OPTIMIZED QUALITY FEATURE LEARNING FOR VIDEO QUALITY ASSESSMENT

Ngai-Wing Kwong[1], *Yui-Lam Chan*[1], *Sik-Ho Tsang*[2], *and Daniel Pak-Kong Lun*[1,2]

[1]The Hong Kong Polytechnic University, [2]Centre for Advances in Reliability and Safety Limited

ABSTRACT

Recently, some transfer learning-based methods have been adopted in video quality assessment (VQA) to compensate for the lack of enormous training samples and human annotation labels. But these methods induce a domain gap between source and target domains, resulting in a sub-optimal feature representation that deteriorates the accuracy. This paper proposes the optimized quality feature learning via a multi-channel convolutional neural network (CNN) with the gated recurrent unit (GRU) for no-reference (NR) VQA. First, inspired by self-supervised learning, the multi-channel CNN is pre-trained on the image quality assessment (IQA) domain without using human annotation labels. Then, semi-supervised learning is used to fine-tune CNN and transfer the knowledge from IQA to VQA while considering motion-aware information for better quality feature learning. Finally, all frame quality features are extracted as the input of GRU to obtain video quality. Experimental results demonstrate that our model achieves better performance than state-of-the-art VQA approaches.

***Index Terms*—** Multi-channel convolutional neural network, quality feature learning, no reference video quality assessment, self-supervised learning, semi-supervised learning

1. INTRODUCTION

In recent years, video sharing has grown rapidly on social networks [1]. However, videos will inevitably be distorted after processing and transmission, thereby affecting the human visual experience (HVE) [2]. To provide a better end-user experience, an accurate objective VQA approach is highly required to preserve the quality of service. There are three types of objective VQA methods based on their use of reference video [3]: Full-reference (FR) [4, 5, 6], reduced-reference (RR) [7], and NR [8, 9, 10] methods. Since the reference video is not always available in real VQA applications, the NR-VQA approach is commonly used in practice [11].

Recently, many deep neural network (DNN) models have been proposed that can automatically learn the data representation. However, a video usually contains high spatial resolution and frame rate. It is impractical to directly adopt DNN in the VQA task since it requires high computational power and vast memory size. It is no way to train the DNN model for VQA in an end-to-end manner. To relieve the above issue, most existing deep learning-based NR-VQA models separate the spatial and temporal learning process to avoid high computational power at once. However, in VQA databases, each video only contains one mean opinion score (MOS) as ground truth to represent the overall video quality. A human-annotated label is not available for each frame. To ease the labeling burden of training a DNN from scratch, some pre-trained CNN models, such as ResNet [12], pre-trained on ImageNet [13], are used by state-of-the-art NR-VQA methods, such as VSFA [9] and CNN-TLVQM [10]. These VQA methods learn spatial features from the image classification task to the VQA target domain via transfer learning [14]. However, features learned from the image classification can only result in sub-optimal feature representation due to the domain gap between the source image classification task and target VQA domain.

Motivated by self-supervised learning (SSL), this paper proposes a multi-channel CNN model using non-human annotated supervision signals for image-level spatial feature learning, with a GRU model to take motion-aware information into account to predict the video-level MOS for NR-VQA. First, the multi-channel CNN with a channel attention mechanism is pre-trained on the IQA domain with the distorted images and their corresponding structure-aware maps and saliency maps for learning the image quality feature representation guided by non-human annotated supervision signals, which is inspired by pretext task learning in SSL [15, 16]. This arrangement compensates for the shortage of human-annotated labels on frames. It is well-known that human visual attention is attracted by motion events more than structural details[17], the semi-supervised learning is designed to fine-tune the pre-trained CNN model for domain adaptation. To include the motion-aware information on the frame, the unlabeled distorted frame and its corresponding structure-aware map and motion-aware map are fed into the pre-trained CNN to predict the pseudo label, which is treated as the label of frame quality. Then, data from IQA and data from VQA are used to fine-tune the CNN to transfer the knowledge from IQA to VQA domain. It achieves the optimized frame-level quality feature representation learning while considering motion-aware information on a video frame. Finally, all optimized quality feature representations are extracted as the input of GRU to obtain the final precise predicted video quality. To the best of our knowledge, it is the first VQA approach using both semi-supervised learning and SSL with the non-human annotated supervision signal to learn the optimized quality feature representation in a self-supervising manner, transferring the feature from IQA domain to VQA domain while considering motion-aware information.

The rest of this paper is organized as follows. In Section 2, the details of our proposed model are described. Then, the experimental results and related analysis are presented in Section 3. Finally, Section 4 concludes the paper.

2. PROPOSED METHOD

In this section, we introduce our NR-VQA method that adopts a new multi-channel CNN model with GRU, incorporating motion-aware information. The framework of our proposed model is shown in Fig. 1. We detail each part in the following sections.

2.1. SSL-based Multi-channel CNN Model for VQA

To address the shortage of human-annotated labels on frames, we propose to pre-train the multi-channel CNN model in the IQA database, which is regarded as a kind of SSL-based approach, to

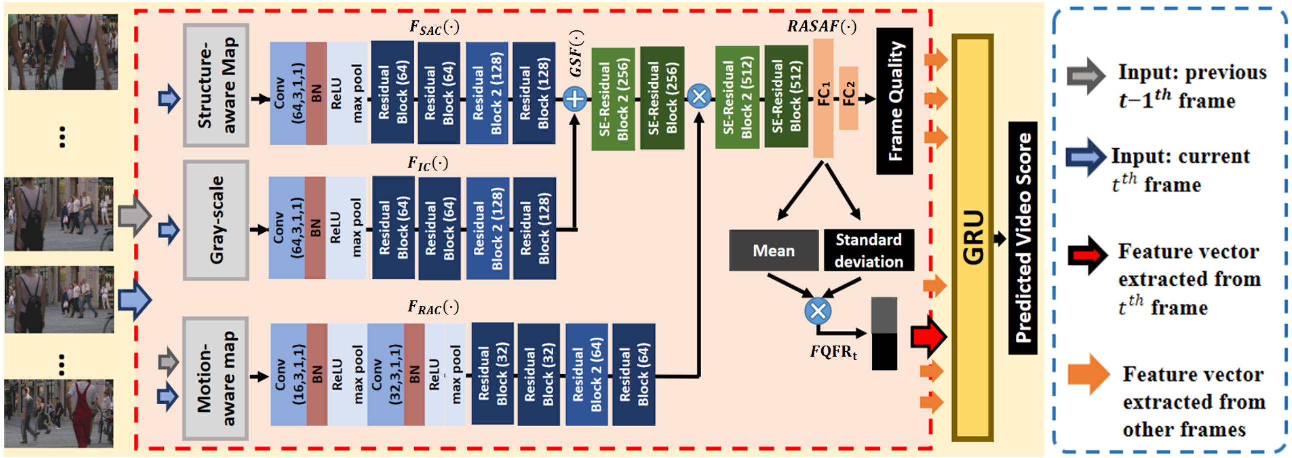

Fig. 1. The framework of our proposed model. $Conv(ch, kn, st, pd)$ represents the 2D convolution operation where ch is the output channel, $kn \times kn$ is the kernel size, st represent the size of stride and pd is the padding size. $ResidualBlock(ch)$ is the structure of residual block with the output channel ch. BN, FC and GP represent the batch normalization operation, fully connected layer and global pooling respectively.

learn the image spatial quality feature representation using distorted images, structure-aware maps, and saliency region maps (SM). In addition, it is known that the HVS is more sensitive to moving objects [17]. Consequently, visual attention should be more attracted by the motion event regions rather than structural details of video. Based on the concept of the region of interest (ROI), we hypothesize that SM can guide the image quality prediction by focusing on vital stillness regions, while the motion-aware region map can guide the frame quality prediction by focusing on motion-aware regions. Therefore, we use the concept of semi-supervised learning on top of the SSL-based multi-channel CNN model and combine data from IQA and data from VQA to fine-tune the CNN model to process the distorted frame, structure-aware maps, and motion-aware region map to estimate the optimized frame-level quality feature representation by considering both spatial and motion-aware information.

2.1.1. *Pre-processing Stage*

Before the training process of the multi-channel CNN model, we first extract the gradient magnitude map (GMM) as the structure-aware map [18]. The GMM of the input distorted images, GMM_d, can be computed by convolving the input distorted images, I^d, with the the Prewitt filters along the horizontal and vertical directions, g_h and g_v, defined in [18], as follows:

$$GMM_d = \sqrt{(I^d * g_h)^2 + (I^d * g_v)^2} \tag{1}$$

Moreover, to extract the SM of the image, the saliency residuals on the spectrum domain are first determined using the method in [19]. Then, we invert saliency residuals from the spectral domain back to the spatial domain to compute the preliminary saliency map (PSM). In addition, we further apply the visual saliency feature (VSF) method in [20] to calculate the center-surround differences on the salient region in the PSM, which can extract fine-grained features and defines borders for the PSM to compute our final SM as follows:

$$SM^d = VSF(PSM(I^d)) \tag{2}$$

Besides, the motion-aware map can be represented by the optical flow map (OFM) since OFM can determine the inter-frame motion variation to represent the motion-aware region. We first convert two adjacent frames into polynomial expansion using the polynomial expansion transformation, $PET(\cdot)$, algorithm in [21]. Since polynomial expansion coefficients can be used to estimate the displacement field, we make good use of this algorithm as the motion estimation method to compute the OFM of two inter-frames as follows:

$$OFM_t = ME(PET(I^d_{t-1}), PET(I^d_t)) \tag{3}$$

where I^d_t and I^d_{t-1} represents the t^{th} and $t-1^{th}$ frame, respectively, and $ME(\cdot)$ is the motion estimation mechanism.

2.1.2. *SSL-based CNN model pre-training*

To train our multi-channel CNN model, motivated by [22, 23] using distortion intensity as the self-supervised signal for regression task in SSL, we use Gradient Magnitude Similarity Deviation (GMSD) [18] as the non-human annotated supervision signal, or so-called pseudo label (PL), PL^d_I=$GMSD$ of unlabeled data from IQA, U^d_I=$\{I^d, GMM^d, SM^d\}$, including the distorted image, and the corresponding GMM and SM, for quality feature learning. By learning GMSD as a pretext task of SSL, our multi-channel CNN model can learn the image quality feature representation. As shown in Fig. 1, in our CNN model, the global spatial features, $GSF(\cdot)$, are extracted by the summation of the features extracted from the channel of GMM^d and the channel of I^d, which is given by:

$$GSF(I^d, GMM^d) = F_{IC}(I^d) \oplus F_{SAC}(GMM^d) \tag{4}$$

where symbol $\oplus$ is the element-wise summation operation, $F_{IC}(\cdot)$ and $F_{SAC}(\cdot)$ represent the feature extraction processes on I^d and GMM^d. Also, we incorporate the squeeze-and-excitation (SENet) block [24] with residual block, which can squeeze the features to be one dimensional data as global information. It can then reinforce the important features and weaken the inconsequence features by the channel-wise multiplication. In the meantime, the region-aware features are extracted from the SM channel and then concatenate with the spatial-aware features, $SAF(\cdot)$=$SEN(GSF(\cdot))$. With the guidance of region-aware features, $SAF(\cdot)$ are weighted by the vital region of the image via the channel attention mechanism. Region-aware and spatial-aware fusion features $RASAF$ is defined as:

$$RASAF(U^d_I) = SEN(SAF(I^d, GMM^d) \otimes F_{RAC}(SM^d)) \tag{5}$$

where symbol $\otimes$ is the concatenation operation, $F_{RAC}(\cdot)$ represents the region-aware features extraction process on SM channel and $SEN(\cdot)$ is the channel attention mechanism. Finally, two fully connected layer, FC_1 and FC_2, are appended to $RASAF$ to predict the non-human annotated supervision signal given by:

$$\hat{p}^L = MultiCNN_{IQA}(U_I^d) = FC_2(FC_1(RASAF(U_I^d))) \quad (6)$$

where $\hat{p}^L$ denote the predicted GMSD of the multi-channel CNN model, $MultiCNN_{IQA}(\cdot)$. The loss function of our $Multi-CNN_{IQA}(\cdot)$ pre-trained in IQA data is defined as:

$$L_{self}(\hat{p}^L, PL_I^d) = \frac{1}{M}\sum_{i=0}^{M-1}(\hat{p}_i^L, PL_{I,i}^d)^2 \quad (7)$$

where M is the batch size, and PL_I^d represent the corresponding supervision signal, GMSD, of the IQA data, U_I^d. With this pretraining, our multi-channel CNN model is able to learn image quality features.

2.1.3. *Semi-supervised learning for fine-tuning*

To transfer the feature representation of our multi-channel CNN models from IQA domain to VQA domain further, we incorporate the motion-aware information into our model to compute the frame-level quality feature representation by using the semi-supervised learning to fine-tune our pre-trained multi-channel CNN model, $MultiCNN_{IQA}(\cdot)$, in Section 2.1.2.

First, SM is used to guide the image quality prediction by focusing on the stillness salient structure region. With the same concept of ROI, we assume that the OFM can be also used to guide the video frame quality prediction by focusing on the motion-aware region at frame as well. Therefore, we replace SM with OFM as the region-aware map for VQA data. As aforementioned, there is no human-annotated label for each video frame. For each video frame of the training VQA databases, $MultiCNN_{IQA}(\cdot)$ is initially used to generate the pseudo labels PL_V^d of data from VQA, U_V^d=$\{I_t^d, GMM_t^d, OFM_t\}$, including the distorted t^{th} frame, and the corresponding GMM and OFM. Assuming $MultiCNN_{trans}(\cdot)$ is the multi-channel CNN model that is being transferred from IQA to VQA, the PL_V^d of data from U_V^d is then generated as:

$$PL_V^d = MultiCNN_{trans}(U_V^d) \quad (8)$$

During the transfer process, the dataset from IQA, $\{U_I^d, PL_I^d\}$, and the dataset from VQA, $\{U_V^d, PL_V^d\}$, are combined as one dataset for semi-supervised learning to re-train the multi-channel CNN model. By doing so, the features learned from the IQA domain can be transferred to our target VQA domain to optimize better frame-level quality feature representation while considering motion-aware information on a video frame. After that, the re-trained multi-channel CNN model is treated as the new multi-channel CNN model to predict the new PL_V^d of data from U_V^d for the next training process, similar to the semi-supervised image classification in [25]. Hence, the loss function of the entire semi-supervised learning including both IQA dataset, $\{U_I^d, PL_I^d\}$, and VQA dataset, $\{U_V^d, PL_V^d\}$, is defined as:

$$L_{semi} = L_{self}(\hat{p}^L, PL_I^d) + a(k)L_{self}(\hat{u}^L, PL_V^d) \quad (9)$$

where

$$a(k) = \begin{cases} 0 & k \leq K_1 \\ [(k-K_1)/(K_2-K_1)]a_f & K_1 < k < K_2 \\ a_f & k \geq K_2 \end{cases} \quad (10)$$

$\hat{u}^L$ denote the predicted result of VQA data, U_V^d , k is the current epoch, and K_1, K_2, and a_f are the parameters for tuning $a(k)$ at different epochs. As we can see, only $L_{self}(\hat{p}^L, PL_I^d)$ is performed when $k \leq K_1$ since the network is at the pre-training stage mentioned in Section 2.1.2. And $a(k)$ is progressively increased by the epoch to include more VQA data $\{U_V^d, PL_V^d\}$ for pre-training. Thus, the model is gradually fine-tuned with more VQA data, U_V^d, so that the domain gap between IQA and target VQA domain is reduced with the consideration of motion-aware information. Finally, after completing the training process at the last epoch, a well-trained multi-channel CNN model, named as $MultiCNN_{VQA}(\cdot)$, is obtained, which can be used for extracting the optimized frame quality feature representation for VQA task. Specifically, the frame-level quality features, $FQFR$, are extracted at the output of the first fully connected layer, FC_1, in $MultiCNN_{VQA}(\cdot)$, as shown in Fig. 1. In practice, we divide the frame into B non-overlapping frame-blocks and each frame block goes through $MultiCNN_{VQA}(\cdot)$ to obtain $FQFR_b$. At the end, we take the mean and standard deviation of all $FQFR_b$ within the frame which as shown in Fig.1:

$$FQFR_t = \{mean\{FQFR_b\}_{b=1}^{b=B}, sd\{FQFR_b\}_{b=1}^{b=B}\} \quad (11)$$

where $FQFP_b$ is the quality feature representation of a frame-block in t^{th} frame, B is the total number of frame-blocks in t^{th} frame, and $mean(\cdot)$ and $sd(\cdot)$ represent the mean and standard deviation operation, respectively. With our proposed pre-training and fine-tuning strategies, $FQFP_t$ contains 512 dimensions, include both structure-aware features and motion-aware features.

2.2. Video quality prediction via GRU model

A GRU model is a well-known recurrent neural network that has a recurrent nature to process input sequences in an iterative way. This makes GRU extract the temporal feature of data efficiently. Consequently, GRU can make predictions based on time series data and can explore the spatiotemporal regularities of distorted videos for our VQA task. Therefore, we take advantage of the GRU model for VQA to learn the temporal variation of frame-level quality feature representation along with time series to represent the spatiotemporal features of the video. The GRU model can reveal the gradient of temporal features by analyzing the entire temporal data sequences, which can comprehensively reflect the whole video quality.

First, for the c^{th} distorted video, a features vector, FV_c=$\{FQFR_1, FQFR_2, ..., FQFR_{T_c-1}, FQFR_{T_c}\}$ is then generated, where c = 1,2, 3, ..., C, C is the total number of videos in the VQA database, $FQFR_t$ is the feature vector of t^{th} frame, and T_c is the total number of frames of the c^{th} distorted video. After that, we built a GRU model to process the input data, FV_c, sequentially to explore spatiotemporal features to comprehensively evaluate the quality of the whole video. We also perform the pre-padding and masking strategy on FV_c to improve the performance since the memory function of GRU can reduce the influence of padding data placed in front of the actual data and it benefits the gradient descent and let the GRU model focus more on meaningful data when the meaningful data are placed at the back.

3. EXPERIMENTAL RESULTS

3.1. Database and implementation details

To demonstrate the validity and the robustness of our proposed model, three UGC VQA databases [26, 27, 28] were tested on models. KoNViD-1k [26] is an extensive database that contains 1200 real-world video sequences with frame rates of 24, 25, and 30 fps. The large number of video sequences in KoNViD-1k represents a wide variety of content and covers almost all kinds of distortions. LIVE-Qualcomm [27] contains 208 distorted videos. These videos are with six common in-capture distortions: artifacts, color, exposure, focus, blurriness, and camera shaking. All videos have a duration of 15 seconds with a frame rate of 30 fps. LIVE-VQC [28] contains 585 distorted videos. All videos have a duration of 10 seconds with frame rates of 19-30 fps (one is 120fps). These videos contain 18 types of resolutions from 240P to 1080P, unique contents, and different combinations of distortions. To evaluate the performance of our proposed model, we used the Pearson Linear Correlation Coefficient (PLCC) and the Spearman Rank Order Correlation Coefficient (SROCC) to measure the accuracy and monotonic consistency between the objective prediction and subjective assessment. Also, a nonlinear regression process is performed to map the prediction result to the subjective scores with different value domains according to the video quality experts group (VQEG) [29].

In our experiments, each video database was divided into five non-overlapping datasets. Four sets of videos were selected for training and validation, while the remaining set was used for testing. Then, five-fold cross-validation was conducted. Also, of videos for training and validation of each cross-validation, 80 % were used for training, and 20 % were used for validating performance. For the multi-channel CNN model training, the CSIQ IQA database [30] was used to pre-train the multi-channel CNN model using (11) since the CSIQ IQA database is a large-scale dataset containing many distorted images with common and diverse distortion, which is suitable as the baseline for image quality feature representation learning. All images and video frames were split into 128 × 128 image/frame blocks. The parameters a_f, K_1, and K_2 in (10) were empirically set as 3, 100, and 700, respectively, through the experiments. We trained the model for 1000 epochs with an initial learning rate of 0.0001 using (9) as the loss function and an Adam optimizer. For the training process of the GRU model, specifically, we built a GRU model with 3 layers and 75 cell units. We set the maximum frame length of the video in each database (Note that the video of 120 fps in LIVE-VQC was not used in our experiment) as the length of FV_c with pre-padding data. Similarly, we used an Adam optimizer to train the model for 500 epochs with an initial learning rate of 0.0001.

3.2. Performance evaluation on three UGC VQA databases

Seven NR-VQA methods, TLVQM [8], VSFA [9], CNN-TLVQM [10], HEKE [31], RAPIQUE [32], VIDEVAL [33], and MDTVSFA [34] were included to evaluate the performance of our proposed model. In particular, VSFA, CNN-TLVQM, RAPIQUE, and MDTVSFA use the CNN model pre-trained on ImageNet classification task to extract the content-aware features via transfer learning. The mean performances of PLCC and SROCC results for the mentioned competitors and the proposed model are given in Table I. As shown in Table I, our proposed model significantly outperforms other NR-VQA methods and achieves the best performance in terms of PLCC and SROCC in these three UGC databases representing that our model is robust and effective. Especially, compared with the second-best performance method in the KoNViD-1k database, the PLCC and SROCC of our proposed method are superior to CNN-TLVQM about 0.018 and 0.012, respectively. Also, as compared with other NR-SCVQA methods, our proposed model improves 0.016 PLCC and 0.005 SROCC in the LIVE-Qualcomm, which confirms the effectiveness of our proposed model.

Table 1. Performance comparison of NR-VQA models on the three UGC databases. The boldfaced entries indicate the best model on each database for each performance metric.

Method	KoNViD-1k		LIVE-Qualcomm		LIVE-VQC	
	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
TLVQM [8]	0.776	0.784	0.802	0.788	0.801	0.803
VSFA [9]	0.757	0.761	0.719	0.726	0.736	0.703
CNN-TLVQM [10]	0.817	0.819	0.813	0.827	0.826	0.817
HEKE [31]	0.739	0.716	0.702	0.718	0.751	0.745
RAPIQUE [32]	0.796	0.804	0.668	0.691	0.743	0.768
VIDEVAL [33]	0.766	0.781	0.698	0.723	0.735	0.729
MDTVSFA [34]	0.784	0.792	0.811	0.807	0.786	0.744
Proposed	**0.835**	**0.831**	**0.829**	**0.832**	**0.827**	**0.823**

Table 2. Generalization performance of NR-VQA models on the three UGC databases.

Method	Directly Average		Weighted Average	
	PLCC	SROCC	PLCC	SROCC
TLVQM [8]	0.793	0.792	0.786	0.790
VSFA [9]	0.737	0.730	0.747	0.740
CNN-TLVQM [10]	0.819	0.821	0.819	0.819
HEKE [31]	0.731	0.726	0.739	0.725
RAPIQUE [32]	0.736	0.754	0.767	0.782
VIDEVAL [33]	0.733	0.744	0.750	0.760
MDTVSFA [34]	0.794	0.781	0.787	0.779
Proposed	**0.830**	**0.829**	**0.832**	**0.829**

Moreover, to explore the generalization ability of models, we following the method in [9] to compute the directly and weighted average performance of models as shown in Table II. From the results in Table II, it is further evident that our proposed model outperforms other NR-VQA methods and exhibits better effectiveness and generalization performance on all three video databases. It can prove that our proposed NR-VQA method is more robust and effective than other transfer learning/pre-trained model-based methods.

4. CONCLUSIONS

This work developed a quality feature learning through multi-channel CNN using non-human annotated labels, and GRU by taking motion-aware information of NR-VQA in considering. First, we overcome limitations of the lack of available human-annotated label data for the VQA task by our SSL-based multi-channel CNN approach. Second, we bridge the domain gap between the IQA and VQA tasks using semi-supervised learning and fine-tuning strategies. Finally, a GRU model is used to explore spatiotemporal features of video to estimate the video quality by the sequence of optimized frame-level quality feature representation. Also, experimental results exhibit the robustness and generalization of our proposed model.

5. ACKNOWLEDGEMENT

The work presented in this article is supported by the Hong Kong Research Grants Council under Research Grant PolyU 152069/18E, and Centre for Advances in Reliability and Safety (CAiRS) admitted under AIR@InnoHK Research Cluster.

6. REFERENCES

[1] Cisco, "Cisco visual networking index: Forecast and trends, 2017 to 2022," in *Whitepaper*, 2017.

[2] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Proc.*, vol. 19, pp. 1427–1441, June 2010.

[3] H. C. Soong and P. Y. Lau, "Video quality assessment: a review of full-referenced, reduced referenced and no-referenced methods," in *Proc. IEEE Int. Collo. Signal Process. Applicat. (CSPA)*. IEEE, 2017, pp. 232–237.

[4] P. V. Vu and D. M. Chandler, "Vis3: An algorithm for video quality assessment via analysis of spatial and spatiotemporal slices," *J. Electron. Image*, vol. 23, pp. 013016, Feb. 2014.

[5] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Process.*, vol. 19, pp. 335–350, Feb. 2010.

[6] J. Wu, Y. Liu, W. Dong, G. Shi, and W. Lin, "Quality assessment for video with degradation along salient trajectories," *IEEE Trans. Multimedia*, vol. 21, pp. 2738–2749, Nov. 2019.

[7] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, pp. 684–694, April 2013.

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

[9] D. Li, T. Jiang, and M. Jiang, "Quality assessment of in-the-wild videos," in *Proc. of the 27th ACM Int. Conf. on Multimedia*, 2019, pp. 2351–2359.

[10] J. Korhonen, Y. Su, and J. You, "Blind natural video quality prediction via statistical temporal features and deep spatial features," in *Proc. of the 28th ACM Int. Conf. on Multimedia*, 2020, pp. 3311–3319.

[11] T.-J. Liu, W. Lin, and C.-C. J. Kuo, "Recent developments and future trends in visual quality assessment," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Submit Conf. (APSIPA ASC)*, 2011, pp. 1–10.

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

[13] J. Deng, W. Dong, R. Socherand L. Li, K. Li, and F.F. Li, "Imagenet: A large-scale hierarchical image database," in *Proc. of the IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2009, p. 248–255.

[14] S. Kornblith, J. Shlens, and Q. V. Le, "Do better imagenet models transfer better?," in *Proc. of the IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 2656–2666.

[15] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *arXiv preprint arXiv:1803.07728*, 2018.

[16] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Proc. European conf. on computer vision (ECCV)*, 2016, pp. 69–84.

[17] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proc. 14th Annu. ACM Int. Conf. Multimedia*, 2006, pp. 815–824.

[18] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Trans. Image Process.*, vol. 23, pp. 684–695, Feb. 2014.

[19] X. Hou and L. Zhang, "Saliency detection: a spectral residual approach," in *Proc. of the IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2007, pp. 1–8.

[20] S. Montabone and A. Soto, "Human detection using a mobile platform and novel features derived from a visual saliency mechanism," in *Image and Vision Computing*, 2010, p. 391–402.

[21] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Scandinavian Conf. on Image analysis*, 2003, pp. 363–370.

[22] K. Sheng, W. Dong, M. Chai, G. Wang, P. Zhou, F. Huang, and C. Ma, "Revisiting image aesthetic assessment via self-supervised feature learning," in *Proc. of the AAAI Conf on Artificial Intelligence*, 2020, pp. 5709–5716.

[23] J. Pfister, K. Kobs, and A. Hotho, "Self-supervised multi-task pretraining improves image aesthetic assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 816–825.

[24] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 2011–2023, April 2019.

[25] D. H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," *Workshop on challenges in representation learning*, vol. 3, 2013.

[26] 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 *Proc. 9th Int. Conf. Quality Multimedia Exper. (QoMEX)*, 2017, pp. 1–6.

[27] D. Ghadiyaram, J. Pan, A. C. Bovik, A. K. Moorthy, P. Panda, and K. C. Yang, "Live-qualcomm mobile in-capture video quality database," 2017.

[28] Z. Sinno and A.C. Bovik, "Large-scale study of perceptual video quality," *IEEE Trans. on Image Process.*, vol. 28, pp. 612–627, Feb. 2019.

[29] Video Quality Experts Group, "Final report from the video quality experts group on the validation of objective quailty metrics for video quality assessment," Available: http://www.its.bldrdoc.gov/vqeg/vqeg-home.aspx.

[30] E.C. Larson and D.M. Chandler, "The csiq image database," Available online at: http://vision.okstate.ed.

[31] Y. Liu, J. Wu, L. Li, W. Dong, J. Zhang, and G. Shi, "Spatiotemporal representation learning for blind video quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, pp. 3500–3513, June. 2022.

[32] Z. Tu, X. Yu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "Rapique: Rapid and accurate video quality prediction of user generated content," *IEEE Open Journal of Signal Process.*, vol. 2, pp. 425–440, 2021.

[33] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "Ugc-vqa: Benchmarking blind video quality assessment for user generated content," *IEEE Trans. on Image Process.*, vol. 30, pp. 4449–4464, 2021.

[34] D. Li, T.Jiang, and M. Jiang, "Unified quality assessment of in-the-wild videos with mixed datasets training," *Int. Journal of Computer Vision*, pp. 1238–1257, 2021.