although DeepBVQA⁹ and Zhang's¹² also use deep learning and CNN algorithms, the proposed method, VQA-LSTM, outperforms them in terms of PLCC and SROCC. As compared to other FR-VQA, our VQA-LSTM also surpasses other FR-VQA metrics, including Zhang's⁵. Since the experimental results show that our VQA-LSTM has a better correlation with human perception than other state-of-the-art approaches, it proves that our VQA-LSTM is robust and effective.

Table 1. Ablation performance of VQA-LSTM variants on MCL-V Video Database.

Method	PLCC	SROCC
Post-padding	0.768	0.782
Pre-padding	0.868	0.856
Post-padding with masking	0.881	0.896
Pre-padding with masking (VQA-LSTM)	0.913	0.953

Table 2. SROCC and PLCC on the (a) MCL-V and (b) LIVE Video Database comparing with other NR/FR-VQA metrics.

MCL-V Video Database						LIVE Video Database					
NR Method	PLCC	SROCC	FR Method	PLCC	SROCC	NR Method	PLCC	SROCC	FR Method	PLCC	SROCC
VIIDEO ¹¹	0.711	0.664	SSIM ¹⁹	0.650	0.648	VIIDEO ¹¹	0.692	0.674	SSIM ¹⁹	0.699	0.718
V-BLIINDS ¹³	0.861	0.746	VIF ²⁰	0.660	0.655	V-BLIINDS ¹³	0.843	0.827	VIF ²⁰	0.759	0.765
Wang's ⁷	0.759	0.806	STMAD ²	0.634	0.623	DeepBVQA ⁹	0.857	0.851	STMAD ²	0.845	0.868
Zhu's ⁸	0.778	0.784	VADM ³	0.742	0.752	Zhang's ¹²	0.863	0.887	Suen's ⁴	0.836	0.859
VQA-LSTM	0.913	0.953	Zhang's ⁵	0.931	0.933	VQA-LSTM	0.893	0.899	Zhang's ⁵	0.875	0.891

(a)

(b)

4. CONCLUSION

In this paper, we developed a general-purpose NR-VQA metric VQA-LSTM. On the one hand, the pre-padding strategy and masking layer are used to ease the training and improve the performance when adopting the LSTM model in VQA metrics. On the other hand, a general LSTM model is built to learn the sequence of universal spatial and temporal features to evaluate the quality of videos. Also, the experimental results demonstrate that the VQA-LSTM is effective and correlation-well with human perception.

5. ACKNOWLEDGEMENTS

This work was supported by...

REFERENCES

- [1] H. C. Soong, and P. Y. Lau, "Video Quality Assessment: A Review of Full-Referenced, Reduced Referenced and No-Referenced Methods," Proc. IEEE Int. Collo. Signal Process. Applicat. (CSPA), pp. 232-237, Mar. 2017.
- [2] P. V. Vu, C. T. Vu, and D. M. Chandler, "A Spatiotemporal Most-Apparent-Distortion Model for Video Quality Assessment," Proc. IEEE Int. Conf. Image Process. (ICIP), pp. 2505-2508, Sep. 2011.
- [3] S. Li, L. Ma, and K. N. Ngan, "Full-Reference Video Quality Assessment by Decoupling Detail Losses and Additive Impairments," IEEE Trans. Circuits Syst. Video Technol., vol. 22, no. 7, pp. 1100-1112, Mar. 2012.
- [4] W.-J. Suen, H.-H. Liu, S.-C. Pei, K.-H. Liu, and T.-J. Liu, "Spatial-temporal visual attention model for video quality assessment," in Proc. IEEE Int. Symp. Circuits Syst. (ISCAS), May 2019, pp. 1-5.
- [5] Yu Zhang, Xinbo Gao, Lihuo He, Wen Lu, Ran He, "Objective Video Quality Assessment Combining Transfer Learning With CNN", IEEE Trans. Neural Netw. Learn. Syst., vol. 31, no. 8, pp. 2716-2730, 2020.
- [6] R. Soundararajan, and A. C. Bovik, "Video Quality Assessment by Reduced Reference Spatio-Temporal Entropic Differencing," Proc. IEEE Trans. Circuit Syst. Video Technol., vol. 23, no. 4, pp. 684-694, Apr. 2012.
- [7] J. Wang, Z. Wang, F. Wang, R. Tariq, and Z. Fei, "A No-Reference Video Quality Assessment Method for VoIP Applications," Proc. IEEE Int. Conf. Signal Process. (ICSP), pp. 644-648, Chengdu, China, Nov. 2016.
- [8] K. Zhu, K. Hirakawa, V. Asari, and D. Saupe, "A No-Reference Video Quality Assessment Based on Laplacian Pyramids," Proc. IEEE Int. Conf. Image Process. (ICIP), pp. 49-53, Melbourne, Australia, Sep. 2013.
- [9] S. Ahn, and S. Lee, "Deep Blind Video Quality Assessment Based on Temporal Human Perception," Proc. IEEE Int. Conf. Image Process. (ICIP), pp. 619-623, Athens, Greece, Oct. 2018.
- [10] Y. Zhang, X. Gao, L. He, W. Lu, and R. He, "Blind video quality assessment with weakly supervised learning and resampling strategy," IEEE Trans. Circuits Syst. Video Technol., vol. 29, no. 8, pp. 2244-2255, Aug. 2019
- [11] A. Mittal, M. A. Saad, and A. C. Bovik, "A Completely Blind Video Integrity Oracle," IEEE Trans. Image Process., vol. 25, no. 1, pp. 289-300, Jan. 2016.
- [12] K. Zhu, C. Li, V. Asari, and D. Saupe, "No-Reference Video Quality Assessment Based on Artifact Measurement and Statistical Analysis," IEEE Trans. Circuit Syst. Video Technol., vol. 25, no. 4, pp. 533-546, Apr. 2015.
- [13] 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, Mar. 2014.
- [14] J. Immerkaer, "Fast Noise Variance Estimation," J. Comput. Vision Image Understanding, vol. 64, no. 2, pp. 300-302, Sep. 1996.
- [15] S. Minami, and A. Zakhor, "An Optimization Approach for Removing Blocking Effects in Transform Coding," IEEE Trans. Circuit Syst. Video Technol., vol. 5, no. 2, pp. 74-82, Apr. 1995.
- [16] M. Dwarampudi, and N.V.S. Reddy, "Effects of Padding on LSTMs and CNNs," arXiv preprint arXiv:1903.07288, pp. 1-5, Mar. 2019.
- [17] J. Y. Lin, R. Song, C.-H. Wu, T. Liu, H. Wang and C.-C. J. Kuo, "MCL-V: A Streaming Video Quality Assessment Database," J. Visual Commun. Image Representation, vol. 30, pp. 1-9, Jul. 2015.
- [18] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," IEEE Trans. Image Process., vol. 19, no. 6, pp. 1427-1441, Jun. 2010.
- [19] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image Quality Assessment: from Error Bisibility to Structural Similarity," IEEE Trans. Image Process., vol. 13, no. 4, pp. 600-612, Apr. 2004.
- [20] H. Sheikh, and A. Bovik, "Image Information and Visual Quality," IEEE Trans. Image Process., vol. 15, no. 2, pp. 430-444, Feb. 2006.