NTIRE 2023 Quality Assessment of Video Enhancement Challenge

NTIRE 2023 Quality Assessment of Video Enhancement Challenge

Xiaohong Liu* Xiongkuo Min* Wei Sun* Yulun Zhang * Kai Zhang*
Radu Timofte* Guangtao Zhai* Yixuan Gao* Yuqin Cao* Tengchuan Kou*
Yunlong Dong* Ziheng Jia * Yilin Li† Kai Zhao† Heng Cong† Hang Shi†
Zhiliang Ma† Mirko Agarla† Zhiwei Huang† Hongye Liu† Ironhead Chuang†
Haotian Fan† Shiqi Zhou† Yu Lai† Wenqi Wang† Haoning Wu†
Chunzheng Zhu† Shiling Zhao† Hanene Brachemi Meftah† Tengfei Shi†
Azadeh Mansouri †

Abstract

This paper reports on the NTIRE 2023 Quality Assessment of Video Enhancement Challenge, which will be held in conjunction with the New Trends in Image Restoration and Enhancement Workshop (NTIRE) at CVPR 2023. This challenge is to address a major challenge in the field of video processing, namely, video quality assessment (VQA) for enhanced videos. The challenge uses the VQA Dataset for Perceptual Video Enhancement (VDPVE), which has a total of 1211 enhanced videos, including 600 videos with color, brightness, and contrast enhancements, 310 videos with deblurring, and 301 deshaked videos. The challenge has a total of 167 registered participants. 61 participating teams submitted their prediction results during the development phase, with a total of 3168 submissions. A total of 176 submissions were submitted by 37 participating teams during the final testing phase. Finally, 19 participating teams submitted their models and fact sheets, and detailed the methods they used. Some methods have achieved better results than baseline methods, and the winning methods have demonstrated superior prediction performance.

1. Introduction

The importance of video quality assessment (VQA) in the field of video processing is self-evident. It can guide the development of video processing algorithms such as video capture, enhancement, transmission, and display. Therefore, VQA methods have been widely used to evaluate the quality of various videos, such as user-generated content (UGC) videos [64], high dynamic range (HDR) videos [53], tone-mapped videos [84], compressed videos [38], and so on. Recently, after capturing a large number of videos, people would like to first enhance certain attributes of the videos, such as contrast, brightness, and color, and then upload these enhanced videos to social medias. Therefore, many video enhancement methods have been proposed [8,21,39,40,59,60,92]. However, the quality levels of these videos processed by various video enhancement methods are different, and evaluating the quality of these enhanced videos is not easy. Therefore, it is very important to propose an efficient VQA method to accurately predict the quality of enhanced videos.

This NTIRE 2023 Quality Assessment of Video Enhancement Challenge aims to promote the development of the VQA methods for enhanced videos to guide the improvement and enhancement of the performance of video enhancement methods, thereby improving the viewing experience of videos [42, 58]. We use the VQA Dataset for Perceptual Video Enhancement (VDPVE) [15] for this challenge. This dataset has 1211 videos with different enhancements, which can be divided into three sub-datasets: the first sub-dataset has 600 videos with color, brightness, and contrast enhancement; the second sub-dataset has 310 videos with deblurring; and the third sub-dataset has 301 deshaked videos. Each enhanced video in the VDPVE has 20 subjective opinion scores.

This is the first time that a quality assessment of video enhancement challenge is held at the NTIRE workshop. The challenge has a total of 167 registered participants. 61 participating teams submitted their prediction results during the development phase, with a total of 3168 submissions. A total of 176 prediction results were submitted by 37 participating teams during the final testing phase. Finally, 19

*The organizers of the NTIRE 2023 Quality Assessment of Video Enhancement Challenge.
†The valid participating teams of the NTIRE 2023 Quality Assessment of Video Enhancement Challenge.
The NTIRE 2023 website: https://cvlai.net/ntire/2023/

valid participating teams submitted their final models and fact sheets. They have provided detailed introductions to their VQA methods for enhanced videos. We provide the detailed results of the challenge in Secion 4 and Secion 5. We hope that this challenge can promote the development of VQA methods for video enhancement.

This challenge is one of the NTIRE 2023 Workshop [1] series of challenges on: night photography rendering [61], HR depth from images of specular and transparent surfaces [89], image denoising [35], video colorization [28], shadow removal [70], quality assessment of video enhancement [41], stereo super-resolution [74], light field image super-resolution [76], image super-resolution ($\times 4$) [94], 360° omnidirectional image and video super-resolution [5], lens-to-lens bokeh effect transformation [10], real-time 4K super-resolution [11], HR nonhomogenous dehazing [3], efficient super-resolution [34].

2. Related Work

2.1. VQA dataset

The successful construction of VQA datasets is the foundation for proposing effective VQA models. The first successful VQA dataset is the LIVE Video Quality Database [57], which has 160 videos with compression and transmission distortions. IVP [91] provides 138 videos with compression and transmission distortions. MCL-V [37] contains 96 distorted videos with two typical video distortion types: compression and compression followed by scaling. MCL-JCV [73] is an H.264/AVC coded video quality dataset consisting of 30 video clips of a wide content variety. In recent years, due to the explosion in the number of UGC videos, many researchers have created VQA datasets for UGC videos. For example, Nuutinen *et al*. introduced the CVD2014 [54], which consists of 234 videos captured by 78 different video capture devices. The authors in [23] constructed one of the most famous VQA datasets for UGC videos, called KoNViD-1k. This dataset has 1200 UGC videos with authentic distortions. The other two popular VQA datasets for UGC videos are the LIVE-VQC [62] and the YouTube-UGC [75], with 585 and 1380 UGC videos, respectively. Besides, we provided a VQA dataset for video enhancement called the VDPVE [15]. The videos in this dataset were processed by various video enhancement methods, including 600 videos with color, brightness, and contrast enhancements, 310 videos with deblurring, and 301 deshaked videos. This dataset is used to test the performance of methods proposed by different participating teams in this challenge.

[1] https://cvlai.net/ntire/2023/

2.2. VQA model

The traditional VQA methods are handcrafted feature-based models. This kind of methods first calculate the quality of each frame of a video by extracting quality features, and then obtain the video quality score [16, 51, 90]. For example, V-BLIINDS [55] is a spatio-temporal natural scene statistics (NSS) model, which can quantify motion coherency in video scenes. TLVQM [30] is based on the idea of calculating features at two levels, that is, first calculating the low complexity features of the entire sequence, and then extracting high complexity features from subsets of representative video frames. VIDEVAL [68] calculates video quality by extracting abundant spatio-temporal features such as motion, jerkiness, blurriness, noise, blockiness, color, and so on. RAPIQUE [69] combines the advantages of both quality-aware scene statistics features and semantics-aware deep convolutional features to calculate video quality.

In addition to traditional VQA methods, deep learning-based VQA methods also attract researchers' attention [6, 16, 49, 65, 85, 95]. For example, VSFA [32] first extracts semantic features from a pre-trained convolutional neural network (CNN), and then uses a gated recursive unit network to extract the temporal relationship between semantic features of video frames to predict video quality. BVQA [44] uses a feature encoder to directly extract spatio-temporal representations from videos to predict video quality. SimpleVQA [64] trains an end-to-end spatial feature extraction network to directly learn quality-aware spatial features from video frames, and extracts motion features to measure temporally related distortions that cannot be modeled by spatial features at the same time to predict video quality.

3. NTIRE 2023 Quality Assessment of Video Enhancement Challenge

We organize the NTIRE 2023 Quality Assessment of Video Enhancement Challenge in order to promote the development of objective VQA methods for video enhancement. The main goal of the challenge is to predict the perceptual quality of enhanced videos, which can also promote the development of video enhancement methods. Details about the challenge are as follows:

3.1. Overview

The challenge has only one track, that is, the task of predicting the perceptual quality of an enhanced video based on a set of prior examples of videos and their perceptual quality labels. The challenge uses the training, validation, and testing sets as defined in the VDPVE [15]. As the final result, the participants in the challenge are asked to submit predicted scores for the given testing set.

3.2. Dataset

The VDPVE has 1211 videos with different enhancements, which can be divided into three sub-datasets: the first sub-dataset has 600 videos with color, brightness, and contrast enhancements; the second sub-dataset has 310 videos with deblurring; and the third sub-dataset has 301 deshaked videos. The resolution of all videos in the VDPVE is 1280×720. The video length is 8s or 10s.

In the first sub-dataset, eight enhancement methods are utilized to enhance the color, brightness, and contrast of 79 videos: ACE [18], AGCCPF [20], BPHEME [72], MBLLEN [50], SGZSL [99], DCC-Net [96], and two commercial software: CapCut and Adobe Premiere Pro. In the second sub-dataset, we utilize five enhancement methods to deblur the 62 blurred videos, including ESTRNN [100], DeblurGANv2 [31], FGST [36], BasicVSR++ [8], and Adobe Premiere Pro. In the third sub-dataset, seven enhancement methods are utilized to stabilize 43 videos, including GlobalFlowNet [26], DIFRINT [9], PWStableNet [98], Yu [88], CapCut (most stable mode), CapCut (minimum cropping mode), and Adobe Premiere Pro.

We invited 21 subjects (20 valid subjects) to rate all enhanced videos in the VDPVE. After normalizing and averaging the subjective opinion scores, the mean opinion score (MOS) of each video can be obtained. Furthermore, we randomly split the enhanced videos in the VDPVE into a training set, a validation set, and a testing set according to the ratio of $7:1:2$. The enhanced videos generated from the same original video are divided into the same set. The numbers of enhanced videos in the training set, validation set, and testing set are 839, 119, and 253, respectively.

3.3. Evaluation protocol

In the challenge, the main scores are utilized to determine the rankings of participating teams. We ignore the sign and calculate the average of Spearman rank-order correlation coefficient (SRCC) and Person linear correlation coefficient (PLCC) as the main score:

$$\text{Main Score} = (|\text{SRCC}| + |\text{PLCC}|)/2. \quad (1)$$

SRCC measures the prediction monotonicity, while PLCC measures the prediction accuracy. Better VQA methods should have larger SRCC and PLCC values. Before calculating PLCC index, we perform the third-order polynomial nonlinear regression. By combining SRCC and PLCC, the main scores can comprehensively measure the performance of participating methods.

3.4. Challenge phases

The whole challenge consists of two phases: the developing phase and the testing phase. In the developing phase, the participants can access to the enhanced videos of the training set and the corresponding MOSs. Participants can be familiar with dataset structure and develop their VQA methods. We also release the enhanced videos of the validation set without corresponding MOSs. Participants can utilize their VQA methods to predict the quality scores of the validation set and upload the results to the server. The participants can receive immediate feedback and analyze the effectiveness of their methods on the validation set. The validation leaderboard is available. In the testing phase, the participants can access to the enhanced videos of the testing set without MOSs and upload the final predicted scores of the testing set before the challenge deadline. Each participating team needs to submit a source code/executable and a fact sheet, which is a detailed description file of the proposed method and the corresponding team information. The final results are then sent to the participants.

4. Challenge Results

A total of 37 teams participated in the testing phase of NTIRE 2023 Quality Assessment of Video Enhancement Challenge, and 19 teams submitted their final codes/executables and fact sheets. Table 1 summarizes the main results and important information of the 19 valid teams. The methods of these teams are briefly introduced in Section 5 and the team members are listed in Appendix B.

4.1. Baselines

We compare the performance of submitted methods with several state-of-the-art NR VQA methods on the testing set, including V-BLIINDS [55], TLVQM [30], VIDEVAL [68], RAPIQUE [69], FastVQA [79], VSFA [32], BVQA [44], and SimpleVQA [64]. V-BLIINDS [55], TLVQM [30], VIDEVAL [68], and RAPIQUE [69] are the handcrafted feature-based VQA models. FastVQA [69], VSFA [32], BVQA [44] and SimpleVQA [64] are deep learning-based VQA models. VSFA [32], BVQA [44], and SimpleVQA [64] utilize CNN models as the network backbone, while FastVQA [69] utilizes the transformer as the network backbone.

4.2. Discussion

The main results of 19 teams' methods and the baseline methods are shown in Table 1, It can be seen that most of existing NR VQA methods are not ideal on VDPVE testing set, while the submitted methods have basically achieved good results. It means that these methods are closer to human visual perception when used to evaluate enhanced videos. 9 teams achieve relatively better performance than FastVQA, which has good performance on the in-the-wild VQA task. Furthermore, the main scores of 4 teams exceed 0.8. The championship team achieves the SRCC score of 0.8576 and the PLCC score of 0.8396. Figure 1 shows scatter plots of predicted scores versus MOSs for the 19 teams'

Table 1. Quantitative results for the NTIRE 2023 Quality Assessment of Video Enhancement Challenge.

Rank	Team	Leader	Main Score	SRCC	PLCC
1	TB-VQA	Yilin Li	0.8576	0.8493	0.8659
2	QuoVadis	Kai Zhao	0.8396	0.8408	0.8383
3	OPDAI	Heng Cong	0.8289	0.8261	0.8317
4	TIAT	Hang Shi	0.8199	0.8163	0.8236
5	VCCIP	Zhiliang Ma	0.7994	0.7962	0.8026
6	IVL	Mirko Agarla	0.7859	0.7896	0.7822
7	HXHHXH	Zhiwei Huang	0.7850	0.7879	0.7821
8	fmgtv	Hongye Liu	0.7727	0.7756	0.7698
9	KK-ARC	Ironhead Chuang	0.7635	0.7663	0.7607
10	DTVQA	Haotian Fan	0.7325	0.7357	0.7294
11	sqiyx	Shiqi Zhou	0.7302	0.7246	0.7358
12	402Lab	Yu Lai	0.7136	0.7150	0.7123
13	one_for_all	Wenqi Wang	0.6990	0.7087	0.6893
14	NTU-SLab	Haoning Wu	0.6972	0.7019	0.6924
15	HNU-LIMMC	Chunzheng Zhu	0.6923	0.6975	0.6872
16	Drealitym	Shiling Zhao	0.6863	0.6900	0.6826
17	LION_Vaader	Hanene Brachemi Meftah	0.6596	0.6674	0.6518
18	Caption Timor	Tengfei Shi	0.6499	0.6524	0.6475
19	IVP-LAB	Azadeh Mansouri	0.5851	0.5887	0.5814
Baseline		FastVQA	0.7330	0.7350	0.7310
		BVQA	0.6835	0.6995	0.6674
		SimpleVQA	0.6347	0.6340	0.6354
		VSFA	0.5648	0.5871	0.5424
		V-BLIINDS	0.5578	0.5652	0.5503
		TLVQM	0.5492	0.5474	0.5509
		RAPIQUE	0.5414	0.5434	0.5393
		VIDEVAL	0.4865	0.5005	0.4724

methods on VDPVE testing set. The curves shown in Figure 1 are obtained by a four-order polynomial nonlinear fitting. We can observe that the predicted scores obtained by the top team methods have higher correlations with the MOSs. They can not only meet the need to predict quality scores for enhanced videos but also contribute to improving the performance of video enhancement methods.

5. Challenge Methods

5.1. TB-VQA

Team TB-VQA is the final winner of the challenge. They propose a NR VQA method [82] based on Swin Transformer with improved spatio-temporal feature fusion and data sample augmentation strategy. The network is developed on top of SimpleVQA [64] and is composed of two key components: the spatial feature extraction module and spatio-temporal feature fusion module. In the spatial feature extraction module, inspired by its strong modeling capabilities and representative performance of visual priors including hierarchy, locality, and translation invariance, they exploit Swin Transformer V2 [45], pre-trained on ImageNet-1K [12], as the backbone network to extract spatial features. Specially, they adopt the features extracted from the last two Transformer blocks to take advantage of deep semantic information in video quality representation. In the spatio-temporal feature fusion module, they introduce an 1×1 convolutional layer, which deepens the spatial features extracted from the intermediate stages of the pre-trained network, to mitigate the gap between shallow and deep features. The semantic features, the deepened features, and the temporal features (originally from the motion feature extraction module in [64]) are flattened and fused as the final features for video quality prediction. To further enhance the robustness of the model, a data augmentation strategy is performed to increase the number of video frames in the training phase. Specially, they first train their model on the LSVQ [86] dataset. Then, they build a dataset similar to

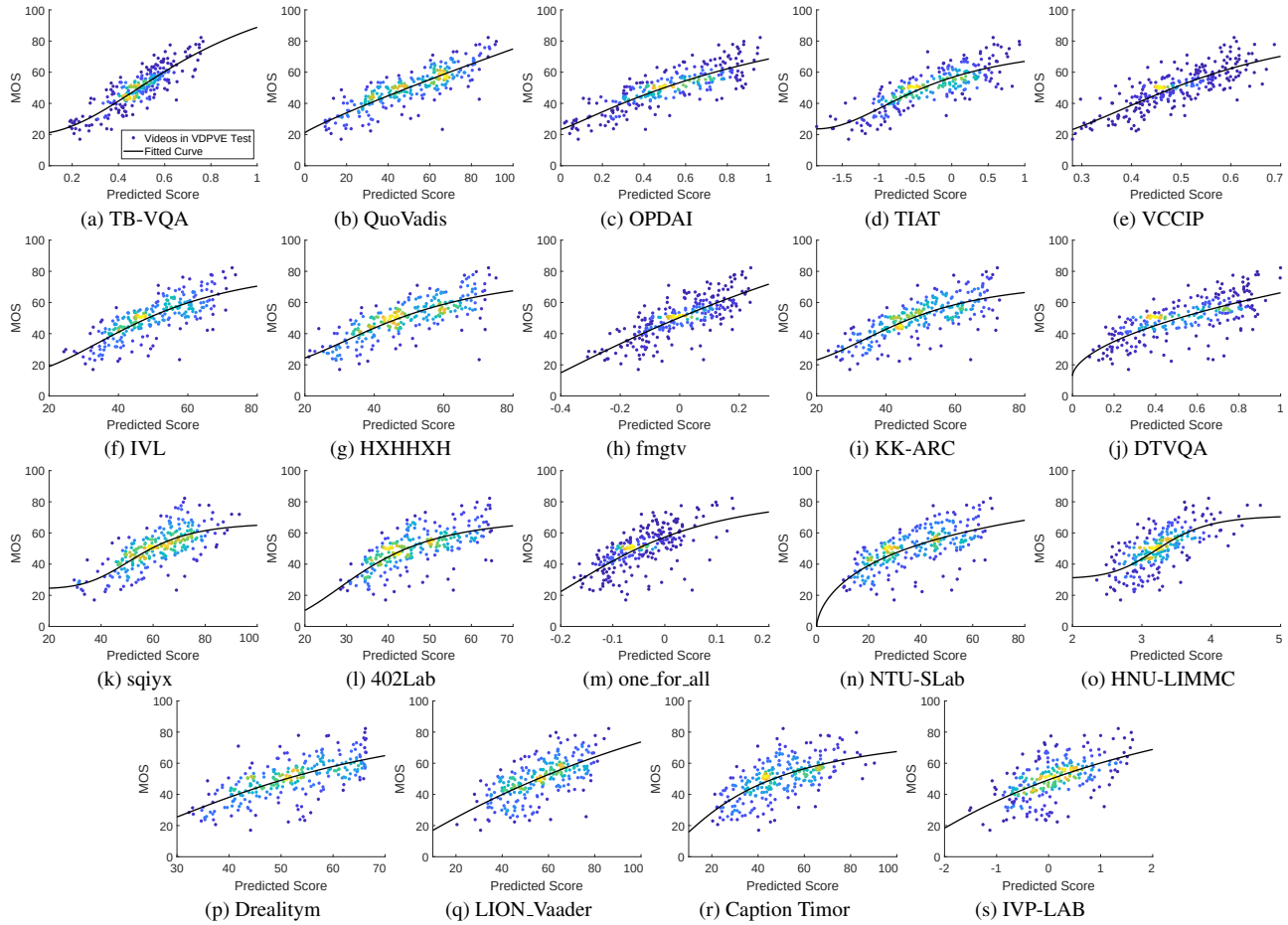


Figure 1. Scatter plots of the predicted scores vs. MOSs. The curves are obtained by a four-order polynomial nonlinear fitting

the VDPVE training set and further train the model on the dataset they built, and then fine-tune it on the VDPVE training set. Particularly, multiple temporal frames are randomly sampled from each segment of videos, which efficiently increases the volume of the training dataset. As a comparison, only one frame per segment with fixed sampling order is utilized as a training sample in [64].

There are 29.77 million trainable parameters in the model. In the training phase, the input frames are resized to 320×320 and randomly cropped with a resolution of 256×256. Batch size is set as 16 and Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$ is utilized for optimization. The learning rate is initialized as 10^{-5} and decayed by $\gamma = 0.95$ every two epochs. During testing, the input frames are resized to 320×320, and the "torchvision.transforms.TenCrop" function is used to crop 10 image patches with a resolution of 256×256, which are located at the four corners and the center, respectively, as well as the horizontally flipped version of the previous crops.

5.2. QuoVadis

Team QuoVadis wins second place in the challenge. They propose a dual-branch VQA network for enhanced videos [97]. As is shown in Figure 2, the overall architecture consists of two parts: the image-based network and the video-based network. The image-based network receives single images as input and generates quality prediction in a global view, while the video-based one obtains shuffled fragments and predicts prediction focusing on textural distortions.

Specially, the architecture used for the image-based network is ConvNext-Tiny [47], following a regression head. To analyze the overall quality of the video, they uniformly sample two frames per second and input them into the network. For each frame, they scale the shorter side to 512 and maintain the same scaling ratio for the longer side. They then crop a 320×320 region from the center as the network input to obtain the global quality information of the image. They further attach a patch-weighted quality prediction head as [83]. And the final prediction is generated

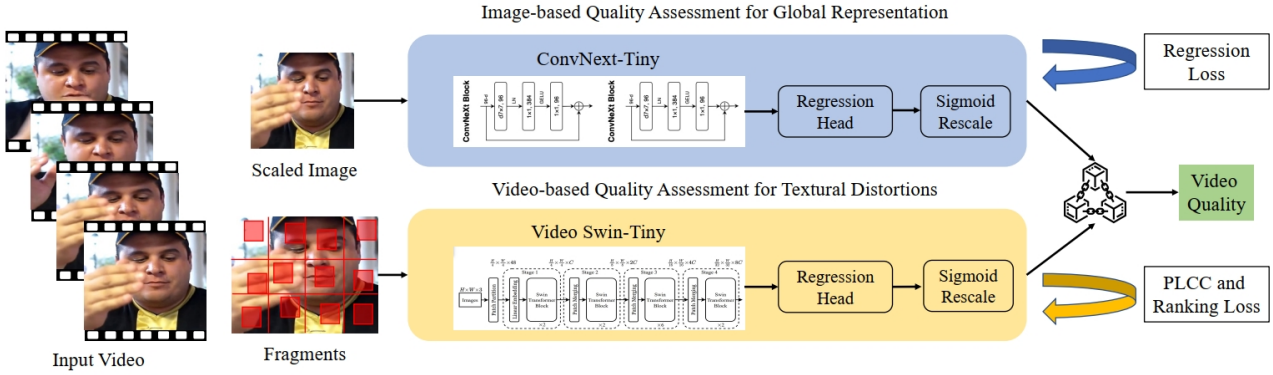


Figure 2. The overview of QuoVadis team proposed dual-branch VQA network for enhanced videos.

by the multiplication of each patch's score and weight. The outputs generated by all frames are averaged as the whole quality of the video. During training, the network is optimized using a smooth $\mathcal{L}_1$ regression loss. Given a video X and corresponding sampled frames $\{x_1, x_2, ..., x_n\}$, the optimization objective can be written as:

$$min\ \mathcal{L}_{reg} = min\ \frac{1}{n}\sum_{i=1}^{n} ||\mathcal{F}(x_i) - y||_1,$$

where $\mathcal{F}(\cdot)$ is the mapping function of the network, and y is the labeled MOS for the video.

Unlike the above image-based network, the video-based network receives video clips as input. Following [79], each clip contains 32 frames sampled uniformly. To preserve the original video quality and obtain local texture information which benefits quality assessment, they utilize the sampling strategy of fragments used in FastVQA. The fragments are obtained through uniform grid mini-patch sampling. This method vastly reduces the computational cost by 97.6% compared with computing attention on the whole resolution. The image-based network structure described above can perceive global semantic information. To complement this, they redefine the form of the fragments and further randomly shuffle the positions of the mini-patches in space, allowing the network to pay more attention to low-quality texture information (such as noise, blur, block effects, etc.) and reduce its focus on higher-level semantics. Then the shuffled fragments are sent into an attention network of Video Swin-Tiny [48]. During training, specifically, a PLCC-induced loss and a ranking-based loss are utilized. Assume there are m videos in the training batch. Given the predicted quality scores $\{y'_1, y'_2, ..., y'_m\}$ and the MOS values $\{y_1, y_2, ..., y_m\}$, the PLCC-induced loss is defined as:

$$\mathcal{L}_{plcc} = (1 - \frac{\sum_{i=1}^{m}(y'_i - a')(y_i - a)}{\sqrt{\sum_{i=1}^{m}(y'_i - a')^2 \sum_{i=1}^{m}(y_i - a)^2}})/2,$$

where a' and a are the mean values of m predicted quality scores and MOSs respectively. And the ranking-based loss can be denoted as:

$$\mathcal{L}_{rank} = \frac{1}{m^2}\sum_{i=1}^{m}\sum_{i=1}^{m} max(0, |y_i - y_j| - e(y_i, y_j)\cdot(y'_i - y'_j)),$$

where $e(y_i, y_j)$ is 1 if $y_i \geq y_j$, else is -1 if $y_i < y_j$. And the optimization objective can be written as:

$$min\ \mathcal{L}_{plcc} + \beta \cdot \mathcal{L}_{rank},$$

and β is the coefficient for balancing. In practice, it is set to 0.3.

There are 55 million parameters in the entire model. In ConvNext-Tiny, the batch size is set to 32, the epoch is set to 30, the learning rate is initialized as 10^{-4}, and AdamW with 10^{-2} weight decay is utilized for optimization. In Video Swin-Tiny, the batch size is set to 16, the epoch is set to 30, the learning rate is initialized as 10^{-3}, and AdamW with 10^{-2} weight decay is utilized for optimization. During the testing phase, the video is analyzed using the dual-branch network structure based on both images and videos, and the prediction results from both branches are averaged to obtain the final prediction quality.

5.3. OPDAI

Team OPDAI wins third place in the challenge. They apply image quality assessment into VQA. Specifically, they combine one VQA method and four image quality assessment methods for this competition. The VQA is based on DOVER [81]. The four image quality assessment models are main model SwinTransfromer [46], main model ConvnextV2 [77], using SwinTransformer as backbone extracting feature and using stacked transformer to regress quality score, and main model CDCNN [93], respectively. Then, they bagging 5 models to obtain the final results. The mean

absolute error (MAE), mean square error (MSE), norm-in-norm loss and Kullback-Leibler (KL) divergence loss are used for training. It is worth noting that not only the provided NTIRE 2023 dataset, but also part of Youtube-UGC [75] dataset and PIPAL [27] are used for pretraining the model. They used a cosine annealing learning rate descent method with warming-up. Minibatch size is set to 60, and the learning rate is initialized as 4×10^{-5}. They utilize AdamW optimizer setting $\beta_1 = 0.9, \beta_2 = 0.999$.

5.4. TIAT

TIAT proposes a deep learning based VQA model. In order to try their best to make use of the information contained in the target video, the spatial information and temporal information in each video are extracted and fused through segmented processing of the target video. The model is based on SimpleVQA [64] and they modify the spatial feature extraction module. Specifically, they utilize Swin-b [46] network pre-trained on ImageNet [12] and slowfast r50 [14] network pre-trained on Kinetics [7] as the feature extraction model.

There are totally 87.30 million parameters in the model. In the training phase, minibatch size is set to 6, the training epoch is set to 40, and the learning rate is initialized as 10^{-5}. They utilize Adam optimizer setting $\beta_1 = 0.9, \beta_2 = 0.999$.

5.5. VCCIP

Team VCCIP proposes a VQA network model for channel fusion, which integrates information from different channels of three consecutive frames in time. This network takes into account both spatial and temporal information, which enables better performance. Additionally, they add a subtask to identify the category of enhancement method to optimize and train the network model more effectively.

The architecture of CF-VQA is illustrated in Figure 3, which comprises of a Swin Transformer [46] backbone and a quality score regression module. The network structure is kept simple. To leverage the powerful learning ability of the Transformer structure, this method tends to utilize a pretrained Swin Transformer base as the backbone. After extracting quality perception features through the effective Swin Transformer backbone, a regression model is incorporated to map these features to the quality score. Firstly, the global average feature pooling (GAP) is applied to generate a feature vector with a dimension of $P \times 2$, where P represents the number of final feature maps. Then, two fully connected (FC) layers, consisting of 512 neurons and 2 neurons, respectively, are utilized to map the feature vector to the predicted quality score and enhancement method category. Finally, CF-VQA can be trained on VDPVE [15] using an end-to-end training method with $\mathcal{L}_1$ loss function,

$$\mathcal{L}_1 = \frac{1}{N} \sum_{i=1}^{N} ||q_{score} - q_{label}||_1,$$

where q_{score} and q_{label} denote the predicted score and MOS of the $i\text{-}th$ training patch, and N represents the total number of training patches.

There are 87.41 million parameters in the model. Swin Transformer Base Network pre-trained using ImageNet [12] is utilized as the feature extraction sub-model. Batch size is set to 4, the learning rate is set to 10^{-5}, and AdamW optimizer is adopted with a weight decay of 5×10^{-4}. In addition, they use the cosine decay learning rate with the minimum learning rate of 10^{-7}, and use linear preheating in first 2 epochs with start learning rate 5×10^{-7}. When training on VDPVE, they randomly sample and horizontally flipping with size 384×384 pixels from each training image for augmentation. In the test phase, one hundred video clips with 384×384 pixels are randomly cropped from each video, and the final quality score of a video is the average score of all clips.

5.6. IVL

Team IVL proposes a VQA method [1] inspired by DOVER [81] and the NR-VQA model introduced in [2]. As depicted in Figure 4, it consists of three components: the feature extractor module includes the technical quality and aesthetic encoders and the quality-related attribute encoder; the feature combination module provides a temporal aggregation part for the frame-level features of the quality-related attribute encoder and reduces the feature vectors obtained by all the encoders to a fixed size; the quality prediction module exploits a support vector regression (SVR) machine for mapping the feature vector into the video quality score.

The feature extraction module contains DOVER to model features capturing information about distortion perception (technical quality) and preferences (aesthetics), and a quality-related attribute encoder to model quality features capturing various quality attributes, such as brightness, contrast and sharpness. The DOVER architecture is modified so that its outputs are feature vectors rather than quality scores. The aesthetic encoder consists of a tiny inflated-ConvNext [47] as backbone while the technical quality encoder exploits a tiny Video Swin Transformer [48]. 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 480p resolution to increase efficiency. The technical quality encoder divides the video into two parts, and selects 32 frames in the first part and 32 frames in the last one with a stride of 4 to increase video coverage. Here five-crop video fragments are used, *i.e.*, each frame is cropped at

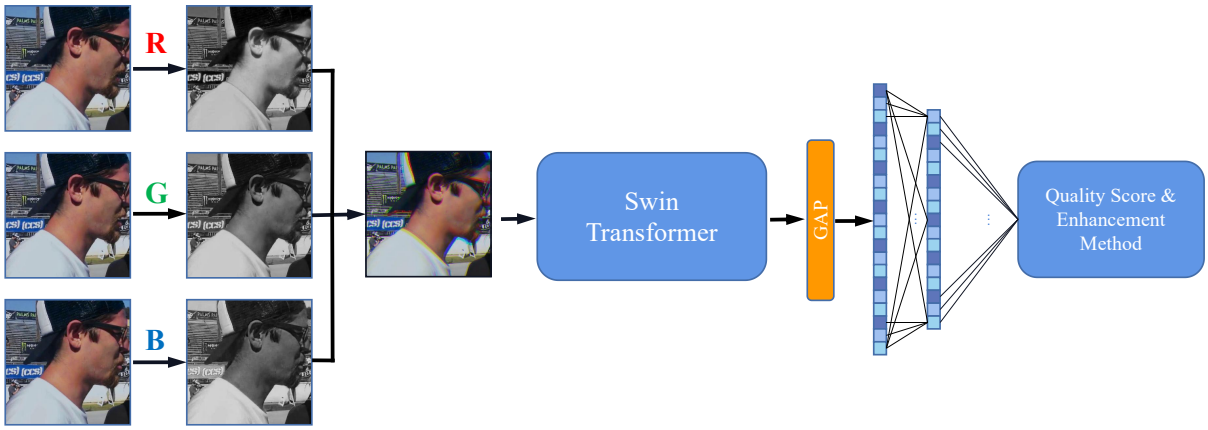


Figure 3. The framework design of VCCIP team's method.

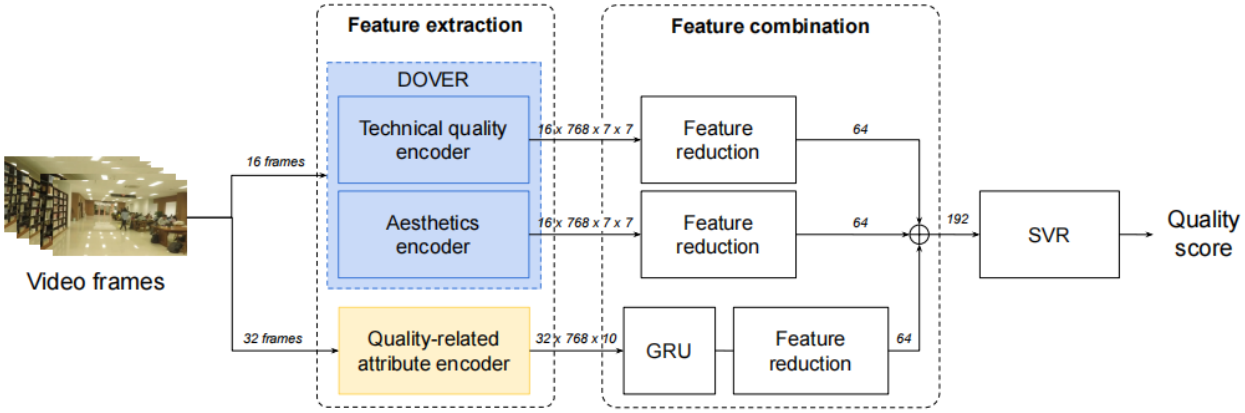


Figure 4. The framework design of IVL team's method.

the four corners and at the central, and fragments are generated from these crops. Note that each crop has a 480p resolution, thus the fragments can cover about 80% of the entire frame spatial dimension. Moreover, downscaling is performed using soft pooling [63] to better preserve video distortions and avoid masking effects. The fragments are processed independently from the others, and the final features are obtained by averaging the features extracted from each crop. The quality-related attribute encoder is similar to the one proposed in [2], but the MobileNet-v2 [56] backbone is replaced with EfficientNet-v2 [66] because of its higher capability in capturing relevant quality information. As in [2], the model is trained using the images from the CID [71] and the SPAQ [13] datasets with the aim of predicting the scores related to quality attributes, *i.e.*, sharpness, groundness, lightness, saturation, brightness, colorfulness, noisiness, contrast, and MOS. In order to obtain the quality features for a video, 32 frames (the same used by the aesthetic encoder) are selected and processed by the network, and the features obtained before the FC layers of each branch are used. The resulting feature vector is obtained by concatenating the feature vectors produced by each branch. The feature combination module reduces the dimensions of the extracted features and prepares them to be used for quality score prediction. The quality-related attribute encoder extracts quality features frame-by-frame. Therefore, a GRU module is used to map frame-level quality features into a single vector capturing temporal dependency among them, and its dimension is later reduced by a FC layer. Then, these features are concatenated with the aesthetic features and the technical quality features related to the first half of the video, and processed by an additional FC layer. The same happens for the features obtained considering the technical features related to the second half of the video. The two outputs are later concatenated. Finally, in the quality prediction module the obtained feature vec-

tor is mapped into the final video quality score through an SVR.

There are about 78 million parameters in the DOVER architecture and 20 million parameters in the quality-related attribute encoder. Pretrained-models are utilized. Specifically, Video Swin Transformer is pretrained on LSVQ [86], inflated-ConvNext is pretrained on the AVA dataset [52], and EfficientNet-v2 is pretrained on the combination of CID [71] and SPAQ [13]. They not only utilize the provided VDPVE [15] dataset to train the whole model, but also take the CID and the SPAQ datasets as supplements. In the training phase, andom video fragments are used at training time, while five-crop video fragments are used at inference time. Video fragments are generated as described in [81]. The batch size is set to 5, the learning rate is set to 10^{-4} for the technical quality encoder and 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 and SPAQ 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 10^{-4} and later decreased by a factor of 10 after 5K iterations. Random horizontal flip is used as data augmentation. Both the encoders are trained using PLCC loss and rank loss, using the ground-truth scores as target. The rank loss has a weight of 0.3 in the total loss for training stability. The SVR for the final score prediction is trained on the VDPVE training set. The required hyperparameters, *i.e.*, $\gamma = 12.20$ and $C = 364.83$, are selected via Bayesian optimization using Leave-One-Out cross-validation. In the testing phase, they average the quality scores obtained for the original video, and its horizontally flipped version.

5.7. HXHHXH

Team HXHHXH proposes a new pre-training method. To extract better spatial distortion features, they use the Live in the wild [19] dataset to pre train the feature extraction network. Additionally, considering combination of temporal and spatial features can better achieve accurate quality assessment, they design a new temporal pre-training strategy. Concretely, for a video with length L, they sample new video segments of $L/2, L/4$ and $L/8$, and use these different video segments for pre-training to achieve better results.

The whole model architecture is based on FastVQA [79], along with a VGG network pre-trained using ImageNet as the feature extraction sub-model. The number of parameters is around 3 million. For training details, they use Adam optimizer setting $\beta_1 = 0.9, \beta_2 = 0.999$. They set minibatch size as 12, and learning rate as 0.0002. The training process takes around 0.5 hour on Nvidia GeForce RTX 3090.

5.8. fmgtv

The method proposed by team fmgtv can be divided into three parts: image quality assessment (IQA) module, VQA module, and fusion module. For IQA module, they extract each frame of the video and use a picture classification network for regression training, while the VQA module extracts N frames at each interval and uses a video classification network for regression training. Specifically, they use ResNet-101, Convnext pre-trained using ImageNet as the feature extraction sub-model, Swin3D and Xclip pre-trained using LSVQ [86] dataset as backbone. Finally, they directly fuse the IQA and VQA results as their final resutls. The fusion strategy brings 3% improvement (0.74 to 0.77).

During training, they use LSVQ [86] training dataset as extra data in addition to provided NTIRE 2023 training data. They use cosine learning rate initialized as 0.01, and set batch size as 32. The training process takes around twelve hours for 30 epochs on Nvidia GeForce RTX 3090.

5.9. KK-ARC

Team KK-ARC proposes a method [25] by stacking ensemble on three VQA models: FastVQA-B [79], FastVQA-M [79], and FasterVQA [80]. All three models use Swin-Transformer [48] as backbone, and are pre-trained on LSVQ [86] dataset. The provided NTIRE 2023 training dataset is used for finetuning on those models separately. XGBoost is then applied for stacking ensemble.

5.10. DTVQA

Team MTVQA proposes a self-attention based perception VQA method. The overview of their framework is shown in 5. At the beginning, they divide VQA problem into authentic/aesthetic video quality and synthetic distortion VQA. After extracting frames from both parts, they use ResNet which is pre-trained using ImageNet as feature extraction sub-model. Different stages of feature map are extracted and concatenated together. They then implement dimensionality reduction by averaging the feature map. At the last step, a self-attention module is used to model the time dimension video quality.

Furthermore, they find that by adding more public available VQA datasets during training can overcome overfitting. Specially, they combine two multi-dataset training strategy: 1) method proposed by MDTVSFA [33], which described a dataset-specific alignment method for training different datasets; 2) a multi-stage training strategy. After training the model on dataset A, they load the trained checkpoint and finetune the model on dataset B, and so on. The additional public available VQA datasets that are used are KoNVid-1k [23], YouTube-UGC [75], and MSU CVQA dataset [4].

The number of parameters for their model achieves approximately 1.71 million. During training, the model is

trained for 50 epochs using Adam optimizer with an initial learning rate 10^{-4}. The batch size is set by 256 for each dataset. Whole training and testing processes are conducted on 4 Nvidia Tesla v100. It takes six hours for training, and 16ms per image during testing.

It is worth mentioning that, to further overcome overfitting, they conduct ensemble by averaging the outputs of the proposed method and DOVER [81] to get the final result.

5.11. sqiyx

Team members from sqiyx indicate that the provided data may not be enough to train a VQA model. As a result, they turn to an IQA model, MANIQA [83]. They introduce the fragment technology in FastVQA [79] to their method, which can keep the resolution information of the image from being lost, and can be regarded as a kind of data augmentation. To further augment data, they utilize cutMix to fuse two random images from the same video. Last but no least, they use a more large model than MANIQA [83] to increase the model capability. The feature extraction submodel is a Vit Large Patch16 Network pre-trained using ImageNet.

During training, they use Adam optimizer by setting $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Minibatch size is set as 8, and learning rate is initialized as 10^{-5}. They use cosine scheduler to update learning rate with parameters Tmax and etamin set to 50 and 0. It takes approximately one day for training. During testing, they use pyav to extract key frames and calculate the results of each frame, and finally take the average to get the final score of the video. All experiments are conducted on one Nvidia Tesla v100.

5.12. 402Lab

The method proposed by team 402Lab includes four parts, the spatial feature extraction module, the motion feature extraction module, the spatial-motion features fusion module, and the quality regression module. This method takes the first frame per second as input, and does not require any cropping operations, which results in low complexity and integral image information. ResNetv2-50 [22] pre-trained on ImageNet is utilized as the feature extraction backbone network. Contestants propose a two-branch feature fusion network strategy, which effectively combines the advantages of CNN network and Transformer network,and fully integrates local and global information. In addition, they propose a patch attention module to make quality assessment more focused on effective information. After motion features and spatial features are extracted, we fuse the two features using the spatial-motion feature fusion module, which is used to compensate for temporal-related distortions that cannot be modeled by spatial features.

The whole training is divided into two procedures. The model is first pre-trained on KonIQ-10K [24] and later finetuned on NTIRE 2023 training dataset. During pre-training, the initial learning rate of the backbone network is 10^{-5} and the rest is 10^{-4}. They also randomly horizontally flipped images with a given probability of 0.5 during pre-training. During finetuning process, they use the SlowFast R50 [14] as the motion feature extraction model for the whole experiments. The weights of the SlowFast R50 [14] are fixed by training on the Kinetics 400 [29] dataset. The initial learning rate of the network is 10^{-5}, and is reduced by 10 after 80 epochs. AdamW optimizer is used in both pre-training and finetuning processes by setting $\beta_1 = 0.9, \beta_2 = 0.999$. And minibatch size is set as 8. Furthermore, to maintain the image ratio, they first resize the image to 640×360, after which we use the same preprocessing method as in [87] to fill the image to a resolution of 640×384. The same operation is conducted in testing process.

5.13. one_for_all

This team proposes a VQA method based on the multi-clips ensemble, which contains two steps: data filtering and partitioning based on video embedding clustering; and quality content decoupled regression headers.

Their network contains about 55 million parameters. They only use the training set of the VDPVE to train their network. No additional data has been used. They use the DOVER backbone pre-trained using LSVQ as the feature extraction sub-model. For optimization,they use the AdamW optimizer by setting β_1=0.9, β_2=0.999. They set the minibatch size as 8. The learning rate is initialized as 10^{-3}.

5.14. NTU-SLab

Team NTU-SLab proposes a network based on the DOVER, which consists of an aesthetic branch and a technical branch. Moreover, they ensemble the DOVER result with raw feature tuning from CLIP-RN50 visual backbone.

Their network contains around 75 million parameters. They only use the training set of the VDPVE to train their network. No additional data has been used. For optimization, they use Adam optimizer by setting β_1=0.9, β_2=0.999. They set minibatch size as 8 and train for 30 epochs. The learning rate is initialized as 10^{-3} and kept unchanged during training. They ensemble the DOVER and CLIP-RN50 results with 2:1 ratio.

5.15. HNU-LIMMC

In order to improve the sensitivity of the model to enhanced video perception, they propose a novel contrastive learning method for VQA based on the idea of self-supervision to improve the performance of the model. At the same time, they introduce a video degradation space. Specifically, they believe that different frames of the same

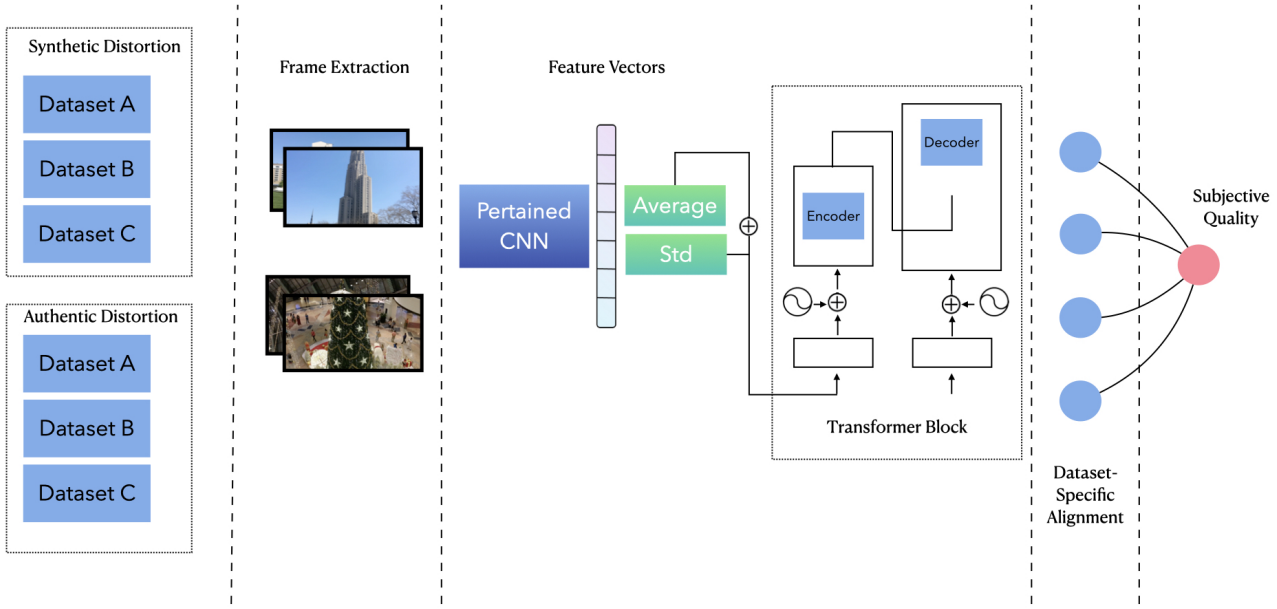


Figure 5. The framework design of DTVQA team's method.

video should be similar and can effectively represent the video. The quality of different degradation methods of the same frame is not similar.

Through this framework, an end-to-end learning method can be effectively established. The degraded video information in the VQA dataset is used to simulate the learning method in the enhanced scene, which is conducive to the generalization ability of the model and improves the effect of the model on the non-natural VQA dataset.

Their network contains about 2.472 million parameters. They only use the training set of the VDPVE to train their network. No additional data has been used. They choose SimpleVQA as the backbone of their model, which is pretrained on LSVQ and fine-tuned on VDVPE. For optimization, they use Adam optimizer. They set the minibatch size as 6. The learning rate is initialized as 10^{-5} and the weight-decay is 10^{-8}.

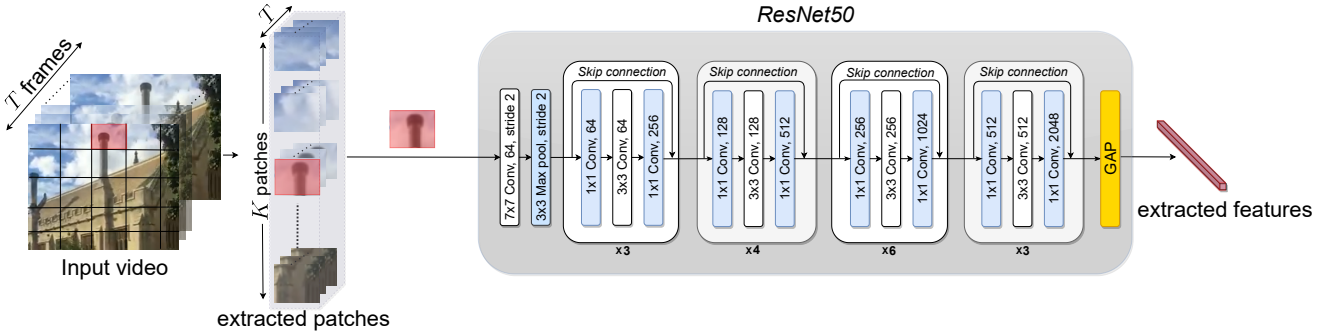
5.16. Drealitym

They propose a NR VQA method for enhanced videos based on the framework of Adaptive Token-Selection ViT (ATSViT) that the process of energy competition between visual information from the psychological perspective to predict video quality scores. They propose a block-level sampling strategy, called timing block sampling (TBS), that takes into account the uneven distribution of local quality distortions focused on by the human eye in the original sequence, increasing the information density of the sampled frame set and reducing the loss possibility of important spatial features through HVS based [43, 67] fine-grained sampling. They construct transformer-based Stage-wise adaptive Screening Network (SSNet) based on based on the filter theory of attention [78], dividing the process of visual information processing into four stages where efficient energy distribution strategies are used, exploiting the attention-based bottlenecks of different sizes, to select the features of tokens that advance to the next stage.

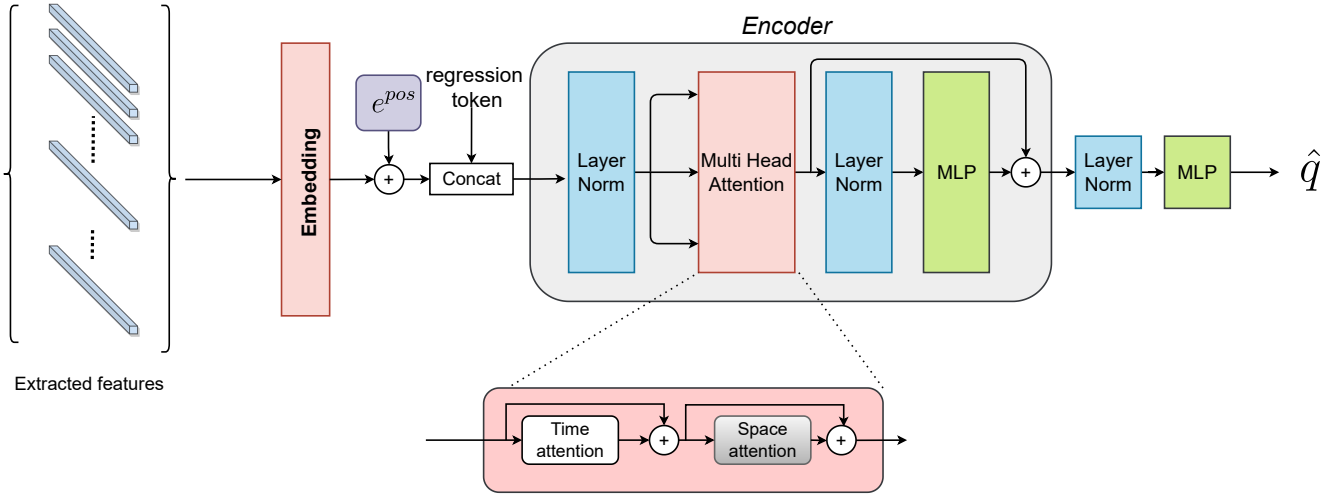
They only use the training set of the VDPVE to train their network. No additional data has been used. They use swin-B Network pretrained on the Kinetics-400 dataset to initialize the backbone in SSNet. For optimization, they use the AdamW optimizer by setting $\beta_1 = 0.9$, $\beta_2 = 0.999$. They set minibatch size as 5. The learning rate is initialized as 5×10^{-6} and use the custom warmup cosine step decay to updates the learning rate. The Weight decay is set as 1.5×10^{-3}.

5.17. LION_Vaader

They adopt the well-known ResNet-50 model and exploit it to perform transfer learning via a fine-tuning technique using the weights of the ImageNet dataset. First, they perform a temporal sub-sampling of the video by select-



(a) Illustration of the features extraction process



(b) Illustration of the feature pooling process

Figure 6. The framework design of LION_Vaader team's method.

ing one frame per second. Then, from each of the selected frames, a set of patches is extracted in a sliding window fashion so that their dimensions match the standard architecture's input shape.Second, each patch is fed to the CNN backbone for feature extraction and fits into the spatio-temporal pooling module in a time-distributed fashion. The extracted features are then fed to the spatio-temporal pooling module. The use of this pooling type is motivated by the fact that the visibility of an artifact depends highly on its location and its neighboring regions in the current frame and the adjacent frames, resulting in the so-called masking phenomenon. Thus, the overall quality of a given video is affected by both, spatial artifacts that occur in some regions of the frame, as well as temporal artifacts that affect a range of sequential frames.

Most of the pooling techniques are effective at capturing short-range patterns within local spatio-temporal regions, whereas they can only model space-time dependencies of at most a handful of seconds, not video whole. To address that, they use a spatio-temporal transformer namely TimeSformer as pooling architecture, which exploits space-time attention. Thus, the feature pooling module captures short-term dependencies between neighboring patches as well as long-range correlations between distant patches. In addition, this pooling can analyze the video over much longer time spans.

Their network contains 25.6 million parameters in the CNN resnet50 model and 122.13 million parameters in the TimeSformer architecture (vary according to the input's number of frames and patches). They use the All Combined dataset that the authors in [68] proposed by merging the KoNViD-1k, LIVE-VQC and YouTube-UGC datasets. They use the ResNet50 pre-trained on ImageNet as the feature extraction sub-model. For optimization, they use the Adam optimizer with default parameters. They set mini-batch size as 1 due to the diversity of the number of frames and patches that can be extracted per video. The learning rate is initialized as 10^{-3}, and a reduce learning rate on plateau call back is used with patience parameter set to 5. Features are only extracted once, and used directly to train the transformer architecture.

5.18. Caption Timor

This team utilized SimpleVQA to directly extract spatial and mobile features, while applying random rotation, flipping, and cutting as data augmentation techniques. They optimized the model by adjusting parameters and taking the average value of five models obtained from five-fold cross-validation training. The loss function used was MSE loss.

Their network consists of 2.6 million parameters and was trained solely on the VDPVE training set without additional data. They employed a ResNet50 pre-trained on ImageNet as the feature extraction sub-model. The Adam optimizer was used for optimization, with β_1 set to 0.9 and β_2 set to 0.999. The minibatch size was set at 8, with a learning rate initialized at 10^{-5} and halved every 5×10^4 minibatch updates. During the testing phase, the batch size was set to 1.

5.19. IVP-LAB

They introduce a novel method to acquire frame level deep features for assessing the quality of videos. To accomplish this, they focus on the deep feature maps correlations of a pre-trained network, or more specifically, their similarity as a helpful tool for assessing video quality. The covariance matrix *i.e.* the Gram matrix, which depicts the correlation between all feature maps of a specific mid-layer, can be stated as deep feature relationships. The structural details of frames' appearance are reflected in these relations and significantly correlate with the perceived quality of a given video. The extracted feature maps relations in different granularities can effectively illustrate the influence of various distortions. Every feature map reflects a different structural detail of the source image. It is shown in [17] that almost flawless reconstruction is possible from the network's lower layers whereas the detailed pixel information is insufficiently maintained in the network's upper layers. In this case, each layer's output of a convolutional neural network can be represented by a collection of feature maps that show the input pixels' structural data. In the proposed method, the extracted Gram Matrix of mid-level convolutional layer is employed as the frame level feature. The proposed method exploits the correlation between the deep feature maps derived from each network's layers to assess the video's quality.

Their network contains 21.7855 million parameters. They only use the training set of the VDPVE to train their network. No additional data has been used. They use the inception-v3 pre-trained on ImageNet as the feature extraction sub-model. The only trainable model is a linear SVR model that is trained using features extracted by a pretrained inception-v3 model. They set epsilon value as 0.3. The first test phase involves extracting spatial information from video frames and merging it into a feature vector. Then this feature vector was given to an SVR model to predict the quality score.

Acknowledgments

We thank Peng Cheng Laboratory for sponsoring this NTIRE 2023 challenge and the NTIRE 2023 sponsors: Sony Interactive Entertainment, Meta Reality Labs, ModelScope, ETH Zürich (Computer Vision Lab) and University of Würzburg (Computer Vision Lab).

A. NTIRE 2023 Organizers

Title:
NTIRE 2023 Quality Assessment of Video Enhancement Challenge
Members:

Xiaohong Liu[1] *(xiaohongliu@sjtu.edu.cn)*, Xiongkuo Min[1], Wei Sun[1], Yulun Zhang[2], Kai Zhang[2], Radu Timofte[2,3], Guangtao Zhai[1], Yixuan Gao[1], Yuqin Cao[1], Tengchuan Kou[1], Yunlong Dong[1], Ziheng Jia[1]
Affiliations:
[1] Shanghai Jiao Tong University, China
[2] ETH Zürich, Switzerland
[3] University of Würzburg, Germany

B. Teams and Affiliations

TB-VQA

Title:
Video Quality Assessment Based on Swin Transformer with Spatio-Temporal Feature Fusion and Data Augmentation
Members:

Yilin Li[1] *(gustav.lyl@alibaba-inc.com)*, Wei Wu[1], Shuming Hu[1], Sibin Deng[1], Pengxiang Xiao[1†], Ying Chen[1], Kai Li[1]
Affiliations:
[1] Department of Tao Technology, Alibaba Group

QuoVadis

Title:
A Dual-branch Network for Enhanced Video Quality Assessment
Members:

Kai Zhao[1] *(zhaokai05@kuaishou.com)*, Kun Yuan[1], Ming Sun[1]
Affiliations:
[1] Kuaishou Technology

[†]Pengxiang Xiao is also with Vrobotit Lab, Beijing University of Posts and Telecommunications, and the work is primarily done during an internship at Alibaba Group.

OPDAI

Title:
Apply Image Quality Assessment into Video Quality Assessment
Members:
Heng Cong[1] *(congheng@corp.netease.com)*, Hao Wang[1], Lingzhi Fu[1], Yusheng Zhang[1], Rongyu Zhang[1]
Affiliations:
[1] Interactive Entertainment Group of Netease Inc, Guangzhou, China

TIAT

Title:
A Deep Learning based Video Quality Assessment Model
Members:
Hang Shi[1] *(hang.shi@transsion.com)*, Qihang Xu[1], Longan Xiao[1]
Affiliations:
[1] Transsion

VCCIP

Title:
Channel Fusion for Video Quality Assessment
Members:
Zhiliang Ma[1] *(mzl@mail.hfut.edu.cn)*
Affiliations:
[1] Hefei University of Technology

IVL

Title:
Video Quality Assessment Guided by Aesthetics and Technical Quality Attributes
Members:
Mirko Agarla[1] *(m.agarla@campus.unimib.it)*, Luigi Celona[1], Claudio Rota[1], Raimondo Schettini[1]
Affiliations:
[1] Department of Informatics Systems and Communication, University of Milano - Bicocca

HXHHXH

Title:
A Multi-interval Sampling Strategy Pre-training for Video Quality Assessment
Members:
Zhiwei Huang[1] *(huangzhiwei10@xiaomi.com)*, Ya'nan Li[1], Xiaotao Wang[1], Lei Lei[1]
Affiliations:
[1] Xiaomi Inc., China

fmgtv

Title:
Integration of IQA and VQA
Members:
Hongye Liu[1] *(liuhongye1998@163.com)*, Wei Hong[2]
Affiliations:
[1] China Ji Liang University
[2] FreeTech

KK-ARC

Title:
Stacking Ensemble with FastVQA-B, FastVQA-M, and FasterVQA
Members:
Ironhead Chuang[1] *(ironheadchuang@kkcompany.com)*, Allen Lin[1], Drake Guan[1], Iris Chen[1], Kae Lou[1], Willy Huang[1], Yachun Tasi[1], Yvonne Kao[1]
Affiliations:
[1] Advanced Research Center, KKCompany, Taiwan

DTVQA

Title:
Self-Attention based Perception Video Quality Assessment Method
Members:
Haotian Fan[1] *(fanhaotian@bytedance.com)*, Fangyuan Kong[1]
Affiliations:
[1] ByteDance

sqiyx

Title:
No Title
Members:
Shiqi Zhou[1] *(408172566@qq.com)*, Hao Liu[1]
Affiliations:
[1] MGTV

402Lab

Title:
Triple-Branch Feature Fusion Network, Spatial Feature Extraction Subnetwork, Spatial-Motion Features Fusion Mdule
Members:
Yu Lai[1] *(2672339375@qq.com)*, Shanshan Chen[1]
Affiliations:
[1] Fuzhou University

one_for_all

Title:
Video Quality Assessment based on Multi-Clips Ensemble
Members:
Wenqi Wang[1]*(wwenki1992@163.com)*,
Affiliations:
[1] Shopee Information Technology Co., Ltd.

NTU-SLab

Title:
DOVER-CLIP-RN50
Members:
Haoning Wu[1]*(haoning001@e.ntu.edu.sg)*, Chaofeng Chen[1]
Affiliations:
[1] S-Lab, Nanyang Technological University

HNU-LIMMC

Title:
Exploring Comparative Learning-Inspired Strategies for Video Quality Evaluation for Enhance Video
Members:
Chunzheng Zhu[1] *(1724735214@qq.com)*, Zekun Guo[1]
Affiliations:
[1] Hunan University

Drealitym

Title:
Video Transformer based Video Quality Assessment
Members:
Shiling Zhao[1] *(yiyiaiou@163.com)*, Haibing Yin[1], Hongkui Wang[1]
Affiliations:
[1] Hangzhou Dianzi University

LION_Vaader

Title:
Quality Assessment for Video Enhancement using Joint Space-Time Attention
Members:
Hanene Brachemi Meftah[1]*(hanene.brachemi@insa-rennes.fr)*, Sid Ahmed Fezza[2], Wassim Hamidouche[1], Olivier Déforges[1]
Affiliations:
[1] INSA Rennes, CNRS, IETR - UMR 6164, Rennes, France
[2] National Higher School of Telecommunications and ICT, Oran, Algeria

Caption Timor

Title:
Five Fold and Data Augmentation.
Members:
Tengfei Shi[1] *(tengfeishibh@163.com)*
Affiliations:
[1] BUAA

IVP-LAB

Title:
Feature Maps Correlation-based Video Quality Assessment
Members:
Azadeh Mansouri[1]*(a_mansouri@khu.ac.ir)*, Hossein Motamednia[2], Amir Hossein Bakhtiari[1], Ahmad Mahmoudi Aznaveh[3]
Affiliations:
[1] Department of Electrical and Computer Engineering Faculty of Engineering Kharazmi University, Tehran, Iran
[2] High Performance Computing Laboratory School of Computer Science Institute for Research in Fundamental Sciences Tehran, Iran
[3] Cyberspace Research Institute Shahid Beheshti University, Tehran, Iran

References

[1] Mirko Agarla, Claudio Celona, Luigi ad Rota, and Raimondo Schettini. Quality assessment of enhanced videos guided by aesthetics and technical quality attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 7

[2] Mirko Agarla, Luigi Celona, and Raimondo Schettini. An efficient method for no-reference video quality assessment. *Journal of Imaging*, 7(3):55, 2021. 7, 8

[3] Codruta O Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, Radu Timofte, et al. NTIRE 2023 challenge on nonhomogeneous dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[4] Anastasia Antsiferova, Sergey Lavrushkin, Maksim Smirnov, Aleksandr Gushchin, Dmitriy Vatolin, and Dmitriy Kulikov. Video compression dataset and benchmark of learning-based video-quality metrics. *Advances in Neural Information Processing Systems*, 35:13814–13825, 2022. 9

[5] Mingdeng Cao, Chong Mou, Fanghua Yu, Xintao Wang, Yinqiang Zheng, Jian Zhang, Chao Dong, Ying Shan, Gen Li, Radu Timofte, et al. NTIRE 2023 challenge on 360° omnidirectional image and video super-resolution: Datasets, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[6] Yuqin Cao, Xiongkuo Min, Wei Sun, and Guangtao Zhai. Attention-guided neural networks for full-reference and no-reference audio-visual quality assessment. *IEEE Transactions on Image Processing*, 2023. 2

[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 7

[8] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5972–5981, 2022. 1, 3

[9] Jinsoo Choi and In So Kweon. Deep iterative frame interpolation for full-frame video stabilization. *ACM Transactions on Graphics*, 39(1):1–9, 2020. 3

[10] Marcos V Conde, Manuel Kolmet, Tim Seizinger, Thomas E. Bishop, Radu Timofte, et al. Lens-to-lens bokeh effect transformation. NTIRE 2023 challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[11] Marcos V Conde, Eduard Zamfir, Radu Timofte, et al. Efficient deep models for real-time 4k image super-resolution. NTIRE 2023 benchmark and report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009. 4, 7

[13] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3677–3686, 2020. 8, 9

[14] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6202–6211, 2019. 7, 10

[15] 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, 2, 7, 9

[16] Yixuan Gao, Xiongkuo Min, Wenhan Zhu, Xiao-Ping Zhang, and Guangtao Zhai. Image quality score distribution prediction via alpha stable model. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 2

[17] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016. 13

[18] Pascal Getreuer. Automatic color enhancement (ACE) and its fast implementation. *Image Processing On Line*, 2:266–277, 2012. 3

[19] Deepti Ghadiyaram and Alan C Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, 2015. 9

[20] Bhupendra Gupta and Mayank Tiwari. Minimum mean brightness error contrast enhancement of color images using adaptive gamma correction with color preserving framework. *Optik*, 127(4):1671–1676, 2016. 3

[21] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Space-time-aware multi-resolution video enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2859–2868, 2020. 1

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645. Springer, 2016. 10

[23] 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 *Proceedings of the 2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE, 2017. 2, 9

[24] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020. 10

[25] Ding-Jiun Huang, Yu-Ting Kao, Tieh-Hung Chuang, Ya-Chun Tsai, Jing-Kai Lou, and Shuen-Huei Guan. Sb-vqa: A stack-based video quality assessment framework for video enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 9

[26] Jerin Geo James, Devansh Jain, and Ajit Rajwade. Globalflownet: Video stabilization using deep distilled global motion estimates. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5078–5087, 2023. 3

[27] Gu Jinjin, Cai Haoming, Chen Haoyu, Ye Xiaoxing, Jimmy S Ren, and Dong Chao. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In *Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 633–651. Springer, 2020. 7

[28] Xiaoyang Kang, Xianhui Lin, Kai Zhang, Zheng Hui, Wangmeng Xiang, Jun-Yan He, Xiaoming Li, Peiran Ren, Xuansong Xie, Radu Timofte, et al. NTIRE 2023 video colorization challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[29] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 10

[30] Jari Korhonen. Two-level approach for no-reference consumer video quality assessment. *IEEE Transactions on Image Processing*, 28(12):5923–5938, 2019. 2, 3

[31] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8878–8887, 2019. 3

[32] 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. 2, 3

[33] Dingquan Li, Tingting Jiang, and Ming Jiang. Unified quality assessment of in-the-wild videos with mixed datasets training. *International Journal of Computer Vision*, 129:1238–1257, 2021. 9

[34] Yawei Li, Yulun Zhang, Luc Van Gool, Radu Timofte, et al. NTIRE 2023 challenge on efficient super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[35] Yawei Li, Yulun Zhang, Luc Van Gool, Radu Timofte, et al. NTIRE 2023 challenge on image denoising: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[36] Jing Lin, Yuanhao Cai, Xiaowan Hu, Haoqian Wang, Youliang Yan, Xueyi Zou, Henghui Ding, Yulun Zhang, Radu Timofte, and Luc Van Gool. Flow-guided sparse transformer for video deblurring. *arXiv preprint arXiv:2201.01893*, 2022. 3

[37] Joe Yuchieh Lin, Rui Song, Chi-Hao Wu, TsungJung Liu, Haiqiang Wang, and C-C Jay Kuo. MCL-V: a streaming video quality assessment database. *Journal of Visual Communication and Image Representation*, 30:1–9, 2015. 2

[38] Wentao Liu, Zhengfang Duanmu, and Zhou Wang. End-to-end blind quality assessment of compressed videos using deep neural networks. In *Proceedings of the ACM Multimedia*, pages 546–554, 2018. 1

[39] Xiaohong Liu, Lei Chen, Wenyi Wang, and Jiying Zhao. Robust multi-frame super-resolution based on spatially weighted half-quadratic estimation and adaptive btv regularization. *IEEE Transactions on Image Processing*, 27(10):4971–4986, 2018. 1

[40] Xiaohong Liu, Lingshi Kong, Yang Zhou, Jiying Zhao, and Jun Chen. End-to-end trainable video super-resolution based on a new mechanism for implicit motion estimation and compensation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2416–2425, 2020. 1

[41] Xiaohong Liu, Xiongkuo Min, Wei Sun, Yulun Zhang, Kai Zhang, Radu Timofte, Guangtao Zhai, Yixuan Gao, Yuqin Cao, Tengchuan Kou, Yunlong Dong, Ziheng Jia, et al. NTIRE 2023 quality assessment of video enhancement challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[42] Xiaohong Liu, Kangdi Shi, Zhe Wang, and Jun Chen. Exploit camera raw data for video super-resolution via hidden markov model inference. *IEEE Transactions on Image Processing*, 30:2127–2140, 2021. 1

[43] Yongxu Liu, Jinjian Wu, Aobo Li, Leida Li, Weisheng Dong, Guangming Shi, and Weisi Lin. Video quality assessment with serial dependence modeling. *IEEE Transactions on Multimedia*, 24:3754–3768, 2021. 11

[44] Yongxu Liu, Jinjian Wu, Leida Li, Weisheng Dong, Jinpeng Zhang, and Guangming Shi. Spatiotemporal representation learning for blind video quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(6):3500–3513, 2021. 2, 3

[45] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12009–12019, 2022. 4

[46] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 6, 7

[47] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 5, 7

[48] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 6, 7, 9

[49] Wei Lu, Wei Sun, Xiongkuo Min, Wenhan Zhu, Quan Zhou, Jun He, Qiyuan Wang, Zicheng Zhang, Tao Wang, and Guangtao Zhai. Deep neural network for blind visual quality assessment of 4k content. *IEEE Trans. Broadcast.*, 2022. 2

[50] Feifan Lv, Feng Lu, Jianhua Wu, and Chongsoon Lim. MBLLEN: Low-light image/video enhancement using cnns. In *BMVC*, volume 220, page 4, 2018. 3

[51] Xiongkuo Min, Guangtao Zhai, Ke Gu, Yutao Liu, and Xiaokang Yang. Blind image quality estimation via distortion aggravation. *IEEE Transactions on Broadcasting*, 64(2):508–517, 2018. 2

[52] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2408–2415. IEEE, 2012. 9

[53] Manish Narwaria, Matthieu Perreira Da Silva, and Patrick Le Callet. Study of high dynamic range video quality assessment. In *Proceedings of the Applications of Digital Image Processing XXXVIII*, volume 9599, pages 289–301, 2015. 1

[54] Mikko Nuutinen, Toni Virtanen, Mikko Vaahteranoksa, Tero Vuori, Pirkko Oittinen, and Jukka Häkkinen.

CVD2014—A database for evaluating no-reference video quality assessment algorithms. *IEEE Transactions on Image Processing*, 25(7):3073–3086, 2016. 2

[55] 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. 2, 3

[56] 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. 8

[57] Kalpana Seshadrinathan, Rajiv Soundararajan, Alan Conrad Bovik, and Lawrence K Cormack. Study of subjective and objective quality assessment of video. *IEEE Transactions on Image Processing*, 19(6):1427–1441, 2010. 2

[58] Zhihao Shi, Xiaohong Liu, Chengqi Li, Linhui Dai, Jun Chen, Timothy N Davidson, and Jiying Zhao. Learning for unconstrained space-time video super-resolution. *IEEE Transactions on Broadcasting*, 68(2):345–358, 2021. 1

[59] Zhihao Shi, Xiaohong Liu, Kangdi Shi, Linhui Dai, and Jun Chen. Video frame interpolation via generalized deformable convolution. *IEEE Transactions on Multimedia*, 24:426–439, 2021. 1

[60] Zhihao Shi, Xiangyu Xu, Xiaohong Liu, Jun Chen, and Ming-Hsuan Yang. Video frame interpolation transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17482–17491, 2022. 1

[61] Alina Shutova, Egor Ershov, Georgy Perevozchikov, Ivan A Ermakov, Nikola Banic, Radu Timofte, Richard Collins, Maria Efimova, Arseniy Terekhin, et al. NTIRE 2023 challenge on night photography rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

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

[63] Alexandros Stergiou, Ronald Poppe, and Grigorios Kalliatakis. Refining activation downsampling with softpool. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10357–10366, 2021. 8

[64] Wei Sun, Xiongkuo Min, Wei Lu, and Guangtao Zhai. A deep learning based no-reference quality assessment model for ugc videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 856–865, 2022. 1, 2, 3, 4, 5, 7

[65] Wei Sun, Tao Wang, Xiongkuo Min, Fuwang Yi, and Guangtao Zhai. Deep learning based full-reference and no-reference quality assessment models for compressed ugc videos. In *Proc. Int. Conf. Multimedia and Expo Worksh.*, pages 1–6. IEEE, 2021. 2

[66] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *Proceedings of the International conference on machine learning*, pages 10096–10106. PMLR, 2021. 8

[67] Jan Theeuwes, Arthur F Kramer, Sowon Hahn, and David E Irwin. Our eyes do not always go where we want them to go: Capture of the eyes by new objects. *Psychological Science*, 9(5):379–385, 1998. 11

[68] 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. 2, 3, 12

[69] 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. 2, 3

[70] Florin-Alexandru Vasluianu, Tim Seizinger, Radu Timofte, et al. NTIRE 2023 image shadow removal challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[71] 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. 8, 9

[72] Chao Wang and Zhongfu Ye. Brightness preserving histogram equalization with maximum entropy: a variational perspective. *IEEE Transactions on Consumer Electronics*, 51(4):1326–1334, 2005. 3

[73] Haiqiang Wang, Weihao Gan, Sudeng Hu, Joe Yuchieh Lin, Lina Jin, Longguang Song, Ping Wang, Ioannis Katsavounidis, Anne Aaron, and C-C Jay Kuo. MCL-JCV: a jnd-based h. 264/avc video quality assessment dataset. In *Proceedings of the IEEE International Conference on Image Processing*, pages 1509–1513. IEEE, 2016. 2

[74] Longguang Wang, Yulan Guo, Yingqian Wang, Juncheng Li, Shuhang Gu, Radu Timofte, et al. NTIRE 2023 challenge on stereo image super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[75] Yilin Wang, Sasi Inguva, and Balu Adsumilli. YouTube UGC dataset for video compression research. In *Proceedings of the 2019 IEEE 21st International Workshop on Multimedia Signal Processing*, pages 1–5. IEEE, 2019. 2, 7, 9

[76] Yingqian Wang, Longguang Wang, Zhengyu Liang, Jungang Yang, Radu Timofte, Yulan Guo, et al. NTIRE 2023 challenge on light field image super-resolution: Dataset, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[77] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. ConvNeXt v2: Co-designing and scaling convnets with masked autoencoders. *arXiv preprint arXiv:2301.00808*, 2023. 6

[78] Noelle L Wood and Nelson Cowan. The cocktail party phenomenon revisited: attention and memory in the classic selective listening procedure of cherry (1953). *Journal of Experimental Psychology: General*, 124(3):243, 1995. 11

[79] Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Fast-vqa: Efficient end-to-end video quality assessment

with fragment sampling. In *Proceedings of the Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*, pages 538–554. Springer, 2022. 3, 6, 9, 10

[80] Haoning Wu, Chaofeng Chen, Liang Liao, Jingwen Hou, Wenxiu Sun, Qiong Yan, Jinwei Gu, and Weisi Lin. Neighbourhood representative sampling for efficient end-to-end video quality assessment. *arXiv preprint arXiv:2210.05357*, 2022. 9

[81] 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. 6, 7, 9, 10

[82] Wei Wu, Shuming Hu, Pengxiang Xiao, Sibin Deng, Yilin Li, Ying Chen, and Kai Li. Video quality assessment based on swin transformer with spatio-temporal feature fusion and data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 4

[83] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1191–1200, 2022. 5, 10

[84] Hojatollah Yeganeh, Shiqi Wang, Kai Zeng, Mahzar Eisapour, and Zhou Wang. Objective quality assessment of tone-mapped videos. In *Proceedings of the IEEE International Conference on Image Processing*, pages 899–903, 2016. 1

[85] Fuwang Yi, Mianyi Chen, Wei Sun, Xiongkuo Min, Yuan Tian, and Guangtao Zhai. Attention based network for no-reference ugc video quality assessment. In *Proc. IEEE Int. Conf. Image Process.*, pages 1414–1418. IEEE, 2021. 2

[86] Zhenqiang Ying, Maniratnam Mandal, Deepti Ghadiyaram, and Alan Bovik. Patch-vq:'patching up'the video quality problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14019–14029, 2021. 4, 9

[87] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3575–3585, 2020. 10

[88] Jiyang Yu and Ravi Ramamoorthi. Learning video stabilization using optical flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8159–8167, 2020. 3

[89] Pierluigi Zama Ramirez, Fabio Tosi, Luigi Di Stefano, Radu Timofte, et al. NTIRE 2023 challenge on hr depth from images of specular and transparent surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[90] Guangtao Zhai and Xiongkuo Min. Perceptual image quality assessment: a survey. *Science China Information Sciences*, 63:1–52, 2020. 2

[91] Fan Zhang, Songnan Li, Lin Ma, Yuk Chung Wong, and King Ngi Ngan. Ivp subjective quality video database. *The Chinese University of Hong Kong, http://ivp. ee. cuhk. edu. hk/research/database/subjective*, 2011. 2

[92] Fan Zhang, Yu Li, Shaodi You, and Ying Fu. Learning temporal consistency for low light video enhancement from single images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4967–4976, 2021. 1

[93] Taohong Zhang, Junnan Hu, Suli Fan, and Yixuan Yu. CD-CNN: a model based on class center vectors and distance comparison for wear particle recognition. *IEEE Access*, 8:113262–113270, 2020. 6

[94] Yulun Zhang, Kai Zhang, Zheng Chen, Yawei Li, Radu Timofte, et al. NTIRE 2023 challenge on image super-resolution (x4): Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[95] Zicheng Zhang, Wei Lu, Wei Sun, Xiongkuo Min, Tao Wang, and Guangtao Zhai. Surveillance video quality assessment based on quality related retraining. In *Proc. IEEE Int. Conf. Image Process.*, pages 4278–4282. IEEE, 2022. 2

[96] Zhao Zhang, Huan Zheng, Richang Hong, Mingliang Xu, Shuicheng Yan, and Meng Wang. Deep color consistent network for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1899–1908, 2022. 3

[97] Kai Zhao, Kun Yuan, Ming Sun, and Xing Wen. Zoom-vqa: Patches, frames and clips integration for video quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 5

[98] Minda Zhao and Qiang Ling. Pwstablenet: Learning pixel-wise warping maps for video stabilization. *IEEE Transactions on Image Processing*, 29:3582–3595, 2020. 3

[99] Shen Zheng and Gaurav Gupta. Semantic-guided zero-shot learning for low-light image/video enhancement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 581–590, 2022. 3

[100] Zhihang Zhong, Ye Gao, Yinqiang Zheng, and Bo Zheng. Efficient spatio-temporal recurrent neural network for video deblurring. In *Proceedings of the European Conference on Computer Vision*, pages 191–207, 2020. 3