Quality assessment of enhanced videos guided by aesthetics and technical quality attributes

Mirko Agarla Luigi Celona Claudio Rota
Raimondo Schettini

Department of Informatics Systems and Communication
University of Milano - Bicocca
Viale Sarca 336, Building U14, Milan, Italy
{first_name.second_name}@unimib.it

Abstract

In this work we propose a novel method to evaluate the quality of enhanced videos. Perceived quality of a video depends on both technical aspects, such as the presence of distortions like noise and blur, and non-technical factors, such as content preference and recommendation. Our approach involves the use of three deep learning based models that encode video sequences in terms of their overall technical quality, quality-related attributes, and aesthetic quality. The resulting feature vectors are adaptively combined and used as input to a Support Vector Regressor to estimate the video quality score. Quantitative results on the recently released VQA Dataset for Perceptual Video Enhancement (VDPVE) introduced for the NTIRE 2023 Quality Assessment of Video Enhancement Challenge demonstrates the effectiveness of the proposed method.

1. Introduction

Evaluating the perceived quality of in-the-wild videos is a critical task in the fields of video capture, transmission, compression, reproduction, and processing. Video Quality Assessment (VQA) methods are tools that use computational models to mimic the human perception about video quality. Conventionally, VQA studies have focused on measuring the technical quality of User-Generated Content (UGC) videos that might be corrupted by *in-capture* artifacts, such as motion blur, noise, and color artifacts [2, 19, 28–30].

Several video enhancement methods have been developed to reduce or remove in-capture artifacts in videos, from the perspective of color, contrast, brightness and stability, to bring people a more comfortable viewing experience [11, 17, 18]. Therefore, how to evaluate the best video quality consistently with human visual perception becomes pivotal. The quality assessment of enhanced versus UGC videos has different facets, which can lead current VQA methods to unsatisfactory results. First, VQA methods typically model videos that are very different from each other, and as a result, may not detect small variations between several versions of the same video content. Second, enhancement methods may introduce artifacts that do not belong to typical in-capture distortions and which are not correctly encoded by data-driven VQA models [7].

To address the need to evaluate different video enhancement methods, it is therefore necessary to collect datasets where human judgments are gathered for various versions of the same video and to develop VQA metrics for this specific problem. For this purpose, VQA Dataset for Perceptual Video Enhancement (VDPVE) [5] was recently proposed for the NTIRE 2023 Quality Assessment of Video Enhancement Challenge [6].

This work presents our method to the aforementioned NTIRE challenge, in which each participant is provided with a set of low quality videos processed with multiple enhancement techniques. For each enhanced video, human judgments (i.e. Mean Opinion Score, MOS) were gathered. The aim is therefore to obtain a method capable of estimating quality scores that highly correlate with the corresponding MOS. To this end, we propose a solution that encodes a video in terms of: technical aspects, such as the presence of distortions like noise and blur; non-technical factors, such as content preference and recommendation. The Disentangled Objective Video Quality Evaluator (DOVER) was designed precisely to model video quality with respect to previous perspectives [29]. In our method we exploit the DOVER to encode the video frames also making some changes that improve its effectiveness. As demonstrated in [1, 4], a model trained for the estimation of the overall qual-

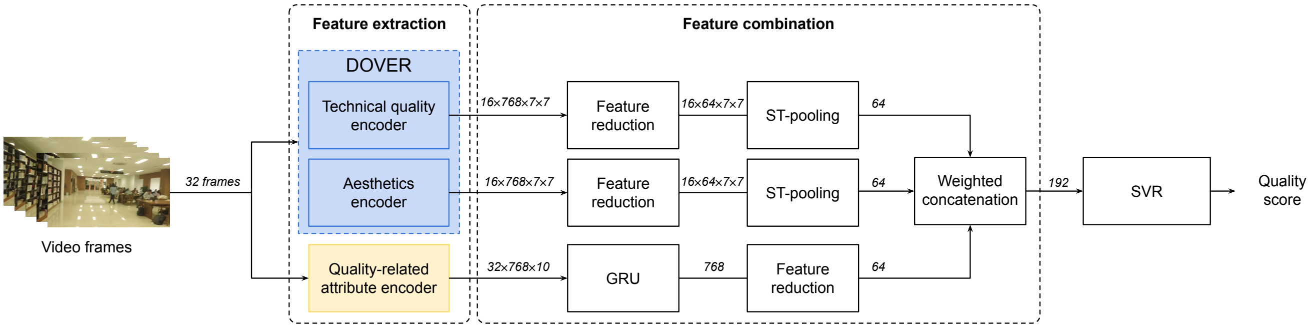


Figure 1. A graphical representation of the proposed method.

ity together with the estimation of color and artifact degradations improves quality prediction. For this reason, in the proposed method we encode video frames also regarding of quality-related attributes. The proposed method therefore encodes the videos in terms of three aspects, namely overall technical quality, quality-related attributes, and aesthetic quality. The resulting feature vectors are processed and concatenated. Finally, a Support Vector Regression (SVR) machine maps the feature vector into a video quality score.

The proposed method ranks sixth at the NTIRE 2023 Quality Assessment of Video Enhancement Challenge.

2. Method

As depicted in Figure 1, the proposed method consists of three components: (i) the *feature extraction* module includes three encoders capturing both technical and non-technical aspects of video sequences; (ii) the *feature combination* module provides a feature reduction and spatio-temporal pooling of the extracted features obtained by all the encoders; (iii) the *quality prediction* module exploits a SVR for mapping the feature vector into the video quality score.

2.1. Feature extraction module

The feature extraction module encodes video frames in terms of overall technical quality, quality-related attributes and aesthetics as described in the following sections.

2.1.1 Technical quality

The technical quality encoder exploits a tiny Video Swin Transformer [15] which takes as input sequences of 32 fragments of 224×224 resolution as in DOVER.

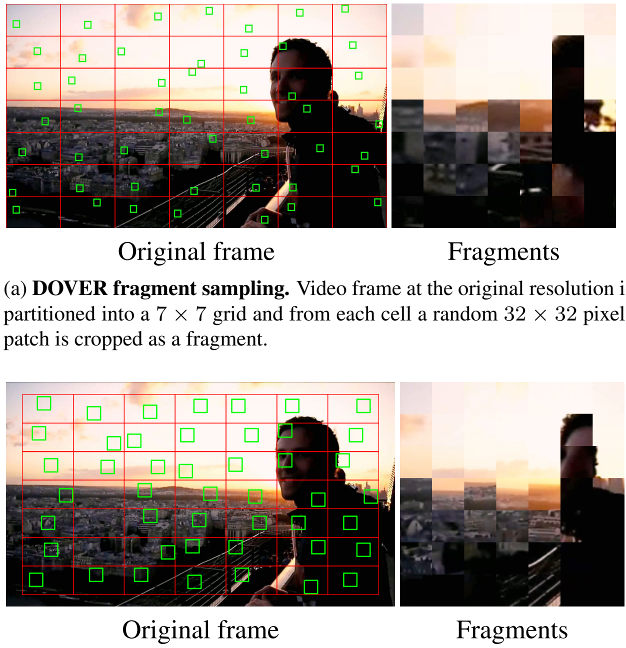
The frames of the sequence are sampled with a stride of 4. During training, the beginning of the sequence is randomly chosen. At inference time, two sequences are sampled, one at the beginning of the video and one at the end of the video, each sequence is processed independently and the predicted scores are averaged.

The fragment sampling pipeline is different from the one proposed in [29]. For videos with a resolution equal to or higher than 1080p, a crop is first performed to a size of 1708×960 (that is about 80% of the original frame area) and then a soft pooling operation [22] with kernel 2×2 is applied to halve the size at 480p. For videos with a resolution lower than 1080p, we directly crop a portion of the frame to 480p. Each 480p frame is logically divided into a 7×7 grid, patches of size 32×32 pixels are randomly selected from each grid cell to obtain a fragment of size 224×224 pixels. The proposed fragment sampling pipeline compared to that of DOVER ensures a higher coverage of the frame content, i.e. about 15% versus 3%, respectively. This means that more of the semantic content of the frame is preserved as can be seen in Figure 2. Additionally, using soft pooling to downscale video frames helps to better preserve video distortions and avoid masking effects.

Given a sequence of T fragments with resolution $H \times W$, the technical quality encoder first applies a 3D patch partitioning layer to obtain $\frac{T}{2} \times \frac{H}{4} \times \frac{W}{4} \times C$ tokens, where each token C consists of 96 features. After the previous layer, a stack of operations that reduce spatial resolution while preserving temporal resolution is applied to the feature blocks. The previous operations output a feature map $\mathbf{F}_t$ with shape $\frac{T}{2} \times \frac{H}{32} \times \frac{W}{32} \times 8C$. For a sequence of $T = 32$ fragments with resolution 224×224, we obtain a feature map with shape $16 \times 7 \times 7 \times 768$.

2.1.2 Quality-related attributes

The quality-related attribute encoder uses an EfficientNet-v2 [23] backbone followed by ten heads used to encode the overall technical quality as well as ten quality-related attributes, i.e. brightness, colorfulness, contrast, graininess, lightness, noisiness, saturation, and sharpness (two distinct heads because this attribute is present in the two datasets used for training). Each head has a convolutional part followed by fully connected layers. Our architecture is similar to the one proposed in [1], where the MobileNet-v2 [20]



(a) **DOVER fragment sampling.** Video frame at the original resolution is partitioned into a 7 × 7 grid and from each cell a random 32 × 32 pixels patch is cropped as a fragment.

(b) **Our fragment sampling.** A subregion covering 80% of the original frame area is cropped and then downscaled through soft pooling to 480p. The resulting frame is partitioned into a 7 × 7 grid and from each cell a random 32 × 32 pixels patch is cropped as a fragment.

Figure 2. Comparison between the DOVER and our fragment sampling pipeline on a 1080p video frame.

backbone is replaced with EfficientNet-v2 [23] because of its higher capability in capturing relevant quality information.

For a sequence of 32 equally-spaced frames, the quality-related attribute encoder extracts the features at the frame-level. For each frame, the central 720p crop is obtained and passed to the model. The features obtained before the fully connected layers of each head are flattened and then concatenated. The resulting feature map $\mathbf{F}_r$ with shape $32 \times 768 \times 10$ is used as input for the next processing steps.

2.1.3 Aesthetics

The aesthetic encoder consists of a tiny inflated-ConvNext [14] as defined in DOVER [29]. The aesthetic encoder has an overall view of the video as it uses a total of 32 equally-spaced frames covering the entire video sequence. The frames processed by this encoder are downscaled to a 224×224 resolution to increase efficiency and reduce sensitivity to technical distortions, such as blur and noise, which are captured instead by the technical quality encoder.

Given a sequence of T frames with resolution $H \times W$ pixels, the aesthetic encoder applies a 3D patch partitioning layer to obtain $\frac{T}{2} \times \frac{H}{4} \times \frac{W}{4} \times C$ tokens, where each token C consists of 96 features. After the previous layer, a stack of operations that reduce spatial resolution while preserving temporal resolution is applied to the feature blocks. Previous operations output a feature map $\mathbf{F}_t$ with shape $\frac{T}{2} \times \frac{H}{32} \times \frac{W}{32} \times 8C$. In our method a sequence of frames $T = 32$ with resolution 224×224 pixels is fed to the aesthetic encoder which produces in output a map with $16 \times 7 \times 7 \times 768$ features.

2.2. Feature combination

The feature combination module refines and combines the representations coming from the previously described encoders.

First, the temporal relationship between the frame-level features $\mathbf{F}_r$ is modeled by a Gated Recurrent Unit (GRU) [3]. Specifically, a 2-layer GRU is applied on the 32 feature vectors obtained by flattening the 768×10 matrix into 7680-dimensional vectors. The 768-dimensional output of the last frame is then reduced to a feature vector $\mathbf{f}_r$ with 64 dimensions by using a fully connected layer.

Second, the number of channels for technical, $\mathbf{F}_t$, and aesthetic, $\mathbf{F}_a$, features is reduced through two convolutional layers with kernel size and stride equal to 1. The resulting feature maps having shape $16 \times 64 \times 7 \times 7$ are both spatially and temporally reduced by average pooling.

Finally, the three feature vectors, namely $\mathbf{f}_t \in \mathbb{R}^{64}$, $\mathbf{f}_a \in \mathbb{R}^{64}$ and $\mathbf{f}_r \in \mathbb{R}^{64}$, are combined into the feature vector, $\mathbf{f} \in \mathbb{R}^{192}$, by applying the weighted concatenation procedure described in Algorithm 1.

Algorithm 1 Weighted concatenation procedure.

```
Input: Overall technical quality features $\mathbf{f}_t$, quality-related attribute features $\mathbf{f}_r$, aesthetic features $\mathbf{f}_a$, $\mathbf{W} \in \mathbb{R}^{128 \times 64}$ and $\mathbf{b} \in \mathbb{R}^{64}$.
Concatenate technical and aesthetic feature vectors, i.e. $\mathbf{f}_{t+a} = \mathbf{f}_t \oplus \mathbf{f}_a$.
Compute the scale vector $\mathbf{s} = \sigma(\mathbf{W}\mathbf{f}_{t+a} + \mathbf{b})$.
Scale quality-related features, $\hat{\mathbf{f}}_r = \mathbf{f}_r \circ \mathbf{s}$.
Concatenate feature vectors, $\mathbf{f} = \mathbf{f}_t \oplus \mathbf{f}_a \oplus \hat{\mathbf{f}}_r$.
Return: $\mathbf{f}$
```

2.3. Quality prediction

In the quality prediction module, the 192-dimensional feature vector is mapped into the final video quality score through a SVR with Radial Basis Function (RBF) kernel.

3. Experimental setup

In this section, the VDPVE dataset [5] is presented. The training procedure of the three encoders and the overall method is then described. Finally, the implementation details and evaluation metrics are detailed.

3.1. Dataset

In our experiments, we use the new VDPVE dataset [5] proposed for the NTIRE 2023 Quality Assessment of Video Enhancement Challenge. The dataset consists of 1211 enhanced videos and is divided into three subsets: the first one contains 600 videos with color, brightness, and contrast enhancements applied using 8 different methods; the second one contains 310 videos with deblurring performed using 5 different methods; the third one contains 301 videos deshaked by using 7 different methods. The dataset is split into training set (839 videos), validation set (119 videos) and test set (253 videos). Ground-truth scores and the information regarding the applied enhancement method are only available for the training videos. Both information is instead kept private for the validation and test videos by the challenge organizers. For this reason, we internally split the training set to conduct our experiments into an internal training set (668 videos) and an internal test set (171 videos), where each scene is either in the training set or in the test set.

3.2. Training procedure

The technical quality and aesthetic encoders of DOVER are pre-trained by mixing together the LSVQ [30], LIVE-VQC [21], KoNViD-1k [8], CVD2014 [16] and YouTube-UGC [27] datasets and fine-tuned using the VDPVE training set. Random video fragments are used at training time, as done in [29], while five-crop video fragments and 2 non-overlapped views of the video are used at inference time. The batch size is set to 5, the learning rate is set to 1×10^{-4} for the technical quality encoder and 1×10^{-3} for the rest of the network with a cosine decay. The model is trained for a total of 50 epochs.

The quality-related attribute encoder is trained using the CID [26] and SPAQ [4] datasets. Images are first randomly cropped to the closest resolution that is multiple of 720p, and then soft pooling is applied to obtain a 720p resolution. The batch size is set to 8. The model is trained for a total of 10K iterations. The learning rate is initially set to 1×10^{-4} and later decreased by a factor of 10 after 5K iterations. Random horizontal flip is used as data augmentation.

The three encoders are trained using the Norm-in-Norm (NiN) loss with monotonicity regularization [13] between the MOS, $\hat{y}$, and the predicted scores, y, as follows:

$$\mathcal{L} = \mathcal{L}_{NiN}(\hat{y}, y) + 0.3\mathcal{L}_{\text{mon}}(\hat{y}, y). \tag{1}$$

The SVR for the final score prediction is trained on the VDPVE training set. The required hyperparameters for the RBF kernel of the SVR are $\gamma = 12.20$ and $C = 364.83$, selected via Bayesian optimization framework[1] using Leave-One-Out cross-validation. The latter uses a surrogate model to approximate the objective function and chooses to optimize it according to some acquisition function. The surrogate model used is Random Forest, while the acquisition function is Upper Confidence Bound (UCB). The search value ranges for C and γ are $[0.01, 1000]$.

3.3. Implementation details

We implement the proposed method in Python3.8 using the PyTorch package with CUDA-v11.6 as back-end. The proposed model is trained on a workstation equipped with an Intel i7-4770 CPU @3.40GHz, 16GB DDR4 RAM 2400MHz, NVIDIA Titan Xp GPU with 3840 CUDA cores.

3.4. Evaluation metrics

The evaluation consists of the comparison of the predicted scores with the reference ground-truth, MOS. The Pearson's Linear Correlation Coefficient (PLCC) and Spearman's Rank Order Correlation Coefficient (SROCC) indexes are used consistently with the literature.

The PLCC is used to evaluate the linear correlation between MOS and predicted scores and it is calculated as follows:

$$PLCC = \frac{\sum_{i=1}^{N}(s_i - \mu_{s_i})(\hat{s}_i - \mu_{\hat{s}_i})}{\sqrt{\sum_{i=1}^{N}(s_i - \mu_{s_i})^2}\sqrt{\sum_{i=1}^{N}(\hat{s}_i - \mu_{\hat{s}_i})^2}} \tag{2}$$

Where N is the number of testing images, s_i and $\hat{s}_i$ respectively indicate the ground-truth and predicted quality scores of i-th image, and μ_{s_i} and $\mu_{\hat{s}_i}$ indicate the mean of them. Let d_i denote the difference between the ranks of i-th test image in ground-truth and predicted quality scores. Before calculating the PLCC index, the third-order polynomial nonlinear regression is performed.

The SROCC measures the monotonic relationship between MOS and predicted scores and it is defined as:

$$SROCC = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)}, \tag{3}$$

with N representing the number of testing images and $d_i = (\text{rank}(s_i) - \text{rank}(\hat{s}_i))$. Both metrics, PLCC and SROCC, are in $[-1, 1]$, and higher values indicate better performance.

The overall estimate of the goodness of a method is expressed in the *MainScore* which is obtained by ignoring the sign and reporting the average of the absolute values ((PLCC + SROCC)/2).

4. Results

In this section, we present the experiments and the related results that led us to the design of the final method. In Section 4.1, we first report the results on the validation set and then conduct a deeper analysis on our internal test set

[1] pyGPGO: https://pygpgo.readthedocs.io/ (last access: 05/04/2023)

Exp.	Main backbone	Fine-tuning head	Soft pool.	Feature combination	Multi-crop evaluation	Quality pred.	Train set	SROCC	PLCC	MainScore
1	FAST-VQA [28]	–		–		MLP	Subset	0.6252	0.6035	0.6144
2		IP-NLR		–		MLP	Subset	0.6951	0.6914	0.6933
3	DOVER [29]	Tech. & Aesth.		–		AVG	Subset	0.7251	0.6852	0.7052
4		Tech.		–		AVG	Subset	0.7365	0.6876	0.7120
5		Tech.	✓	–		AVG	Subset	0.7559	0.7129	0.7344
6		Tech.	✓	Tech. & Aesth.		MLP	Subset	0.7577	0.7394	0.7486
7		Tech.	✓	Tech. & Aesth.	✓	MLP	Subset	0.7660	0.7430	0.7545
8		Tech.	✓	Tech. & Aesth. & Attr.	✓	SVR	Subset	0.7671	0.7666	0.7669
9		Tech.	✓	Tech. & Aesth. & Attr.	✓	SVR	Entire	0.7850	0.7790	0.7820
10	Quality-related attribute encoder					MLP	Other	0.4916	0.5766	0.5341

Table 1. Results for the different configurations of the proposed method on the VDPVE validation set. Here we train the model using the internal training set, and use the internal test set to select the best results. IP-NLR: Intra-Patch Non-Linear Regression Head, MLP: Multi-Layer Perceptron, AVG: Average of the predicted scores, SVR: Support Vector Regression, Other: Model trained on datasets for image quality assessment.

in Section 4.2. The comparison between the performance obtained by our method on the test set is carried out with other state-of-the-art methods (see Section 4.3) and with the other participants in the NTIRE 2023 Quality Assessment of Video Enhancement Challenge (see Section 4.4).

4.1. Ablation study

In order to motivate the different choices adopted for the development of the proposed solution, we construct ablation experiments to show how results change when one or more features are omitted. Table 1 reports the results obtained under different configurations of the proposed method on the validation set of the VDPVE dataset [5].

Main backbone. The choice of backbone for encoding video sequences is crucial. The use of FAST-VQA [28] and DOVER [29] has been experimented due to their effectiveness in estimating the quality of UGC videos. Experiments have been carried out by fine-tuning the whole model or only the head. As it is possible to see from Table 1, DOVER outperforms FAST-VQA and for this reason it is chosen as the backbone.

End-to-end DOVER. The DOVER architecture consists of the technical quality encoder and the aesthetics encoder [29]. When only the technical encoder is fine-tuned on the dataset, the results are better than when both the encoders are fine-tuned. For this reason, we fine-tune only the technical encoder.

Fragment sampling procedure. The technical encoder uses frame fragments to infer the technical video quality. However, since fragments are typically obtained from the original video without any downscaling operation [28, 29], they have a limited view of the scene depicted in a video frame, resulting in poor performance. As we can see, downscaling frames to a smaller resolution (i.e., 1708 × 960) before creating fragments allows to have more content information, leading to a considerable increase in performance (about 2.24%). Moreover, using five-crop fragments for evaluation increases the spatial frame coverage, further boosting the overall performance.

Late vs. early fusion strategy. In DOVER [29], each encoder individually predicts the quality score of a video, and the final score is their simple average. Instead, we propose an early-fusion strategy, where we directly predict one score learned by the combination of the features obtained from all the encoders. This more sophisticated mechanism allows the model to better make use of the encoded information, obtaining a considerable improvement in performance (about 1.42%).

Feature combination. The naive quality-related attribute encoder alone (i.e. the scores are predicted for each frame and the video score is their simple average) without any fine-tuning on the VDPVE dataset [5] obtains a MainScore of 0.5341. It is possible to see that adding the quality-related attributes provides complementary information and contributes to increasing the performance by about 1.24%.

Subset vs. entire training set. Once the method and hyperparameters are defined, the model is trained on the entire training set without considering the internal split. In Table 1, *experiment 9* corresponds to our method trained on the entire training set. We highlight that the availability of more training data allows the model to generalize better on the validation data. In particular, *experiment 9* improves the MainScore by about 2% with respect to *experiment 8*.

4.2. Internal test set results

The configurations presented in Section 4.1 are also evaluated on the internal test set with the aim of obtaining an effectiveness feedback. The quality-related attribute encoder is directly evaluated on VDPVE videos without any fine-tuning, while the overall model is trained with the Bayesian

Exp.	SROCC	PLCC	MainScore
1	0.7140	0.7288	0.7214
2	0.7692	0.7364	0.7528
3	0.7929	0.7818	0.7874
4	0.8298	0.8289	0.8294
5	0.7835	0.7957	0.7896
6	0.8303	0.8181	0.8242
7	0.8282	0.8150	0.8216
8	0.8217	0.8149	0.8183

Table 2. Results for the different configurations of the proposed method related to the internal test set.

Enhancement type	PLCC	SROCC	MainScore
A (Enhanced)	0.8399	0.8450	0.8424
B (Stabilized)	0.5709	0.5684	0.5697
C (Deblurred)	0.7143	0.6784	0.6964

Table 3. Internal test set results of the proposed method grouped by enhancement type.

optimization using Leave-One-Out cross-validation. Then the best hyper-parameters are used for the evaluation on the internal test set.

Table 2 underlines that, although the internal test set is designed to reflect the MOS distribution of the training set, the results are not correlated with the submissions on the validation set reported in Table 1.

Figure 3 presents the scatter plots of the quality-related attribute encoder and the proposed model. A logistic regression function is drawn for highlighting the silhouette of the fit. We can see that the quality predictions of the quality-related attribute encoder are more spread than the ones of the overall model, showing the need for feature combining. This is also exemplified in Figure 4, which presents two examples of predictions where the quality-related attribute encoder (blue line) directly predicts the quality score of each video frame. The frame-level predictions are highly sensitive to the scene changes and the overall frame quality. To this end, the adaptive combination of quality-related features with technical and aesthetic features uniforms and weighs both features to obtain the final video score (red points).

The internal test split allows also to analyze the evaluation metrics between the different enhancement types. From Table 3, it is possible to notice that our model performs better on enhanced videos than on stabilized and deblurred ones. This may be explained by two reasons. First, the number of enhanced videos is about twice the number of stabilized and deblurred videos. Second, the quality-related attribute encoder is trained using images containing various enhancements and therefore can provide more powerful features leading to better performance on enhanced videos than on the other video types.

Method	SROCC	PLCC	MainScore
VIDEVAL [24]	0.5005	0.4724	0.4865
RAPIQUE [25]	0.5434	0.5393	0.5414
TLVQM [9]	0.5474	0.5509	0.5492
V-BLIINDS [19]	0.5652	0.5503	0.5578
VSFA [12]	0.5871	0.5424	0.5648
BVQA [10]	0.6995	0.6674	0.6835
FAST-VQA [28]	0.7350	0.7310	0.7330
Ours	**–**	**–**	**0.7859**

Table 4. Comparison with state-of-the-art methods for video quality assessment on the test set of the VDPVE dataset. The higher the better.

4.3. Comparison with state-of-the-art methods

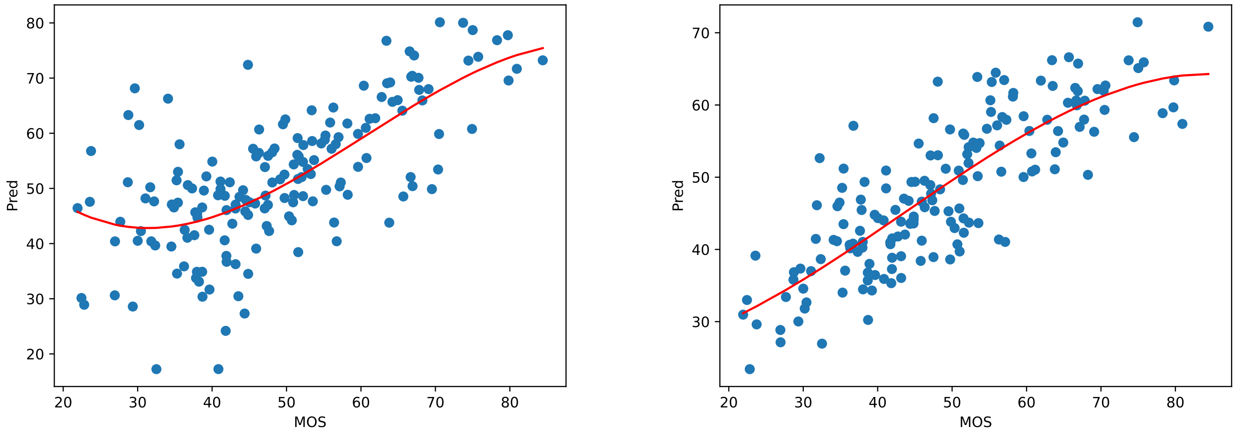
We compare the proposed solution with other state-of-the-art VQA methods. In particular, we use V-BLIINDS [19], TLVQM [9], VIDEVAL [24], RAPIQUE [25], FAST-VQA [28], VSFA [12] and BVQA [10]. The results evaluated in terms of SROCC and PLCC are reported in Table 4. We can see that our solution considerably outperforms the other methods, obtaining a higher mean score of about 0.05 with respect to FAST-VQA [28], which is the second-best method.

4.4. Results of the NTIRE 2023 Quality Assessment of Video Enhancement Challenge

The NTIRE 2023 Quality Assessment of Video Enhancement Challenge has the goal of developing a solution for video quality assessment capable to produce high-quality results with the best correlation to the reference ground truth (i.e., Mean Opinion Score). A total of 19 teams were involved in the final stage of the challenge and were included in the leaderboard. The final results of the competition are reported in Table 5 and are related to the test set split of the VDPVE dataset [5]. Here, the proposed method won sixth place.

5. Conclusion

In this paper, we presented a novel method to evaluate the quality of enhanced videos. Our method relies on three different neural networks to get multiple information from video sequences related to technical quality, aesthetic quality and several quality-related aspects, such as video sharpness, contrast and saturation. The features obtained from these three encoders are adaptively combined by taking into account their relevance in the final predic-



(a)

(b)

Figure 3. Scatter plots of the predicted quality scores versus MOS on the VDPVE internal test set for (a) the quality-related attribute encoder, and (b) the overall proposed model. The PLCC and SROCC for (a) are 0.5352 and 0.5462 and for (b) are 0.8217 and 0.8149.

Team	MainScore	Ranking
TB-VQA	0.8576	1
QuoVadis	0.8396	2
OPDAI	0.8289	3
Transsion Image Algorithm Team	0.8199	4
VCCIP	0.7994	5
IVL	**0.7859**	**6**
HXHHXH	0.7850	7
fmgtv	0.7727	8
KKARC	0.7635	9
DTVQA	0.7325	10
sqiyx	0.7302	11
402Lab	0.7136	12
NTU-SLab	0.6990	13
HNU_LIMMC	0.6972	14
Drealitym	0.6923	15
LION_Vaader	0.6863	16
Caption_Timor	0.6596	17
IVLab	0.6499	18
one_for_all	0.5851	19

Table 5. Results of the NTIRE 2023 Quality Assessment of Video Enhancement Challenge. Our solution won sixth place in the competition.

tion. The video quality score is finally obtained using a Support Vector Regression machine. Experimental results conducted on the new VDPVE dataset [5] show the effectiveness of our method, which considerably outperforms other existing state-of-the-art approaches for video quality assessment. In the context of the NTIRE 2023 Quality Assessment of Video Enhancement Challenge, our method won sixth place, further demonstrating its effectiveness.

References

[1] Mirko Agarla, Luigi Celona, and Raimondo Schettini. An efficient method for no-reference video quality assessment. *MDPI Journal of Imaging*, 7(3):55, 2021. 1, 2

[2] Mirko Agarla, Luigi Celona, and Raimondo Schettini. Predicting video memorability using a model pretrained with natural language supervision. In *MediaEval Multimedia Benchmark Workshop Working Notes*, 2023. 1

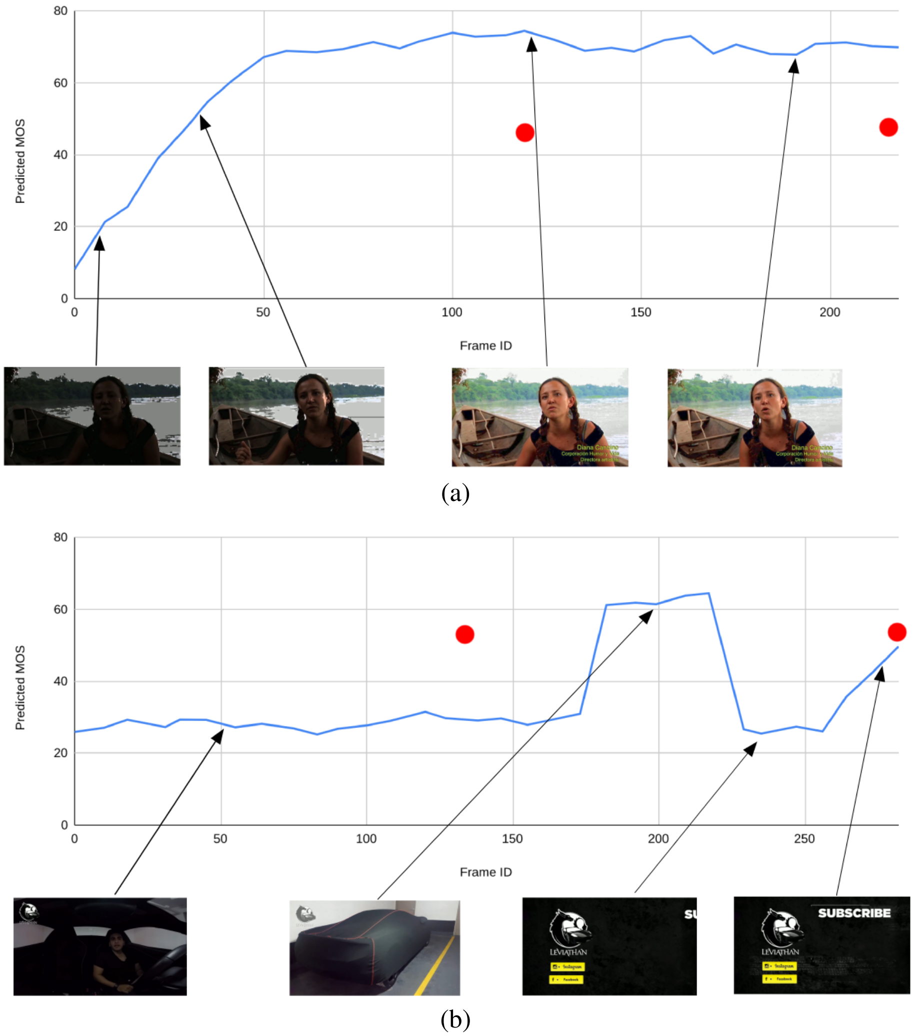
[3] Kyunghyun Cho, Bart van Merriënboer, Çağlar Gulçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *EMNLP*, pages 1724–1734, 2014. 3

[4] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *Conference on Computer Vision and Pattern Recognition*, pages 3677–3686. IEEE/CVF, 2020. 1, 4

[5] Yixuan Gao, Yuqin Cao, Tengchuan Kou, Wei Sun, Yunlong Dong, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. Vdpve: Vqa dataset for perceptual video enhancement. *arXiv preprint arXiv:2303.09290*, 2023. 1, 3, 4, 5, 6, 7

[6] Yixuan Gao, Yuqin Cao, Tengchuan Kou, Wei Sun, Yunlong Dong, Xiaohong Liu, Xiongkuo Min, Guangtao Zhai, et al. Ntire 2023 challenge on quality assessment for video enhancement. 2023. 1

[7] Jinjin Gu, Haoming Cai, Chao Dong, Jimmy S Ren, Radu Timofte, Yuan Gong, Shanshan Lao, Shuwei Shi, Jiahao Wang, Sidi Yang, et al. Ntire 2022 challenge on perceptual image quality assessment. In *Conference on Computer Vision and Pattern Recognition*, pages 951–967. IEEE/CVF, 2022. 1



(a)

(b)

Figure 4. Predictions of video A0002_06 (a) and video C0043_02 (b) with MOS of 37.88 and 59.82 respectively. The frame-level predictions of the quality-related attribute encoder and the predictions of two views of the overall proposed model are depicted with blue line and red points, respectively. The predictions of the two views are related to the frames $[0 - 124]$ and $[126 - 250]$ for (a) and $[9 - 133]$ and $[158 - 282]$ for (b).

[8] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe. The konstanz natural video database (konvid-1k). In *2017 Ninth international conference on quality of multimedia experience (QoMEX)*, pages 1–6. IEEE, 2017. 4

[9] Jari Korhonen. Two-level approach for no-reference consumer video quality assessment. *IEEE Transactions on Image Processing*, 28(12):5923–5938, 2019. 6

[10] Bowen Li, Weixia Zhang, Meng Tian, Guangtao Zhai, and Xianpei Wang. Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9):5944–5958, 2022. 6

[11] Chongyi Li, Chunle Guo, Linghao Han, Jun Jiang, Ming-Ming Cheng, Jinwei Gu, and Chen Change Loy. Low-light image and video enhancement using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9396–9416, 2021. 1

[12] Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2351–2359, 2019. 6

[13] Dingquan Li, Tingting Jiang, and Ming Jiang. Norm-in-norm loss with faster convergence and better performance for image quality assessment. In *International Conference on Multimedia*, pages 789–797. ACM, 2020. 4

[14] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Conference on Computer Vision and Pattern Recognition*, pages 11976–11986. IEEE/CVF, 2022. 3

[15] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Conference on Computer Vision and Pattern Recognition*, pages 3202–3211. IEEE/CVF, 2022. 2

[16] Mikko Nuutinen, Toni Virtanen, Mikko Vaahteranoksa, Tero Vuori, Pirkko Oittinen, and Jukka Häkkinen. CVD2014 - A database for evaluating no-reference video quality assessment algorithms, July 2016. Our research group home page: http://www.helsinki .fi/psychology/groups/visualcognition/. 4

[17] Yunbo Rao and Leiting Chen. A survey of video enhancement techniques. *J. Inf. Hiding Multim. Signal Process.*, 3(1):71–99, 2012. 1

[18] Claudio Rota, Marco Buzzelli, Simone Bianco, and Raimondo Schettini. Video restoration based on deep learning: a comprehensive survey. *Springer Artificial Intelligence Review*, pages 1–48, 2022. 1

[19] Michele A Saad, Alan C Bovik, and Christophe Charrier. Blind prediction of natural video quality. *IEEE Transactions on image Processing*, 23(3):1352–1365, 2014. 1, 6

[20] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 2

[21] Zeina Sinno and Alan Conrad Bovik. Large-scale study of perceptual video quality. *IEEE Transactions on Image Processing*, 28(2):612–627, 2018. 4

[22] Alexandros Stergiou, Ronald Poppe, and Grigorios Kalliatakis. Refining activation downsampling with softpool. In *International Conference on Computer Vision*, pages 10357–10366. IEEE/CVF, 2021. 2

[23] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on Machine Learning*, pages 10096–10106. PMLR, 2021. 2, 3

[24] Zhengzhong Tu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik. Ugc-vqa: Benchmarking blind video quality assessment for user generated content. *IEEE Transactions on Image Processing*, 30:4449–4464, 2021. 6

[25] Zhengzhong Tu, Xiangxu Yu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik. Rapique: Rapid and accurate video quality prediction of user generated content. *IEEE Open Journal of Signal Processing*, 2:425–440, 2021. 6

[26] Toni Virtanen, Mikko Nuutinen, Mikko Vaahteranoksa, Pirkko Oittinen, and Jukka Häkkinen. Cid2013: A database for evaluating no-reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 24(1):390–402, 2014. 4

[27] Yilin Wang, Sasi Inguva, and Balu Adsumilli. Youtube ugc dataset for video compression research. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–5. IEEE, 2019. 4

[28] Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Fastvqa: Efficient end-to-end video quality assessment with fragment sampling. In *European Conference on Computer Vision*, pages 538–554. Springer, 2022. 1, 5, 6

[29] Haoning Wu, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Disentangling aesthetic and technical effects for video quality assessment of user generated content. *arXiv preprint arXiv:2211.04894*, 2022. 1, 2, 3, 4, 5

[30] Zhenqiang Ying, Maniratnam Mandal, Deepti Ghadiyaram, and Alan Bovik. Patch-vq:'patching up'the video quality problem. In *Conference on Computer Vision and Pattern Recognition*, pages 14019–14029. IEEE/CVF, 2021. 1, 4