# Semantic Information Oriented No-Reference Video Quality Assessment

Wei Wu, Qinyao Li, Zhenzhong Chen, *Senior Member, IEEE*, and Shan Liu, *Senior Member, IEEE*

***Abstract*—In this letter, a method called Semantic Information Oriented No-Reference (SIONR) video quality assessment model is developed, which can effectively represent quality degradation of video by taking the variations of semantic information into consideration. Specially, temporal variations of the semantic features between adjacent frames are calculated to consider the inconsistency of the static semantic information. Moreover, low-level features are also applied as a supplementary to take distortions related to local details into consideration. Experimental results demonstrate that our proposed method obtains competitive performance compared with state-of-the-art methods in the two databases. Also, our model achieves good generalization capability. The code is available at: https://github.com/lorenzowu/SIONR.**

***Index Terms*—no-reference video quality assessment, semantic information, temporal variations, low-level features**

## I. Introduction

VIDEO quality assessment (VQA) has been studied as an increasingly important tool used in video processing [1]. VQA methods can be classified into three categories according to the accessibility of reference videos: full-reference VQA (FR-VQA), reduced-reference VQA (RR-VQA), and no-reference (NR-VQA). In reality, the reference video is usually unavailable. Thus, NR-VQA becomes a hot research topic.

For traditional NR-VQA methods, feature descriptors are usually designed based on the properties of HVS or natural scene statistics to conduct quality prediction [2], [3], [4]. However, it's not easy to design proper hand-crafted features to represent quality degradation. Deep learning-based methods have shown great potential in the area of quality assessment due to the powerful feature representation ability of CNN. Early deep learning-based methods [5], [6] are designed based on the traditional synthetic distortion databases [1], [7] which have a small number of unique contents and are degraded by only one or at most two synthetic distortions. To overcome these limitations, in-the-wild videos [8], [9] which contain a mass of content and may suffer from complex mixed real-world distortions are introduced. Unlike quality assessment of synthetically-distorted videos, quality assessment of in-the-wild videos requires to compare cross-content video pairs, which may be more strongly affected by the content. Considering the content-dependency effect on the human visual system (HVS), VSFA [10] utilizes the pre-trained ResNet-50 [11] to extract spatial features of each frame. Then a sequence of GRUs [12] and a temporal pooling are used to learn the temporal-memory effect and predict the quality.

Although recent deep learning-based models have demonstrated promising results on in-the-wild videos, they tend to use static high-level semantic features extracted by pre-trained models. However, they usually focus more on the impairment of the video contents instead of considering the semantic differences caused by quality degradation. In addition, recent deep learning-based methods [10] investigate global abstracts instead of local details of distortions, which play an important role in human's perception of the quality degradation.

In this letter, we propose a novel model named Semantic Information Oriented No-Reference (SIONR) VQA model considering the variations of the semantic contents and the low-level distortions. To represent the distortions reflected by the semantic information, we calculate the temporal variations between adjacent frames based on the features extracted from the pre-trained ResNet-50 [11]. These temporal variations can measure the inconsistency of the semantic meaning. To deal with the low-level distortions, a shallow but effective convolutional neural network is designed to extract low-level features. Moreover, inspired by the fact that the perception of the spatial distortions can be largely influenced by the temporal variations [13], [14], [15], [16], [17], we fuse the spatial features with the temporal variations of the low-level features to determine the contribution of temporal and spatial distortions and their interaction to the overall video quality. Finally, the low-level fused features are combined with the high-level semantic temporal variations to take both semantic variation and local distortions into consideration.

## II. The Proposed Approach

An NR-VQA framework, named SIONR, considering the temporal variations of the semantic information and low-level distortions is proposed in this section, as shown in Fig. 1. First, we extract the high-level features and low-level features separately to take the perception of both the semantic information and low-level distortions into consideration. For the high-level features, the temporal variations are calculated on adjacent frames to deal with the fluctuation of semantic information. For the low-level features, we multiply the temporal variations with the spatial features to compute the mutual effects between them. Later, quality-related features in high-level and low-level channels are fused to predict frame-level quality. Finally, the temporal average pooling is applied to obtain the video-

This work was supported in part by grants from the National Natural Science Foundation of China under Grant 61771348 and Tencent. *(Corresponding author: Zhenzhong Chen.)*

W. Wu, Q. Li, and Z. Chen are with the School of Computer Science, Wuhan University, Wuhan 430072, China (e-mail: zzchen@ieee.org). Q. Li is also with the Hongyi Honor College, Wuhan University, Wuhan 430079, China.

S. Liu is with Tencent, Palo Alto, CA 94306 USA.

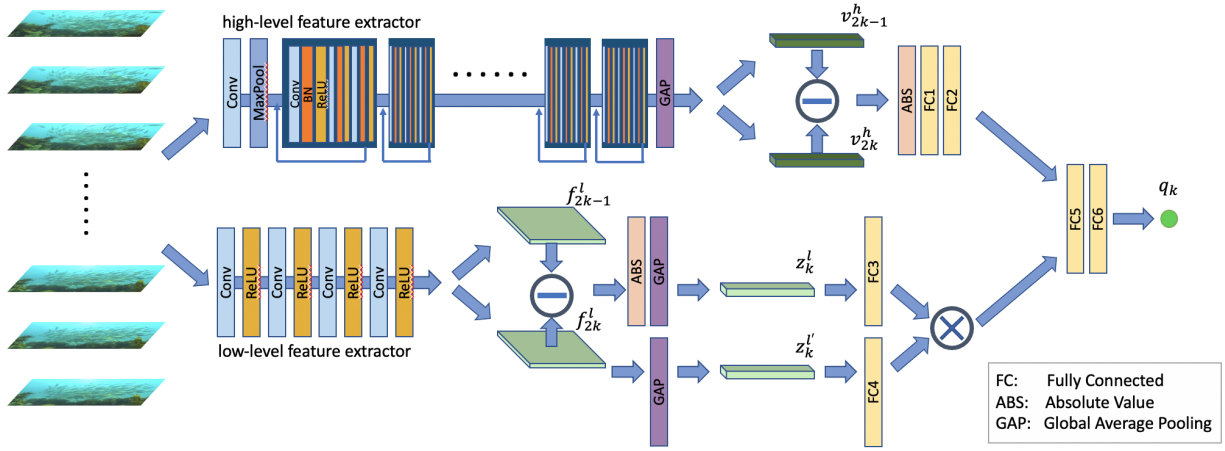


Fig. 1: The framework of our approach. A shallow network is used to extract low-level features while ResNet-50 is used to extract high-level features. Temporal variations are then calculated separately. The low-level temporal features are fused with corresponding spatial features before the combination of different level features. FC layers are applied finally to form a score.

level quality. Details of every module in the framework are described in the following sections.

### A. High-level Features

*1) High-level Feature Extraction:* For in-the-wild videos, semantic content of the videos will have a huge impact on the perceived video quality due to the fact that different video content tends to influence human tolerance thresholds for distortions [10]. So it's rational to extract high-level semantic features, which are content-aware to evaluate the quality degradation of the in-the-wild videos. Therefore, we choose the ResNet-50 [11] pre-trained on ImageNet [18] for the high-level feature extraction. This model possesses the discriminatory power of different content information, thus being able to extract content-aware features.

Firstly, assuming the video contains $T$ frames, we feed the video frame $I_t$ into the pre-trained CNN model and output the high-level semantic feature map $f_t^h$ from its last convolutional layer. $f_t^h$ has a total of $C$ feature maps. Then a global average pooling layer turns the feature maps into a single feature vector (2048 dimensions). The feature vector is used to calculate temporal variations.

*2) Temporal variations of high-level features:* The temporal variations of high-level features refer to the fluctuation of high-level features in time as a measure of the temporal difference in semantic information. The difference between adjacent frames of a good quality video is so small that the semantics of adjacent frames can be regarded as the same to a certain extent. However, severe distortions may lead to a great change of the semantic meaning, which may be detected by the temporal variations of the high-level features. So, it's rational to use the temporal variations of high-level features instead of the high-level features directly to indicate the presence of distortions. What's more, considering the fact that video quality can be influenced by the quality fluctuation in the time domain, a video with better quality tends to have smaller quality variation over time as compared to a video with poorer quality [17]. Based on the above reasons, temporal variations of the semantic information are an important factor to measure the quality degradation. So we first perform Global Average Pooling (GAP) operation and then calculate the temporal variations of high-level features.

$$v_t^h = \text{GAP}\left(f_t^h\right), t \in [1, T] \tag{1}$$

$$z_k^h = \text{ABS}\left(v_{2k}^h - v_{2k-1}^h\right), k \in [1, T//2] \tag{2}$$

where $f_t^h$ is the high-level semantic feature map, $v_t^h$ is the output feature vectors, $z_k^h$ refers to temporal variations in the high level, and ABS represents element-wise absolute value.

Pursuit movement [14], one of the three types of eye movements, refers to the ability of the eyes to smoothly track a moving object. If there exists a severe quality degradation among the temporal axis, the pursuit movement allows the human visual system to detect the changes and then perceive the temporal variations of the distortions. In other words, temporal variations can reflect the distortions of moving objects, thus reflecting the distortions in the temporal domain.

### B. Low-level Features

*1) Low-level Feature Extraction:* Although recent deep learning-based models [19], [10] have demonstrated promising results on quality assessment, they tend to focus largely on extracting high-level semantic features by pre-trained recognition-oriented CNN models [20], [11]. However, image recognition tasks are different from VQA tasks when it comes to their goals and properties. Usually, image recognition tasks pay more attention to visual contents than to distortions while VQA tasks attach great importance to both distortions and visual contents. On the condition of that, the pre-trained recognition-oriented CNN models may overlook the low-level visual features, which concern more on local details and thus play an important role in human's perception of the video quality. So, we design a low-level feature extractor to better deal with local details.

A shallow convolution network is used to extract low-level vision-dependent information as a supplement to facilitate

quality prediction. To do this, the structure of the low-level feature extractor before the feature fusion is Conv1$(3,3,16)$ - Conv2$(3,3,32)$ - Conv3$(3,3,64)$ - Conv4$(3,3,64)$, where Conv$p(s,s,c)$ is the $p$-th 2D convolutional layer with $c$ filters of spatial size $s \times s$. Zero-padding is applied before each convolution to output the desired spatial size of the feature map. Additionally, the output of each convolutional layer is spatially down-sampled by a factor of 2. Supposing that the video contains $T$ frames, the output of the low-level feature map $f_t^l$ will be extracted from the last convolutional layer of the shallow convolutional network when the frame $I_t$ is fed into the network.

*2) Temporal variations of low-level features:* The temporal variations of low-level features refer to the fluctuation of low-level features in time as a measure of the temporal difference in local distortion information. To account for the temporal variations of low-level features, we calculate the temporal variations of low-level spatial features for each location of video frames. Ninassi *et al.* [14] found that the perception towards local feature is related to the residual movement of the eye called Fixation, this eye movement allows HVS to gaze at a particular area of the visual field, which means the spatial distortions are locally evaluated. Different from the high-level feature processing, we first calculate the low-level temporal variations and then apply a GAP to obtain low-level perception in frame-level. This is because low-level vision is position-related and involves local details, while high-level vision is a global abstract. Since we care more about the values of temporal variations instead of the directions, we took the absolute value of the temporal variations of spatial features:

$$z_k^l = \text{GAP}\left(\text{ABS}\left(f_{2k}^l - f_{2k-1}^l\right)\right), k \in [1, T//2] \quad (3)$$

Unlike images, video signals carry information over spatial as well as the temporal domain. Although temporal factors are crucial for VQA, the spatial factors apart from the temporal ones also need to be considered. The detailed method to determine the contributions of the two factors and their interaction will be explained in the next section.

### C. Hierarchical Feature Fusion

*1) Low-level Spatiotemporal Feature Fusion:* Low-level feature extractor is used in our model to extract features which contain distortion information per frame. However, Spatial quality alone may not be sufficient since the temporal factors crucial to VQA are disregarded. Many researchers have pointed out that quality along the temporal axis is also an important factor to be considered [21], [13]. So, it's rational for us to take the contribution of both spatial factors and temporal factors as well as their interaction into consideration. To handle this influence, we first apply an FC layer to the features (64-dimension FC for the low-level temporal features and 1-dimension FC for the low-level spatial features) and then multiply them together to imitate the interaction. Nevertheless, for high-level semantic features, spatial features represent semantic information, while the difference of high-level features represents distortions. Therefore, they cannot be combined directly in the same way as low-level features.

*2) Hierarchical Feature Fusion for Quality Prediction:* Research on neuroscience points out there exists a hierarchical process for visual perception [22], [23]. According to the hierarchical characteristic of visual perception, we propose to combine both the low and high-level features. However, the output dimensions of the high-level and low-level features are not balanced (2048 for high-level features, 64 for low-level features). If we directly concatenate those two features, the expression of low-level information is much likely to be suppressed. As can be seen from Table II, high-level features have a greater contribution to the prediction of video quality. Therefore, we reduce the high-level feature dimension to 128 through fully connected layer FC1 and FC2, which will not only affect the expression of low-level features but also highlight the importance of high-level features. After the concatenation, the dimension of fused features is gradually reduced to 1 through FC5 and FC6, and the output dimension of FC5 is 64. The one dimension feature is the frame-level score $q_k$. Then the temporal average pooling is applied to obtain the video-frame quality.

### D. Implementation Details

We train the whole network in an end-to-end manner. The proposed model is implemented with PyTorch with the initial learning rate set to 0.001. Also, the training batch size is set to 1. We compute the mean square error (MSE) between the predicted score and MOS as the loss and update the model parameters using the Adam [24] optimizer. Additionally, the two feature extraction networks are trained jointly.

## III. Experimental Results and Analysis

### A. Databases

To evaluate the proposed metric, we use four relevant NR-VQA databases: KoNViD-1k [8], LIVE-VQC [9], CVD2014 [25] and LIVE-Qualcomm [26]. KoNViD-1k consists of 1200 public-domain video sequences and LIVE-VQC contains 585 videos. As for the other two datasets, CVD2014 is composed of 234 videos and LIVE-Qualcomm has 208 videos. The CVD2014 and LIVE-Qualcomm datasets are only used in the cross-dataset tests because of the small size.

### B. Performance Evaluation

For a fair comparison, we include a number of representative and state-of-the-art BVQA/BIQA algorithms in our benchmarking evaluation as performance references to be compared against. These baseline models include NIQE [27], BRISQUE [28], CORNIA [29], V-BLIINDS [2], HIGRADE [30], FRIQUEE [4], VSFA [10], TLVQM [31]. Among them, NIQE is "completely blind" (opinion-unaware), since no training is required to build it. The rest of the models are all training-based (opinion-aware) and we retrain the models/features when evaluating on a given dataset.

Pearson Linear Correlation Coefficient (PLCC), Spearman Rank-Order Correlation Coefficient (SROCC) and Root Mean Square Error (RMSE) are the three performance criteria of VQA methods. SROCC indicates the prediction monotonicity,

TABLE I: Performance on the two benchmark datasets

| KoNViD-1k [8] | | | |
|---|---|---|---|
| Model | PLCC (±STD) | SROCC (±STD) | RMSE (±STD) |
| NIQE [27] | 0.5143 (±0.0405) | 0.5139 (±0.0384) | 0.5501 (±0.0255) |
| BRISQUE [28] | 0.6128 (±0.0458) | 0.6445 (±0.0320) | 0.5121 (±0.0323) |
| CORNIA [29] | 0.7135 (±0.0236) | 0.7169 (±0.0245) | 0.4486 (±0.0188) |
| V-BLIINDS [2] | 0.7041 (±0.0356) | 0.7219 (±0.0318) | 0.4578 (±0.0297) |
| HIGRADE [30] | 0.7235 (±0.0325) | 0.7278 (±0.0299) | 0.4319 (±0.0281) |
| FRIQUEE [4] | 0.7325 (±0.0269) | 0.7312 (±0.0271) | 0.4289 (±0.0232) |
| VSFA [10] | 0.7733 (±0.0247) | 0.7727 (±0.0236) | 0.4117 (±0.0205) |
| TLVQM [31] | 0.7546 (±0.0215) | 0.7686 (±0.0221) | 0.4262 (±0.0206) |
| SIONR | 0.8180 (±0.0172) | 0.8109 (±0.0200) | 0.3688 (±0.0160) |
| LIVE-VQC [9] | | | |
| Model | PLCC (±STD) | SROCC (±STD) | RMSE (±STD) |
| NIQE [27] | 0.4636 (±0.0735) | 0.4462 (±0.0794) | 14.8988 (±0.6879) |
| BRISQUE [28] | 0.6436 (±0.0454) | 0.6118 (±0.0494) | 13.1195 (±0.9064) |
| CORNIA [29] | 0.7183 (±0.0420) | 0.6719 (±0.0473) | 11.8329 (±0.7008) |
| V-BLIINDS [2] | 0.7266 (±0.0300) | 0.7055 (±0.0406) | 11.7119 (±0.7538) |
| HIGRADE [30] | 0.6430 (±0.0652) | 0.6274 (±0.0680) | 12.9720 (±0.9045) |
| FRIQUEE [4] | 0.7012 (±0.0546) | 0.6532 (±0.0582) | 12.2458 (±0.9173) |
| VSFA [10] | 0.7512 (±0.0438) | 0.7035 (±0.0454) | 11.3219 (±0.8571) |
| TLVQM [31] | 0.8076 (±0.0316) | 0.8064 (±0.0297) | 10.1127 (±0.7126) |
| SIONR | 0.7821 (±0.0355) | 0.7361 (±0.0446) | 10.4744 (±0.6052) |

while PLCC and RMSE measure the prediction accuracy. Better VQA methods should have larger PLCC/SROCC and smaller RMSE. Note that PLCC and RMSE are computed after performing a nonlinear four-parametric logistic function to map the objective predictions into the same scale of MOS as described in [32].

Following convention, we randomly split the dataset into non-overlapping training, validation and testing sets ($60\%/20\%/20\%$). There is no overlap between these three parts. This procedure is repeated 20 times and the mean and standard deviation of performance values are reported in Table I. For all training-based and our methods, we choose the models with the lowest RMSE values on the validation set during the training phase. On KoNViD-1K, SIONR outperforms other BVQA models by a notable margin. However, on LIVE-VQC, TLVQM shows better performance than SIONR. The reasons may lie in two points. Firstly, LIVE-VQC video dataset (585) is much smaller than KoNViD-1k video dataset (1200), so it may be hard to train deep learning-based methods on this dataset. Besides, LIVE-VQC video dataset contains more (camera) motions than KoNViD-1k video dataset. TLVQM [31] puts motion relevant features into consideration and thus can deal with the motion better. Overall, our model achieves competitive performance compared with other methods.

### C. Ablation Study

*1) High-level and low-level features:* Since our model involves two components, the low-level features and high-level features, we study the impact of each of the components in Fig. 1. We note that high-level features perform better than low-level features. Further, we see that the combination of the high-level and low-level features leads to an improvement in the performance, which means although the high-level features contribute more to the performance, the contribution of low-level features cannot be ignored.

*2) Effect of temporal variations:* In order to verify the effectiveness of the temporal variation operator, we compare the cases with and without this operation. SIONR denotes the features with a temporal variation operator. $\mathrm{SIONR\ w/oTemVar}$ denotes the features are directly fed into fully connected layers. From Table II, we observe that the performance of the temporal variation operator on the two datasets is better that of no operation, which illustrates the effectiveness of the temporal variations.

TABLE II: Effectiveness of low-level features, high-level features and temporal variations on the two benchmark datasets

| KoNViD-1k [8] | | | |
|---|---|---|---|
| Model | PLCC (±STD) | SROCC (±STD) | RMSE (±STD) |
| SIONR w/o TemVar | 0.4361 (±0.1742) | 0.3817 (±0.1807) | 0.5639 (±0.0736) |
| SIONR w/o HighLevel | 0.7290 (±0.0567) | 0.7244 (±0.0538) | 1.4168 (±2.9985) |
| SIONR w/o LowLevel | 0.8112 (±0.0196) | 0.8042 (±0.0209) | 0.3749 (±0.0180) |
| SIONR | 0.8180 (±0.0172) | 0.8109 (±0.0200) | 0.3688 (±0.0160) |
| LIVE-VQC [9] | | | |
| Model | PLCC | SROCC | RMSE |
| SIONR w/o TemVar | 0.4910 (±0.2085) | 0.4111 (±0.2180) | 14.1323 (±2.2719) |
| SIONR w/o HighLevel | 0.6883 (±0.0465) | 0.6376 (±0.0589) | 12.1981 (±0.7096) |
| SIONR w/o LowLevel | 0.7750 (±0.0369) | 0.7218 (±0.0432) | 10.6227 (±0.6243) |
| SIONR | 0.7821 (±0.0355) | 0.7361 (±0.0446) | 10.4744 (±0.6052) |

TABLE III: Cross-dataset comparisons

| Train / Test | Model | PLCC | SROCC | RMSE |
|---|---|---|---|---|
| KoNViD-1k [8] / LIVE-VQC [9] | TLVQM [31] | 0.6962 | 0.6666 | 12.2450 |
| | SIONR | 0.7871 | 0.7554 | 10.5222 |
| KoNViD-1k [8] / CVD2014 [25] | TLVQM [31] | 0.5475 | 0.4937 | 17.9438 |
| | SIONR | 0.7353 | 0.7351 | 14.5326 |
| KoNViD-1k [8] / LIVE-Qualcomm [26] | TLVQM [31] | 0.4997 | 0.4480 | 10.3325 |
| | SIONR | 0.6594 | 0.6528 | 8.9682 |
| LIVE-VQC [9] / KoNViD-1k [8] | TLVQM [31] | 0.6148 | 0.6270 | 0.5054 |
| | SIONR | 0.7511 | 0.7505 | 0.4231 |
| LIVE-VQC [9] / CVD2014 [25] | TLVQM [31] | 0.6426 | 0.6029 | 16.4287 |
| | SIONR | 0.6259 | 0.5926 | 16.7228 |
| LIVE-VQC [9] / LIVE-Qualcomm [26] | TLVQM [31] | 0.6763 | 0.6623 | 8.7868 |
| | SIONR | 0.6260 | 0.6043 | 9.3019 |

### D. Cross-Dataset Generalizability

We also run cross-dataset tests on the KoNViD-1k, LIVE-VQC, CVD2014 and LIVE-Qualcomm datasets to verify the generalizability of the proposed model. We only use KoNViD-1k and LIVE-VQC as training sets due to the reasons that CVD2014 and LIVE-Qualcomm datasets are too small to be trained well. From Table I, we can see that TLVQM is a competitive method, so we use this method as a comparison to verify the generalization of our model. The results are shown in Table III, from which we can find that the proposed model shows good performance. Therefore, we conclude that the proposed model is reasonable and has good generalizability.

## IV. Conclusion

In this letter, we propose a novel SIONR model for NR-VQA. In our model, the temporal variations of high-level features are considered by calculating the temporal difference of the high-level features, so that the quality degradation reflected by the inconsistency of the semantic information can be detected. In addition, low-level features are used as an important supplementary for high-level semantic information to evaluate the local distortion information. The experimental results demonstrate that our model obtains competitive performance compared with state-of-the-art methods in the two databases and achieves good generalization capability. The ablation study confirms the necessity of each module in our framework.

## REFERENCES


[1] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1427–1441, 2010.

[2] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind prediction of natural video quality," *IEEE Transactions on Image Processing*, vol. 23, no. 3, pp. 1352–1365, 2014.

[3] J. Xu, P. Ye, Y. Liu, and D. Doermann, "No-reference video quality assessment via feature learning," in *IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 491–495.

[4] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *Journal of Vision*, vol. 17, no. 1, pp. 32–32, 2017.

[5] Y. Zhang, X. Gao, L. He, W. Lu, and R. He, "Blind video quality assessment with weakly supervised learning and resampling strategy," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 8, pp. 2244–2255, 2018.

[6] Y. Li, L.-M. Po, C.-H. Cheung, X. Xu, L. Feng, F. Yuan, and K.-W. Cheung, "No-reference video quality assessment with 3D shearlet transform and convolutional neural networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 6, pp. 1044–1057, 2015.

[7] P. V. Vu and D. M. Chandler, "ViS3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices," *Journal of Electronic Imaging*, vol. 23, no. 1, pp. 1–25, 2014.

[8] 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 *International Conference on Quality of Multimedia Experience (QoMEX)*, 2017, pp. 1–6.

[9] Z. Sinno and A. C. Bovik, "Large-scale study of perceptual video quality," *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 612–627, 2018.

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

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[12] 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," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734.

[13] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Transactions on Image Processing*, vol. 19, no. 2, pp. 335–350, 2009.

[14] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Considering temporal variations of spatial visual distortions in video quality assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 253–265, 2009.

[15] M. Barkowsky, J. Bialkowski, B. Eskofier, R. Bitto, and A. Kaup, "Temporal trajectory aware video quality measure," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 266–279, 2009.

[16] A. K. Moorthy and A. C. Bovik, "Efficient video quality assessment along temporal trajectories," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 11, pp. 1653–1658, 2010.

[17] M. Narwaria, W. Lin, and A. Liu, "Low-complexity video quality assessment using temporal quality variations," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 525–535, 2012.

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.

[19] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 4041–4056, 2020.

[20] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

[21] Q. Huynh-Thu and M. Ghanbari, "Modelling of spatio–temporal interaction for video quality assessment," *Signal Processing: Image Communication*, vol. 25, no. 7, pp. 535–546, 2010.

[22] S. Hochstein and M. Ahissar, "View from the top: Hierarchies and reverse hierarchies in the visual system," *Neuron*, vol. 36, no. 5, pp. 791–804, 2002.

[23] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature Neuroscience*, vol. 2, no. 11, pp. 1019–1025, 1999.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations (ICLR)*, 2015.

[25] 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 Transactions on Image Processing*, vol. 25, no. 7, pp. 3073–3086, 2016.

[26] 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 Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 9, pp. 2061–2077, 2017.

[27] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2012.

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

[29] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1098–1105.

[30] D. Kundu, D. Ghadiyaram, A. C. Bovik, and B. L. Evans, "No-reference quality assessment of tone-mapped HDR pictures," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2957–2971, 2017.

[31] J. Korhonen, "Two-level approach for no-reference consumer video quality assessment," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 5923–5938, 2019.

[32] VQEG, *Report on the Validation of Video Quality Models for High Definition Video Content, Phase I*, 2010. [Online]. Available: https://www.its.bldrdoc.gov/vqeg/projects/hdtv/hdtv.aspx