International Journal of Computer Vision (2021) 129:1238–1257
https://doi.org/10.1007/s11263-020-01408-w

Unified Quality Assessment of in-the-Wild Videos with Mixed Datasets Training

Dingquan Li[1,2,4,5] · Tingting Jiang[1,3,6] · Ming Jiang[4]

Received: 21 December 2019 / Accepted: 18 November 2020 / Published online: 19 January 2021

Abstract
Video quality assessment (VQA) is an important problem in computer vision. The videos in computer vision applications are usually captured in the wild. We focus on automatically assessing the quality of in-the-wild videos, which is a challenging problem due to the absence of reference videos, the complexity of distortions, and the diversity of video contents. Moreover, the video contents and distortions among existing datasets are quite different, which leads to poor performance of data-driven methods in the cross-dataset evaluation setting. To improve the performance of quality assessment models, we borrow intuitions from human perception, specifically, content dependency and temporal-memory effects of human visual system. To face the cross-dataset evaluation challenge, we explore a mixed datasets training strategy for training a single VQA model with multiple datasets. The proposed unified framework explicitly includes three stages: relative quality assessor, nonlinear mapping, and dataset-specific perceptual scale alignment, to jointly predict relative quality, perceptual quality, and subjective quality. Experiments are conducted on four publicly available datasets for VQA in the wild, *i.e.*, LIVE-VQC, LIVE-Qualcomm, KoNViD-1k, and CVD2014. The experimental results verify the effectiveness of the mixed datasets training strategy and prove the superior performance of the unified model in comparison with the state-of-the-art models. For reproducible research, we make the PyTorch implementation of our method available at https://github.com/lidq92/MDTVSFA.

Keywords Content dependency · In-the-wild videos · Mixed datasets training · Temporal-memory effect · Video quality assessment

1 Introduction

Image/Video quality assessment (I/VQA) is a fundamental and longstanding problem in the image processing and computer vision community. It is involved in benchmarking and optimizing many vision applications, such as image classification (Dodge and Karam 2016), object tracking (Nieto et al. 2019), video compression (Rippel et al. 2019), image inpainting (Isogawa et al. 2019), and super resolution (Zhang et al. 2019a). Because of its importance, I/VQA has attracted significant attention in the past two decades (Wang et al. 2004a; Mittal et al. 2012; Zhang et al. 2014; Kang et al. 2014; Ma et al. 2016; Liu et al. 2017; Kim et al. 2018; Lin and Wang 2018). Videos obtained in the wild are often in low-quality because of many factors, such as out of focus, object motion, camera shakiness, under-/over- exposure, and adverse weather, etc. With the guidance of VQA in the wild, one can automatically identify, cull, repair or enhance low-quality videos before sending them to the subsequent vision applications, so that the applications can work in the real

Communicated by Mei Chen.

✉ Tingting Jiang
ttjiang@pku.edu.cn

Dingquan Li
dingquanli@pku.edu.cn

Ming Jiang
ming-jiang@pku.edu.cn

[1] National Engineering Laboratory for Video Technology, Peking University, Beijing, China

[2] Advanced Institute of Information Technology, Peking University, Hangzhou, China

[3] Department of Computer Science, Peking University, Beijing, China

[4] Laboratory of Mathematics and Its Applications, School of Mathematical Sciences, Peking University, Beijing, China

[5] Beijing International Center for Mathematical Research, Peking University, Beijing, China

[6] Advanced Innovation Center for Future Visual Entertainment, Beijing Film Academy, Beijing, China

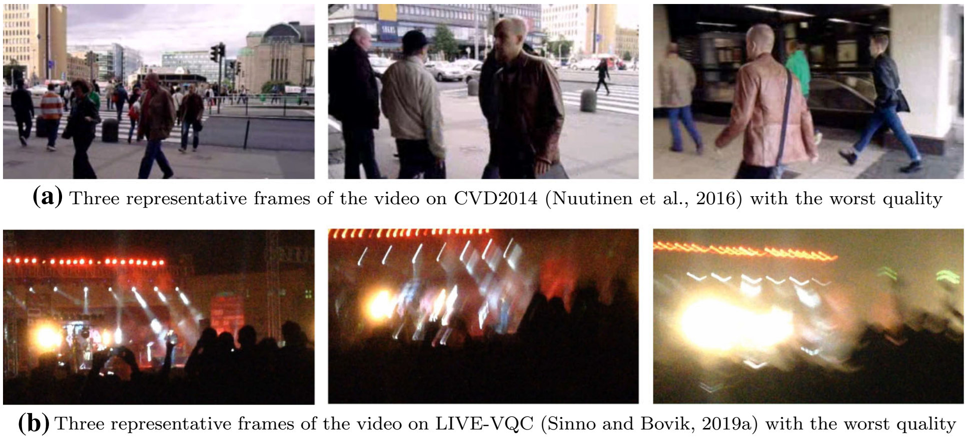
scenario. Thus, VQA in the wild is necessary for computer vision in the wild, but few attention is paid to this task.

VQA in the wild is a challenging task for the reason that the pristine videos are not available, the distortions are complex, and the contents are diverse. Compared to synthetically-distorted videos, in-the-wild videos contain huge amount of contents and may be infected with mixed real-world distortions, especially some of which are temporally heterogeneous (*e.g.*, temporary auto-focus blurs and exposure adjustments). Consequently, modern advanced I/VQA methods, *e.g.*, BRISQUE (Mittal et al. 2012) and VBLIINDS (Saad et al. 2014), validated on synthetically-distorted video datasets (Seshadrinathan et al. 2010; Moorthy et al. 2012), do a poor job in predicting the quality of in-the-wild videos (Men et al. 2017; Ghadiyaram et al. 2018; Nuutinen et al. 2016; Sinno and Bovik 2019a) (see Tables 5 and 6).

Some efforts have been made to generate a better feature for VQA in the wild (You and Korhonen 2019; Korhonen 2019; Li et al. 2019a). Korhonen (2019) obtains well-behaved low-complexity features for all frames and high-complexity features for representative frames, so that good quality predictions can be achieved by the support vector regression or the random forest regression. You and Korhonen (2019) learn effective spatio-temporal features with 3D convolutional neural network (3D-CNN) and predict the video quality by a long-short term memory (LSTM) network. Our previous work (Li et al. 2019a) borrows intuitions from human visual system (HVS), which extracts content-aware and distortion-sensitive features. Although the above mentioned methods achieve superior performances on the benchmark VQA datasets individually, their performances are poor in cross-dataset evaluation setting (See Table 7). For example, when the model is trained on KoNViD-1k (Hosu et al. 2017), the test performance on LIVE-Qualcomm (Ghadiyaram et al. 2018) or CVD2014 (Nuutinen et al. 2016) drops sharply (Korhonen 2019). This may be caused by the over-fitting problem in the training process and the discrepancy of data distribution among the datasets.

To face this cross-dataset evaluation challenge, one possible solution is to mix multiple datasets during the training phase, so that the data-driven model can learn the characteristics of video contents and distortions from all these datasets. Mixed datasets training provides two advantages. First, it provides a single unified model for all datasets/applications instead of multiple models for different datasets. Second, it makes the utmost of existing relevant data for VQA model training, since the largest size of current in-the-wild VQA datasets is only 1200 and acquiring new annotations is time-consuming. However, mixed datasets training is not trivial, since the ranges of subjective quality scores among different datasets are inconsistent. A naïve strategy is the “linear re-scaling”, which maps all subjective score ranges of different datasets to the same range. Nevertheless, the ranges of the inherent video quality among these datasets are not equal in most circumstances. For instance, in Fig. 1, both the two videos are the worst in their corresponding datasets. The video in Fig. 1a has better quality in comparison with the video in Fig. 1b, since the latter one contains more complicated distortions, including motion blur, under-/over-exposure, and grainy noise. However, linear re-scaling leads to the same quality labels for them. Such “inconformity” will disturb the training process, thus a good performance is hard to achieve (see Fig. 6 and Table 5).

To tackle the above inconformity problem, we should align subjective quality scores for different datasets. One way is conducting an additional subjective study to re-align the subjective quality scores. The other way is to learn the alignment of subjective quality scores for these datasets. As the first way is time-consuming and impracticable when more and more datasets are considered, we choose the second one. Before introducing our method, we first introduce three important quality concepts: perceptual quality, subjective quality, and relative quality.


(a) Three representative frames of the video on CVD2014 (Nuutinen et al., 2016) with the worst quality

(b) Three representative frames of the video on LIVE-VQC (Sinno and Bovik, 2019a) with the worst quality

Fig. 1 An illustration of the videos with the worst quality on CVD2014 and LIVE-VQC, respectively (Full videos are provided in https://bit.ly/3csmHYk). The upper video has a better quality in comparison with the lower video. However, linear re-scaling leads to the same quality labels for them. Such “inconformity” will disturb the training process, and lead to a poor performance

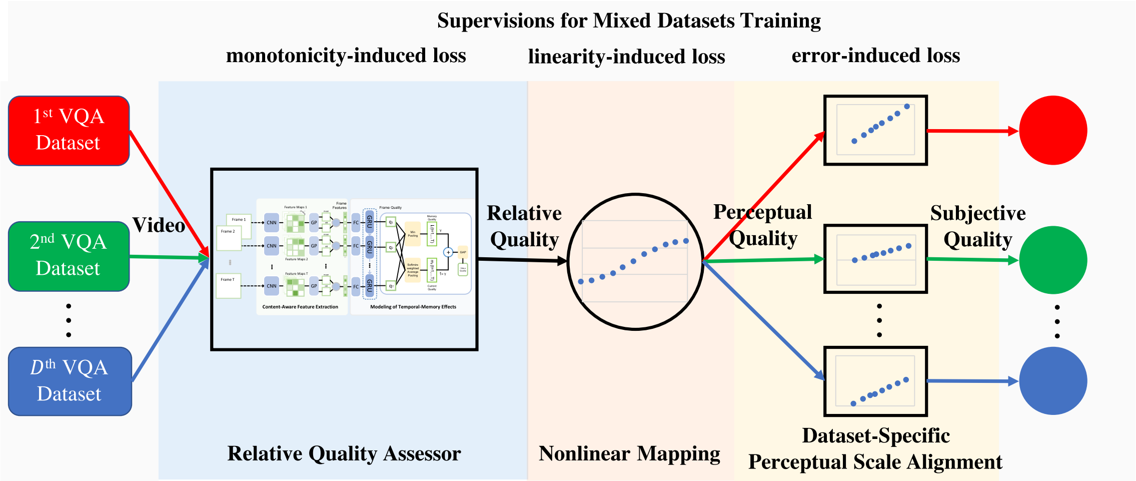

Supervisions for Mixed Datasets Training
monotonicity-induced loss
linearity-induced loss
error-induced loss
1st VQA Dataset
2nd VQA Dataset
Dth VQA Dataset
Video
Content-Aware Feature Extraction
Modeling of Temporal-Memory Effects
Relative Quality
Perceptual Quality
Subjective Quality
Relative Quality Assessor
Nonlinear Mapping
Dataset-Specific Perceptual Scale Alignment

Fig. 2 An overview of the proposed unified framework. It consists of three stages: relative quality assessor, nonlinear mapping, and dataset-specific perceptual scale alignment for predicting relative quality, perceptual quality, and subjective quality, respectively. The supervisions for mixed datasets training at the three stages are monotonicity-induced loss, linearity-induced loss, and error-induced loss, respectively. D is the number of datasets

Perceptual Quality Perceptual quality is an ideal concept that is related to human perception of video quality, and only if we gather all the videos in the wild and conduct the largest scale subjective study can we get the ground-truth of perceptual quality. Perceptual quality can be used for benchmarking and optimizing video processing systems/algorithms, but its ground-truth is impossible to obtain since we cannot conduct such a large-scale subjective study on all videos in the wild.

Subjective Quality As an "approximation" of perceptual quality, subjective quality is considered, whose ground-truth can be accessed by conducting a subjective study on a video dataset of limited size. Although subjective quality is designed to reflect perceptual quality, it may have different ranges for different datasets. In terms of this fact, we can assume the subjective quality to be linearly correlated with the perceptual quality for a single dataset, but the linear transformations between subjective quality and perceptual quality are not necessarily the same for different datasets. Subjective quality can be used as a supervised signal for the prediction of perceptual quality.

Relative Quality Compared to directly rating the quality of a video in the subjective study, it is easier for humans to choose a video with better quality from two videos. In terms of this fact, we define the concept of relative quality, which can be accessed by ranking the quality of videos. Relative quality can be used for benchmarking video processing algorithms. However, due to its nonlinearity to perceptual quality, it might not be directly used for optimization. For example, the optimization might be early stopped when the relative quality is approaching the perfect value while the perceptual quality is far from the perfect one.

With the above three concepts, we show our solution. We decompose the VQA problem into three sub-problems, *i.e.*, predicting relative quality, perceptual quality, and subjective quality in turn (see Fig. 2). For details, our proposed model contains three stages to solve these three sub-problems. First, to predict the relative quality, we use our previous HVS-inspired VQA model (Li et al. 2019a) as the backbone. The relative quality assessor takes the video as input and outputs a relative quality score. This stage focuses on prediction monotonicity, which describes the ability to provide the quality ranking for any list of videos that is consistent with subjective quality. Correspondingly, we propose a monotonicity-induced loss for this stage. Second, to predict the perceptual quality, we adopt the well-known 4-parameter logistic function for characterizing the nonlinearity of human perception on video quality (VQEG 2000). We reformulate this function and design it as a network module. The nonlinear mapping module maps the relative quality of a video to the perceptual quality of a video. The predicted perceptual quality is expected to be linearly correlated with the subjective quality, and we propose a linearity-induced loss as the supervision for this stage. Third, we learn a dataset-specific perceptual scale alignment for each dataset, which tries to map the perceptual quality of a video to the subjective quality of the video on its belonging dataset. With this dataset-specific alignment, an error-induced loss can be used as the supervision without disturbing the training. Under this model, we can use the above three losses for mixed datasets training to solve the "inconformity" problem.

To verify the effectiveness of the proposed unified model with the mixed datasets training strategy, we conduct comparative experiments on four publicly available datasets for VQA in the wild, *i.e.*, KoNViD-1k (Hosu et al. 2017), CVD2014 (Nuutinen et al. 2016), LIVE-Qualcomm (Ghadiyaram et al. 2018), and LIVE-VQC (Sinno and Bovik 2019a). Our method is compared with several modern advanced methods. In terms of prediction monotonicity and prediction accuracy, the superior performances of our method across datasets are verified by the experimental results.

Lastly, we highlight the relationship and difference between our previous work (Li et al. 2019a) and this work. The model design in this work is build upon the model in our previous work. However, there are two major differences between our previous work and this work. First, this work focuses on model optimization with mixed datasets training while our previous work does not consider mixed datasets training. Second, in this work, it is the first time to decompose the VQA problem into three sub-problems: predicting relative quality, perceptual quality, and subjective quality, and we propose a unified VQA framework that explicitly designs three stages to tackle these three sub-problems.

2 Related Work

This section reviews some related work. Section 2.1 overviews several representative VQA methods, especially the VQA methods for in-the-wild videos. Section 2.2 introduces mixed datasets training in computer vision, especially in the tasks of quality assessment.

2.1 Video Quality Assessment

Classical VQA methods are grounded on different cues, such as structures (Wang et al. 2004b, 2012), motion (Seshadrinathan and Bovik 2010; Manasa and Channappayya 2016), energy (Li et al. 2016a), saliency (Zhang and Liu 2017; You et al. 2014), gradients (Lu et al. 2019), or natural video statistics (NVS) (Mittal et al. 2016; Saad et al. 2014; Sinno and Bovik 2019b). Besides, some VQA methods focus on the fusion of primary features (Freitas et al. 2018; Li et al. 2016b). Recently, several VQA methods exploit the use of deep learning in this task (Zhang et al. 2019c; Liu et al. 2018; Kim et al. 2018; Zhang et al. 2020). Kim et al. (2018) obtain the spatio-temporal sensitivity maps by a CNN model. Liu et al. (2018) exploit the 3D-CNN model for multi-task learning of codec classification and quality assessment for compressed videos. Zhang et al. (2019c) and Zhang et al. (2020) make use of both video and image data with transfer learning. However, all these methods are proposed for quality assessment of synthetically-distorted videos, and they are not applicable to in-the-wild videos or their performances are poor on in-the-wild datasets. Note that the relevant concept "streaming video quality-of-experience (QoE)" is beyond the scope of this work, and the interested reader can refer to these two good surveys (Seufert et al. 2014; Juluri et al. 2015).

Quality assessment of in-the-wild videos has been attracting significant attention in recent years. Four relevant datasets have been constructed with corresponding subjective studies, *i.e.*, CVD2014 (Nuutinen et al. 2016), KoNViD-1k (Hosu et al. 2017), LIVE-Qualcomm (Ghadiyaram et al. 2018), and LIVE-VQC (Sinno and Bovik 2019a). Since we cannot access the pristine reference videos in this situation, only no-reference VQA (NR-VQA) methods are applicable. The deep learning-based VQA models described in the last paragraph are unfeasible in this problem since they either need the reference information (Zhang et al. 2019c; Kim et al. 2018; Zhang et al. 2020) or only suit for compression artifacts (Liu et al. 2018). Thus, in our previous work (Li et al. 2019a), we propose a deep learning-based model for predicting the quality of in-the-wild videos. The model extracts content-aware distortion-sensitive features from CNN models trained for image classification tasks, and uses a gated recurrent unit (GRU) followed by a subjectively-inspired temporal pooling layer for modeling the temporal-memory effect. Concurrent works to our previous work are You and Korhonen (2019), Varga (2019) and Varga and Szirányi (2019). Although all of these methods achieve a good performance, they do not enable mixing multiple datasets during the training phase. As a result, their performances are poor in the cross-dataset evaluation setting. The main purpose of this paper is to propose an elegant mixed datasets training strategy. With this strategy, we can train a unified model that learns the characteristics of videos from all datasets and thus further improve the overall performance over the datasets.

2.2 Mixed Datasets Training

Mixed datasets training has two advantages. One is to provide a unified model for all datasets. The other is to take full advantage of existing relevant datasets for improving the model learning. Therefore, many computer vision tasks consider mixed datasets training, such as person re-identification (Lv et al. 2018; Li et al. 2019c), monocular depth estimation (Lasinger et al. 2019), and human parsing (He et al. 2019).

There are some relevant works in quality assessment tasks that consider mixed datasets training. The challenge is that ranges of subjective quality scores are inconsistent across datasets. Korhonen (2019) uses a naïve method to handle this challenge, *i.e.*, linearly re-scaling the subjective quality scores of different datasets to the same range. Pair-wise learning considers only the relative quality score instead of the absolute subjective quality scores, and thus can bypass the "inconformity" problem. Therefore, several I/VQA works

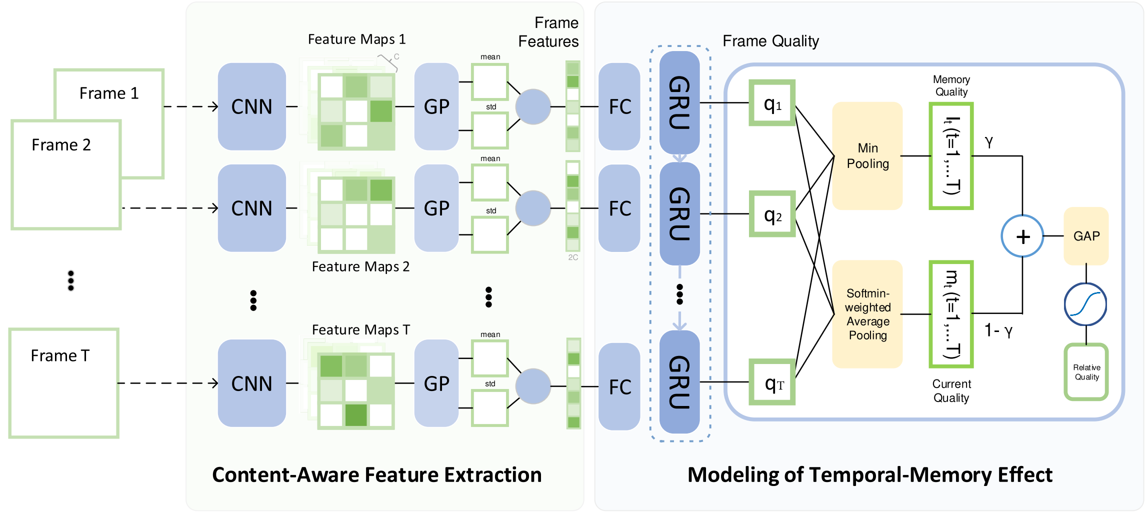


Fig. 3 Relative Quality Assessor. It mainly consists of two modules. Module I includes a pre-trained CNN with effective global pooling (GP) serving as a content-aware feature extractor. Module II models temporal-memory effect and it includes two sub-modules: a GRU network and a subjectively-inspired temporal pooling layer. Note that the GRU network is the unrolled version of one GRU and the parallel CNNs/FCs share weights

consider pair-wise learning for mixed datasets training, while they use different loss functions for training (Yang et al. 2019; Zhang et al. 2019b; Krasula et al. 2020). Yang et al. (2019) use the margin ranking loss and the Euclidean loss. Zhang et al. (2019b) consider the cross entropy loss and the fidelity loss. Krasula et al. (2020) determine different and similar pairs based on statistical analysis on the mean and standard deviation of subjective ratings, and then define the training objective as the correct classification rate of these pairs. However, pair-wise learning will increase the training time. In the next section, we show that our proposed monotonicity-induced loss can be regarded as an extension of the losses in Yang et al. (2019) and Zhang et al. (2019b) with a more efficient implementation. Besides the monotonicity-induced loss, we also propose a linearity-induced loss and assign a dataset-specific perceptual scale alignment to enable mixing multiple datasets during the training phase.

3 Proposed Method

3.1 Overview

Figure 2 shows the overview of the proposed unified VQA framework for quality assessment of in-the-wild videos. Our VQA model consists of three stages: relative quality assessor, nonlinear mapping, and dataset-specific perceptual scale alignment for predicting relative quality, perceptual quality, and subjective quality, respectively.

The flow of our proposed unified framework is as follows. At the first stage, to predict the relative quality, we learn a relative quality assessor with the supervision of a monotonicity-inspired loss, where the monotonicity-induced loss is derived from the monotonicity condition and it is the sum of all pair-wise losses. To account for the content dependency and temporal-memory effects of human perception, we design our relative quality assessor as a deep neural network that includes two key modules: content-aware feature extraction and modeling of temporal-memory effect. At the second stage, to predict the perceptual quality, a nonlinear mapping module is added after the relative quality assessor, to explicitly account for the nonlinearity of human perception. The parameters in this module are learned with the supervision of a linearity-induced loss based on Pearson's linear correlation. At the third stage, to predict the subjective quality, a dataset-specific perceptual scale alignment layer is added to map the predicted perceptual quality to the subjective quality of a video on each dataset. After the alignment, the widely-used error-induced loss is used as the supervision.

Thus, in our mixed datasets training strategy, three kinds of losses are involved. For each dataset, the overall loss is the sum of these three kinds of losses on the dataset. To emphasize the datasets with larger loss values, our final training loss is a softmax-weighted loss over all training datasets. With this strategy, we can learn a single unified VQA model for multiple datasets by mixing them all during the training phase.

3.2 Relative Quality Assessor

This subsection describes the design of the relative quality assessor. The framework of our relative quality assessor is shown in Fig. 3. We adopt the model in our previous work (Li et al. 2019a) as the backbone of the relative quality assessor. It integrates the two eminent effects of human perception into the assessor. One is the content dependency effect, which guides us introducing the content-aware feature extraction module. The other is the temporal-memory effect, which is modeled in the feature level and the quality score level.

3.2.1 Content-Aware Feature Extraction

In the visual quality assessment task, human perception is content dependent (Siahaan et al. 2018; Triantaphillidou et al. 2007; Wang et al. 2017; Bampis et al. 2017; Zhang et al. 2018; Li et al. 2019a, b). This can be attributed to the fact that, the complexity of distortions, the human tolerance thresholds for distortions, and the human preferences could vary a lot in different contents/scenes. Since there are diverse contents in the in-the-wild scenario, a relative quality assessor which mimics human perception, should take this effect into accounts. So we need to extract features that are not only distortion-sensitive but also content-aware. The image classification models pre-trained on ImageNet (Deng et al. 2009) using CNN possess the discriminatory power of different content information. Thus, the deep features extracted from these models, *e.g.*, ResNet (He et al. 2016), are expected to be content-aware. Meanwhile, Dodge and Karam (2016) point out that the deep features are distortion-sensitive. So it is reasonable to extract content-aware and distortion-sensitive features from pre-trained image classification models.

Assuming the video is a stack of T frames $\mathbf{I}_t (t = 1, 2, \dots, T)$, we feed each video frame into a pre-trained CNN model and get the corresponding deep feature maps $\mathbf{M}_t$ from its top convolutional layer:

$$\mathbf{M}_t = \text{CNN}(\mathbf{I}_t). \tag{1}$$

$\mathbf{M}_t$ contains a total of C feature maps. Then, we apply spatial global pooling (GP) for each feature map of $\mathbf{M}_t$. Applying only the spatial global average pooling operation (GP_{mean}) to $\mathbf{M}_t$ discards much information of $\mathbf{M}_t$. We further consider the spatial global standard deviation pooling operation (GP_{std}) to preserve the variation information in $\mathbf{M}_t$. The output feature vectors of GP_{mean}, GP_{std} are $\mathbf{f}_t^{\text{mean}}$, $\mathbf{f}_t^{\text{std}}$ respectively.

$$\mathbf{f}_t^{\text{mean}} = \text{GP}_{\text{mean}}(\mathbf{M}_t), \quad \mathbf{f}_t^{\text{std}} = \text{GP}_{\text{std}}(\mathbf{M}_t). \tag{2}$$

After that, $\mathbf{f}_t^{\text{mean}}$ and $\mathbf{f}_t^{\text{std}}$ are concatenated to serve as content-aware and distortion-sensitive features $\mathbf{f}_t$:

$$\mathbf{f}_t = \mathbf{f}_t^{\text{mean}} \oplus \mathbf{f}_t^{\text{std}}, \tag{3}$$

where $\oplus$ is the concatenation operator and the length of $\mathbf{f}_t$ is $2C$.

3.2.2 Modeling of Temporal-Memory Effect

Temporal-memory effect is another important clue for designing objective VQA models (Park et al. 2013; Seshadrinathan and Bovik 2011; Xu et al. 2014; Choi and Bovik 2018; Kim et al. 2018). It induces that video quality rating is influenced by historic memory. We model the temporal-memory effect in two aspects. In the feature integration aspect, we adopt a GRU network for modeling the long-term dependencies in our method. In the quality pooling aspect, we propose a subjectively-inspired temporal pooling model and embed it into the network.

Long-Term Dependencies Modeling Existing NR-VQA methods cannot well model the long-term dependencies in the VQA task. To handle this issue, we resort to GRU (Cho et al. 2014), a recurrent neural network model with gates control.

The dimension of the extracted content-aware features is very high, which is not suitable for GRU training. Therefore, we perform dimension reduction using a single fully-connected (FC) layer before feeding them into GRU, that is:

$$\mathbf{x}_t = \mathbf{W}_{fx}\mathbf{f}_t + \mathbf{b}_{fx}, \tag{4}$$

where $\mathbf{W}_{fx}$ and $\mathbf{b}_{fx}$ are the parameters in the single FC layer. Without the bias term, it acts as a linear dimension reduction model.

After dimension reduction, the reduced features $\mathbf{x}_t (t = 1, 2, \dots, T)$ are sent to GRU. We consider the hidden states of GRU as the integrated features $\mathbf{h}_t$, whose initial values are $\mathbf{h}_0$. $\mathbf{h}_t$ is calculated as follow.

$$\mathbf{h}_t = \text{GRU}(\mathbf{x}_t, \mathbf{h}_{t-1}). \tag{5}$$

With the integrated features $\mathbf{h}_t$, we predict the frame quality score q_t by adding a single FC layer:

$$q_t = \mathbf{W}_{hq}\mathbf{h}_t + \mathbf{b}_{hq}, \tag{6}$$

where $\mathbf{W}_{hq}$ and $\mathbf{b}_{hq}$ are the weight and bias parameters.

Subjectively-Inspired Temporal Pooling In subjective experiments, subjects are intolerant of poor quality video events (Park et al. 2013). More specifically, temporal hysteresis effect is found in the subjective experiments (Seshadrinathan and Bovik 2011). That is, subjects react sharply to drops in video quality and provide poor quality for such time interval, but react dully to improvements in video quality thereon. Inspired by these observations, to connect the predicted frame-level quality to the video-level quality, we put forward a new differentiable temporal pooling model. Details are as follows.

To mimic the human's intolerance to poor quality events, we define a memory quality element l_t at the tth frame as the minimum of quality scores over the previous several frames:

$$l_t = \begin{cases} q_t & \text{for } t = 1, \\ \min_{k \in V_{\text{prev}}} q_k & \text{for } t > 1, \end{cases} \tag{7}$$

where $V_{\text{prev}} = \{\max(1, t-\tau), \ldots, t-2, t-1\}$ is the index set of the considered frames, and τ is a hyper-parameter relating to the temporal duration.

Accounting for the fact that subjects react sharply to the drops in quality but react dully to the improvements in quality, we construct a current quality element m_t at the tth frame, using the weighted quality scores over the next several frames, where larger weights are assigned for worse quality frames. Specifically, we define the weights w_t^k by a differentiable softmin function, *i.e.*, a composition of the negative linear function and the softmax function.

$$m_t = \sum_{k \in V_{\text{next}}} q_k w_t^k, \quad w_t^k = \frac{e^{-q_k}}{\sum_{j \in V_{\text{next}}} e^{-q_j}}, k \in V_{\text{next}}, \tag{8}$$

where $V_{\text{next}} = \{t, t+1, \ldots, \min(t+\tau, T)\}$ is the index set of the related frames.

In the end, we approximate the subjective frame quality scores by linearly combining the memory quality and current quality elements. The relative quality score Q_r is then calculated by temporal global average pooling (GAP) of the approximate scores and bounded by a sigmoid function:

$$q_t' = \gamma l_t + (1-\gamma) m_t, \tag{9}$$

$$Q_r = \sigma\left(\frac{1}{T}\sum_{t=1}^{T} q_t'\right), \tag{10}$$

where γ is a hyper-parameter to balance the contributions of memory and current elements to the approximate score, and $\sigma(\cdot)$ is the sigmoid function.

3.3 Nonlinear Mapping

For predicting the perceptual quality, we add a nonlinear mapping module after the relative quality assessor to explicitly account for the nonlinearity of human perception on video quality (VQEG 2000). The nonlinear mapping module can be a complex neural network with many parameters, or just a simple nonlinear function with few parameters.

Following the suggestion by Video Quality Experts Group (VQEG 2000), we can use a 4-parameter logistic function for nonlinear mapping.

$$Q_p = f(Q_r) = \frac{\beta_1 - \beta_2}{1 + e^{-\frac{Q_r - \beta_3}{|\beta_4|}}} + \beta_2, \tag{11}$$

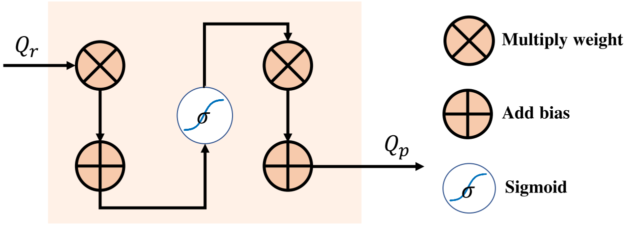


Fig. 4 Illustration of the nonlinear mapping module

where β_1 to β_4 are fitting parameters, Q_r is the relative quality score, and Q_p is the perceptual quality score.

We can reformulate Eq. (11) as the following.

$$Q_p = \beta_1' \sigma(\beta_4' Q_r + \beta_3') + \beta_2', \tag{12}$$

where $\beta_1' \leftarrow \beta_1 - \beta_2$, $\beta_2' \leftarrow \beta_2$, $\beta_3' \leftarrow -\frac{\beta_3}{|\beta_4|}$, and $\beta_4' \leftarrow \frac{1}{|\beta_4|}$. And β_1', β_2' are parameters to control the range of Q_p. β_3', β_4' are parameters to control the normalization of Q_r. Therefore, it is equivalent to "Linear (*i.e.*, Multiply Weights and Add Bias)+Sigmoid+Linear", as shown in Fig. 4.

With the reformulation, we can design the 4-parameter nonlinear mapping as a network module. Since we will handle the scale problem in the next stage, the nonlinear mapping just handles the nonlinearity and does not change the scale, *i.e.*, the ranges of Q_r and Q_p are both [0, 1]. We need to initialize the 4 parameters in this module at the start of the training. Random initialization is not a good choice since we have priors of Q_r and Q_p. Therefore, we can have a better initialization as follows.

$$\begin{aligned} &\beta_1' \leftarrow \sup(Q_p) - \inf(Q_p) = 1, \\ &\beta_2' \leftarrow \inf(Q_p) = 0, \\ &\beta_3' \leftarrow -c * \text{mean}(Q_r)/\text{std}(Q_r), c = 1, \\ &\beta_4' \leftarrow c/\text{std}(Q_r), \end{aligned} \tag{13}$$

where mean$(\cdot)$, std$(\cdot)$, inf$(\cdot)$, sup$(\cdot)$ indicate the mean, standard deviation, infimum, and supremum functions, respectively.

3.4 Dataset-Specific Perceptual Scale Alignment

Since the subjective study is designed to reflect human perception on video quality, based on the concepts of subjective quality and perceptual quality, we can assume that the subjective quality is linearly correlated with the perceptual quality. Thus, the perceptual scale alignment can be simply set as a specific FC layer.

$$Q_s = \xi_1 Q_p + \xi_2, \tag{14}$$

where Q_s is the predicted subjective quality score, and ξ_1, ξ_2 are the scale and shift parameters.

Since different datasets have different ranges of subjective quality scores, we need a dataset-specific alignment of perceptual scale on each dataset. Equation (14) is then modified as follows.

$$Q_s^{(d)} = \xi_1^{(d)} Q_p + \xi_2^{(d)} (d = 1, \dots, D), \tag{15}$$

where $Q_s^{(d)}$ is the predicted subjective quality score on the dth dataset, $\xi_1^{(d)}, \xi_2^{(d)}$ are the scale and shift parameters for the dth dataset, and D is the number of considered datasets. These parameters can be determined by least square regression (LSR) or just jointly learned with other parameters by iterative stochastic gradient decent (SGD) algorithm. The latter way can provide supervision for end-to-end network training and it is adopted in our mixed datasets training strategy.

3.5 Mixed Datasets Training Strategy

We have introduced the unified VQA model in the above. In this subsection, we show how we can enable mixed datasets training when the ranges of subjective quality scores are not consistent among the VQA datasets. For the first and second stages, the relative quality and perceptual quality are not involved with the ranges of subjective quality. We bypass the inconformity problem by designing two losses to supervise the training process of predicting relative quality and perceptual quality. For the third stage, to predict subjective quality of videos on each dataset, we learn a dataset-specific perceptual scale alignment for each dataset to avoid the inconformity caused by the naïve linear re-scaling. Such dataset-specific alignment enables mixing multiple datasets during the training without disturbing the process. Specifically, monotonicity-induced loss is proposed for Stage 1 "relative quality assessor", and linearity-induced loss is adopted for Stage 2 "nonlinear mapping". As for Stage 3 "dataset-specific perceptual scale alignment", we can just use the widely-used error-induced (*i.e.*, normalized L_1) loss as the supervision.

Assume we have D datasets of VQA, and the dth dataset contains N_d videos ($d = 1, \dots, D$). For the ith video of the dth dataset, we denote its predicted relative quality score as $Q_r^{d,i}$, the predicted perceptual quality score as $Q_p^{d,i}$, the predicted subjective quality score as $Q_s^{d,i}$, and ground-truth subjective quality score as $Q^{d,i}$.

3.5.1 Monotonicity-Induced Loss

The goal of relative quality assessor is to achieve the best prediction monotonicity. That is, it aims to give a quality ranking for any list/pair of videos from the same dataset, that is consistent with subjective quality. A natural objective is to maximize the Spearman's rank-order correlation coefficient (SROCC) or Kendall's rank-order correlation coefficient (KROCC). However, they are not applicable to back-propagation based neural network optimization due to their non-differentiable property.

Let us take a close look at the monotonicity condition. For all $i, j = 1, \dots, N_d, d = 1, \dots, D$,

$$(Q_r^{d,i} - Q_r^{d,j})(Q^{d,i} - Q^{d,j}) \geq 0. \tag{16}$$

So we can consider the sum of the pair-wise losses as a surrogate. We call this monotonicity-induced loss, which is defined as follows.

$$\begin{aligned} L_{\text{rel}}^{(d)} &= \frac{2}{N_d(N_d - 1)} \sum_{i<j} E_r^{d,(i,j)}, \\ E_r^{d,(i,j)} &= \max\{(Q_r^{d,i} - Q_r^{d,j}) * \text{sign}(Q^{d,j} - Q^{d,i}), 0\}, \end{aligned} \tag{17}$$

where $E_r^{d,(i,j)}$ is the error term induced by the monotonicity condition, *i.e.*, Eq. (16). Here, we choose the error term as the margin ranking loss used in Yang et al. (2019). It can also be in the form of the fidelity loss or the cross entropy loss as described in Zhang et al. (2019b). Note that compared to pairwise learning, the number of forward operations is reduced from $C_{N_d}^2$ to N_d in our list-wise learning setting. Together with the vectorization form, we provide a much more efficient implementation and save more training time than the pairwise learning used in image quality assessment (Yang et al. 2019; Zhang et al. 2019b).

3.5.2 Linearity-Induced Loss

The goal of the nonlinear mapping module is to achieve the best prediction linearity between the predicted perceptual quality scores and the subjective quality scores. Pearson's linear correlation coefficient (PLCC) is a good objective for characterizing linearity. And it is differentiable, so we can define our linearity-induced loss for nonlinear mapping module as follow.

$$\begin{aligned} L_{\text{lin}}^{(d)} &= (1 - \text{PLCC}_d)/2, \\ \text{PLCC}_d &= \frac{\sum_i (Q_p^{d,i} - \bar{Q}_p^{(d)})(Q^{d,i} - \bar{Q}^{(d)})}{\sqrt{\sum_i (Q_p^{d,i} - \bar{Q}_p^{(d)})^2 \sum_i (Q^{d,i} - \bar{Q}^{(d)})^2}}, \end{aligned} \tag{18}$$

where $\bar{Q}_p^{(d)} = \frac{1}{N_d} \sum_i Q_p^{d,i}$ and $\bar{Q}^{(d)} = \frac{1}{N_d} \sum_i Q^{d,i}$. Note that PLCC-induced loss is also considered in Ma et al. (2018), Liu et al. (2018) and Li et al. (2020).

3.5.3 Error-Induced Loss

After dataset-specific perceptual scale alignment, our goal is to minimize the absolute prediction error. In this stage, any regression error can be used as the loss function. We simply choose the widely-used error-induced (*i.e.*, normalized L_1) loss in this work. More general and robust regression losses may be explored to further improve the optimization performance (Barron 2019). To balance the losses among different datasets, we consider the inverse scale on each dataset as a normalization factor.

$$L_{\text{err}}^{(d)} = \sum_i \frac{1}{N_d} \frac{\left| Q_s^{d,i} - Q^{d,i} \right|}{S_d}, \tag{19}$$

where $S_d = \max(Q^{d,i}) - \min(Q^{d,i})$ is the range of the subjective quality scores on the dth dataset.

3.5.4 Final Loss for The Whole Model

We obtain the loss for the dth dataset ($d = 1, \ldots, D$) from the above three losses $L_{\text{rel}}^{(d)}, L_{\text{lin}}^{(d)}, L_{\text{err}}^{(d)}$.

$$L^{(d)} = L_{\text{rel}}^{(d)} + L_{\text{lin}}^{(d)} + L_{\text{err}}^{(d)}, \tag{20}$$

and the overall final loss for training a single unified model from multiple datasets is defined as a softmax-weighted average of the losses over all datasets.

$$\begin{aligned} L &= \textstyle\sum_d w^{(d)} L^{(d)}, \\ w^{(d)} &= e^{L^{(d)}} / \textstyle\sum_d e^{L^{(d)}}, \end{aligned} \tag{21}$$

where $w^{(d)}$ is the weight of $L^{(d)}$ ($d = 1, \ldots, D$).

3.6 Implementation Details

We choose ResNet-50 (He et al. 2016) pre-trained on ImageNet (Deng et al. 2009) for the content-aware feature extraction, and the feature maps are extracted from its top convolutional layer “res5c”. In this instance, the dimension of $\mathbf{f}_t$ is $2C = 4096$. The feature dimension is then reduced from 4096 to 128, followed by a single-layer GRU network with hidden size 32. τ and γ in the temporal pooling layer are set as 12 and 0.5, respectively. We choose the 4-parameter nonlinear mapping, and the parameters in the module are initialized based on Eq. (13). We freeze the parameters in the pre-trained ResNet-50 to ensure that the content-aware property is not altered, and we train the other part of the network in an end-to-end manner. We train our model using Adam optimizer (Kingma and Ba 2014) for 40 epochs with an initial learning rate 1e-4, a training batch size 32 for each dataset. The proposed model is implemented with PyTorch (Paszke et al. 2019). To support reproducible scientific research, we release the code at https://github.com/lidq92/MDTVSFA.

4 Experiments

This section reports and analyzes the experimental results. We first describe the experimental setup, including the benchmark datasets, compared methods and basic evaluation criteria. Next, we study the effectiveness of our mixed datasets training strategy. After that, the performance comparison is carried out between our method and the state-of-the-art methods. Finally, the computational efficiency is briefly discussed.

4.1 Experimental Setup

Benchmark Datasets Currently, there are four datasets for quality assessment of in-the-wild videos, including CVD2014 (Nuutinen et al. 2016), KoNViD-1k (Hosu et al. 2017), LIVE-Qualcomm (Ghadiyaram et al. 2018), and LIVE-VQC (Sinno and Bovik 2019a). We summarize their brief information in Table 1. We can see that the four datasets have different characteristics and the ranges of mean opinion score (MOS) are different among these datasets. In the default setting, each dataset is split into 80%, and 20% for training and testing, respectively. No overlap is among training and testing data. And 25% of the training data are used for validation. We repeat this procedure 10 times to avoid performance bias.

Compared Methods Only NR methods are applicable for quality assessment of in-the-wild videos. We select five state-of-the-art NR methods for comparison, whose original codes are released by the authors, including VBLIINDS (Saad et al. 2014), VIIDEO (Mittal et al. 2016), BRISQUE (Mittal et al. 2012),[1] NIQE (Mittal et al. 2013), and CORNIA (Ye et al. 2012). Besides, we also show some relevant results reported from previous arts, *e.g.*, TLVQM (Korhonen 2019). Note that the method in Zhang et al. (2019c) needs scores of full-reference methods, methods in Kim et al. (2018) and Zhang et al. (2020) are full-reference methods, and thus they are unfeasible for this problem.

Basic Evaluation Criteria We follow the suggestion from Video Quality Experts Group (VQEG 2000), and report SROCC and PLCC as the criteria of prediction monotonicity and prediction accuracy, respectively. Better VQA methods should have larger values of SROCC and PLCC. When the predicted quality scores are not the same scale as the subjective scores, PLCC is calculated after nonlinear mapping with a 4-parameter logistic function as suggested by VQEG.

[1] Video-level features of BRISQUE are the average pooling of its frame-level features.

Table 1 Brief information of the four benchmark datasets, including the information of the videos and the information of the corresponding subjective study

Dataset	CVD2014 (Nuutinen et al. 2016)	KoNViD-1k (Hosu et al. 2017)	LIVE-Qualcomm (Ghadiyaram et al. 2018)	LIVE-VQC (Sinno and Bovik 2019a)
Number of videos	234	1200	208	585
Number of scenes	5	≈1200	54	≈585
Number of devices	78	–	8	101
Number of users	–	480	–	80
Video orientations	Landscape	Landscape	Landscape	Portrait, landscape
Video resolutions	1280 × 720 or 640 × 480	960 × 540	1920 × 1080	1920 × 1080 to 320 × 240
Number of resolutions	2	1	1	18
Frames per second	11 to 31	24, 25 or 30	30	19–30 (one 120)
Time span	10–25 s	8 s	15 s	10 s
Max video length	830 frames	240 frames	526 frames	1202 frames
Test methodology	Single stimulus	Single stimulus	Single stimulus	Single stimulus
Lab or crowdsourcing	Lab	Crowdsourcing	Lab	Crowdsourcing
Number of participants	210	642	39	4776
Number of ratings	28–33	>50, average 114	18	>200, average 240
Raw ratings provided	Yes	Yes	No	No
Mean opinion score	[−6.50, 93.38]	[1.22, 4.64]	[16.5621, 73.6428]	[6.2237, 94.2865]

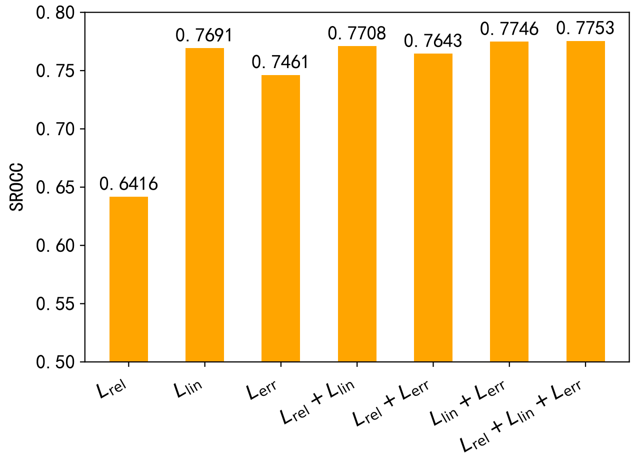


Fig. 5 Median SROCC results for different losses used in our mixed datasets training strategy

4.2 Effectiveness of Mixed Datasets Training Strategy

In this subsection, we conduct experiments to verify the effectiveness of our mixed datasets training strategy in the following four aspects. We first consider different loss combinations in our strategy. Then, we compare our strategy with the naïve linear re-scaling strategy. In the third and fourth aspects, we exploit whether our strategy helps further improving the performance with more training data available.

Different Loss Combinations To verify the effectiveness of the proposed losses, we compare different combinations of monotonicity-induced loss L_{rel}, linearity-induced loss L_{lin}, and error-induced loss L_{err}. We consider mixing all the four datasets (CVD2014, KoNViD-1k, LIVE-Qualcomm, and LIVE-VQC) in this experiment. Figure 5 shows the dataset-size weighted average of median SROCC results over the four datasets. It can be seen that the combination of three losses is better than that of two losses, and the combination of two losses is better than one of the two losses only. The three losses all contribute to the performance gain, but the contribution of linearity-induced loss is the largest.

Comparison with Linear Re-scaling To verify the effectiveness of our dataset-specific perceptual scale alignment, we compare it with the naïve linear re-scaling. Similar to the last experiment, all the four datasets (CVD2014, KoNViD-1k, LIVE-Qualcomm, and LIVE-VQC) are considered. And both our strategy and the linear re-scaling strategy use all three losses. They are compared with the models trained on one of the datasets, *i.e.*, "Trained only on CVD2014/KonViD-1k/LIVE-Qualcomm/LIVE-VQC". Figure 6 shows the dataset-size weighted average of median SROCC results over the four datasets. Models trained on the two larger datasets (KoNViD-1k and LIVE-VQC) achieve better performance than models trained on the two smaller datasets (CVD2014 and LIVE-Qualcomm). Linear rescaling strategy improves the performance to 0.7576, and our mixed datasets training strategy further improves the performance to 0.7753. The further performance gain is contributed by the dataset-specific perceptual scale alignment for avoiding the inconformity due to linear re-scaling.

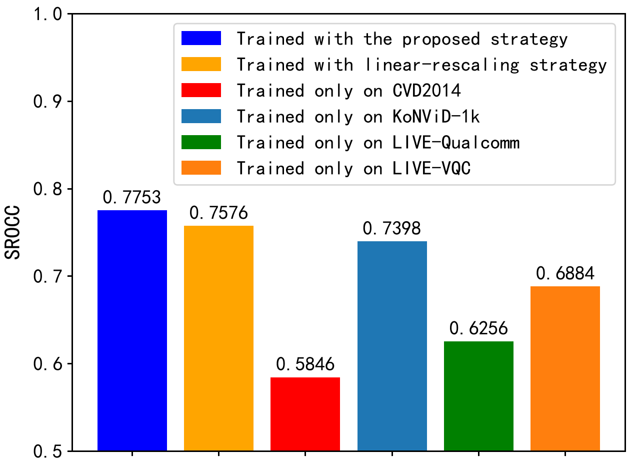


Fig. 6 Median SROCC results for models trained with our strategy and the linear re-scaling strategy in comparison with the models trained only on one of the datasets

Mixing More Datasets In this experiment, we explore the effect of mixing more datasets into the training data. Table 2 shows the median SROCC results in 10 runs for mixing different datasets. Each cell shows the base performance of the model trained on the train sets of D_B and tested on a test set of D_T, and the value in the brackets indicates the performance gain when the train set of D_+ is added into the training data. The "Overall Performance" is the dataset-size weighted average of median SROCC over these datasets. In general, the overall performance over the four datasets is improved in most cases. As for the performance on a single test set, there are mainly three scenarios.

- $D_T = D_+$: The performance values in this scenario, shown in the diagonal blocks of Table 2, all increase a lot. For example, when the train set of CVD2014 (D_+) is added into any train sets of D_B, the performance on the test set of CVD2014 (D_T) is improved (0.1479+ gain).
- $D_T \subseteq D_B$: The performance values in this scenario, marked in a light gray background, mostly decrease a little. For example, when the train set of CVD2014 (D_+) is added into the train set of KoNViD-1k (D_B), the performance on the test set of KoNViD-1k (D_T) drops 0.0067.
- $D_T \cap (D_B \cup D_+) = \emptyset$: The performance values in this scenario, marked in a dark gray background, may increase or decrease. For example, when the train set of LIVE-Qualcomm (D_+) is added into the train set of LIVE-VQC (D_B), the performance on the test set of CVD2014 (D_T) is improved while that on the test set of KoNViD-1k (D_T) drops.

This phenomenon is the consequence of the following two factors: (1) over-fitting problem during the training, (2) the discrepancy of data distribution between the train set and the

Table 2 Performance gain in terms of median SROCC when one more dataset is added into the training data

Train data		Test set of D_T				Overall
train sets of D_B	train set of D_+	CVD2014	KoNViD-1k	LIVE-Qualcomm	LIVE-VQC	Performance
KoNViD-1k	CVD2014	0.6474(+0.2078)	0.7809(-0.0067)	0.6732(-0.0248)	0.7160(-0.0227)	0.7398(+0.0100)
LIVE-Qualcomm		0.5879(+0.2757)	0.6128(+0.0563)	0.7538(+0.0344)	0.6214(-0.0061)	0.6256(+0.0609)
LIVE-VQC		0.4819(+0.3556)	0.7059(+0.0319)	0.6550(+0.0246)	0.7470(-0.0193)	0.6884(+0.0518)
KoNViD-1k+LIVE-Qualcomm		0.6933(+0.1479)	0.7836(-0.0177)	0.8170(-0.0012)	0.6969(-0.0118)	0.7544(+0.0028)
KoNViD-1k+LIVE-VQC		0.6325(+0.1978)	0.7974(-0.0115)	0.6995(+0.0018)	0.7461(-0.0018)	0.7575(+0.0143)
LIVE-Qualcomm+LIVE-VQC		0.5849(+0.2516)	0.6843(+0.0310)	0.8010(+0.0108)	0.7434(-0.0222)	0.7002(+0.0383)
KoNViD-1k+LIVE-Qualcomm+LIVE-VQC		0.6422(+0.1870)	0.7906(-0.0113)	0.8003(+0.0042)	0.7476(-0.0124)	0.7646(+0.0107)
CVD2014	KoNViD-1k	0.8747(-0.0195)	0.6051(+0.1692)	0.3919(+0.2565)	0.4950(+0.1983)	0.5846(+0.1652)
LIVE-Qualcomm		0.5879(+0.1054)	0.6128(+0.1708)	0.7538(+0.0631)	0.6214(+0.0755)	0.6256(+0.1288)
LIVE-VQC		0.4819(+0.1506)	0.7059(+0.0915)	0.6550(+0.0445)	0.7470(-0.0008)	0.6884(+0.0691)
CVD2014+LIVE-Qualcomm		0.8636(-0.0224)	0.6691(+0.0968)	0.7883(+0.0275)	0.6153(+0.0698)	0.6865(+0.0707)
CVD2014+LIVE-VQC		0.8375(-0.0072)	0.7378(+0.0482)	0.6795(+0.0217)	0.7276(+0.0167)	0.7401(+0.0316)
LIVE-Qualcomm+LIVE-VQC		0.5849(+0.0574)	0.6843(+0.1063)	0.8010(-0.0007)	0.7434(+0.0042)	0.7002(+0.0644)
CVD2014+LIVE-Qualcomm+LIVE-VQC		0.8364(-0.0072)	0.7152(+0.0641)	0.8118(-0.0073)	0.7211(+0.0141)	0.7385(+0.0368)
CVD2014	LIVE-Qualcomm	0.8747(-0.0111)	0.6051(+0.0640)	0.3919(+0.3963)	0.4950(+0.1203)	0.5846(+0.1019)
KoNViD-1k		0.6474(+0.0459)	0.7809(+0.0026)	0.6732(+0.1437)	0.7160(-0.0192)	0.7398(+0.0146)
LIVE-VQC		0.4819(+0.1030)	0.7059(-0.0216)	0.6550(+0.1460)	0.7470(-0.0036)	0.6884(+0.0119)
CVD2014+KoNViD-1k		0.8552(-0.0140)	0.7743(-0.0084)	0.6484(+0.1673)	0.6934(-0.0083)	0.7498(+0.0074)
CVD2014+LIVE-VQC		0.8375(-0.0010)	0.7378(-0.0226)	0.6795(+0.1322)	0.7276(-0.0065)	0.7401(-0.0016)
KoNViD-1k+LIVE-VQC		0.6325(+0.0098)	0.7974(-0.0068)	0.6995(+0.1008)	0.7461(+0.0014)	0.7575(+0.0072)
CVD2014+KoNViD-1k+LIVE-VQC		0.8303(-0.0010)	0.7860(-0.0066)	0.7012(+0.1032)	0.7443(-0.0092)	0.7718(+0.0036)
CVD2014	LIVE-VQC	0.8747(-0.0372)	0.6051(+0.1327)	0.3919(+0.2876)	0.4950(+0.2326)	0.5846(+0.1555)
KoNViD-1k		0.6474(-0.0149)	0.7809(+0.0165)	0.6732(+0.0263)	0.7160(+0.0301)	0.7398(+0.0177)
LIVE-Qualcomm		0.5879(-0.0030)	0.6128(+0.0715)	0.7538(+0.0472)	0.6214(+0.1220)	0.6256(+0.0747)
CVD2014+KoNViD-1k		0.8552(-0.0250)	0.7743(+0.0117)	0.6484(+0.0528)	0.6934(+0.0509)	0.7498(+0.0220)
CVD2014+LIVE-Qualcomm		0.8636(-0.0272)	0.6691(+0.0461)	0.7883(+0.0235)	0.6153(+0.1058)	0.6865(+0.0520)
KoNViD-1k+LIVE-Qualcomm		0.6933(-0.0511)	0.7836(+0.0070)	0.8170(-0.0167)	0.6969(+0.0507)	0.7544(+0.0102)
CVD2014+KoNViD-1k+LIVE-Qualcomm		0.8412(-0.0119)	0.7659(+0.0135)	0.8157(-0.0113)	0.6851(+0.0501)	0.7572(+0.0181)

D_+ is the added dataset for training, D_B indicates the base datasets for training before adding D_+, and D_T indicates the dataset for testing. "Overall Performance" is indicated by the dataset-size weighted average of median SROCC. Positive gain is shown in blue, while negative gain is shown in red. The performance values in the scenario $D_T \subseteq D_B$ are marked in a light gray background, and the performance values in the scenario $D_T \cap (D_B \cup D_+) = \emptyset$ are marked in a dark gray background

Table 3 The test performance of a model trained only on a single train set

SROCC	Test			
Train	CVD2014	KoNViD-1k	LIVE-Qualcomm	LIVE-VQC
CVD2014	**0.8747**	0.6051	0.3919	0.4950
KoNViD-1k	0.6474	**0.7809**	0.6732	0.7160
LIVE-Qualcomm	0.5879	0.6128	**0.7538**	0.6214
LIVE-VQC	0.4819	0.7059	0.6550	**0.7470**

Bold value indicates in each column shows the best SROCC values

Table 4 Cross dataset performance gain in terms of median SROCC when KoNViD-1k is added into the training data

Train data	Test dataset		
	CVD2014 (full)	LIVE-Qualcomm (full)	LIVE-VQC (full)
CVD2014 (+KoNViD-1k)	–	0.3390(+0.2579)	0.4751(+0.2128)
LIVE-Qualcomm (+KoNViD-1k)	0.4938(+0.1488)	–	0.5988(+0.0983)
LIVE-VQC (+KoNViD-1k)	0.4662(+0.1584)	0.5888(+0.0521)	–
CVD2014+LIVE-Qualcomm (+KoNViD-1k)	–	–	0.5984(+0.0821)
CVD2014+LIVE-VQC (+KoNViD-1k)	–	0.6459(+0.0087)	–
LIVE-Qualcomm+LIVE-VQC (+KoNViD-1k)	0.5069(+0.1178)	–	–

Note that the testing is conducted on the full dataset, including its train and test sets

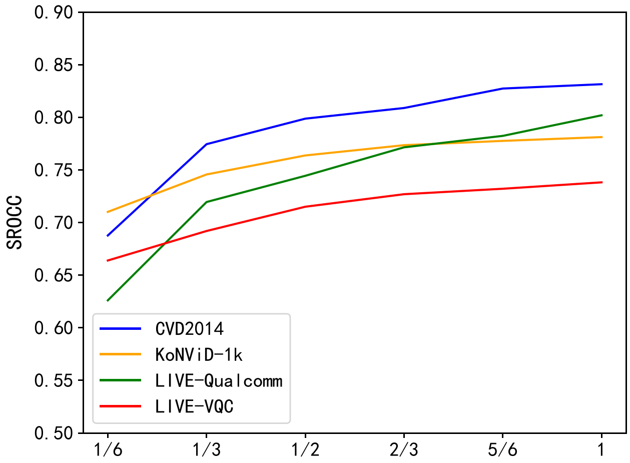


Fig. 7 Mean SROCC results under different training proportions when the model is trained by mixing all datasets

test set. Table 3 shows the performance of the model trained on a single train set and tested on a single test set, which can somehow reflects how well the trained dataset can represent the test set. In Table 3, the diagonal values are always the largest one in its column, *i.e.*, the most similar data set to a test set is its corresponding train set. Thus, adding the train set of D_+ to the train sets of D_B leads to a significant performance improvement on the test set of D_+, but a minor performance drop on the test sets of D_B. However, we can notice that adding one more train set to the LIVE-Qualcomm train set provides a performance gain on the LIVE-Qualcomm test set. This might be attributed to the fact that LIVE-Qualcomm is the smallest dataset among these four datasets and overfitting is most likely to happen during model training on LIVE-Qualcomm. Besides, the performance on the test set of an unseen dataset D_T depends on whether the train set of D_+ or D_B is more similar to the test set of D_T. In this regard, to improve the model performance on unseen datasets, it is critical to collect similar data for training. When datasets with similar data distribution to the test set are added into training data, it is more likely to learn the characteristics that are needed for assessing the quality of the test video in the wild. For example, in Table 4, when KoNViD-1k is added into the training data, the cross-dataset evaluation performance on the unseen dataset is improved.

Different Training Proportions In this experiment, we utilize different proportions of training data from the four datasets (LIVE-VQC, LIVE-Qualcomm, KoNViD-1k, and CVD2014) to train our VQA model with the proposed strategy. Figure 7 shows the test performance on the four datasets under different training proportions of the training data. The performance on each dataset increases as the training proportion increases. Our method can still achieve a good performance even when the training proportion is 1/2, which means only half of the training data are used for training. And the increasing trend indicates that the performance can still be improved when more training data are available.

Based on the above study, we have learned that our mixed datasets training strategy is effective. To sum up, it is helpful for learning characteristics from all datasets and thus improving the overall performance. It also has the potential benefits

Table 5 Overall performance comparison on CVD2014, KoNViD-1k, and LIVE-Qualcomm

Method	SROCC↑ mean (± std)	*p*-value based on SROCC	PLCC↑ mean (± std)	*p*-value based on PLCC
BRISQUE (Mittal et al. 2012)	0.6610 (± 0.0218)	9.6754E−09	0.6032 (± 0.0144)	4.7276E−10
NIQE (Mittal et al. 2013)	0.5255 (± 0.0479)	2.3066E−09	0.5396 (± 0.0430)	6.4720E−10
CORNIA (Ye et al. 2012)	0.5913 (± 0.0253)	5.4983E−10	0.5954 (± 0.0240)	5.0748E−10
VIIDEO (Mittal et al. 2016)	0.2368 (± 0.0595)	7.4623E−11	0.2351 (± 0.0574)	4.4222E−11
VBLIINDS (Saad et al. 2014)	0.6628 (± 0.0321)	7.7577E−08	0.6127 (± 0.0833)	5.1515E−05
TLVQM (Korhonen 2019)	0.77 (± 0.02)*	*	0.77 (± 0.02)*	*
LS-VSFA	0.7603 (± 0.0219)	4.0044E−07	0.7662 (± 0.0238)	1.9500E−06
MDTVSFA	0.7860 (± 0.0202)	–	0.7923 (± 0.0207)	–

Mean and standard deviation (std) of the dataset-size weighted performance values in 10 runs are reported, *i.e.*, mean (± std). The *p*-value is also reported, where $p < 0.001$ indicates our method MDTVSFA is significantly better than the method in that row
*The results are cited from Table VIII of the original paper (Korhonen 2019)
We can not calculate the *p*-value due to the lack of raw SROCC/PLCC values of TLVQM

for cross-dataset evaluation since the characteristics of the test videos are more likely to be learned, if more datasets with similar data distribution to the testing set are added into the training data. Besides, the performance can be further improved with more training data available.

4.3 Performance Comparison

In this section, we compare our method with the state-of-the-art NR methods. For VBLIINDS, BRISQUE and our method, we choose the models with the highest SROCC values on the validation set during the training phase. NIQE, CORNIA, and VIIDEO are tested on the same 20% testing data after fitting the four-parameter logistic function with the training data.

Overall Performance In this part, all the methods are trained using mixed datasets. Similar to Korhonen (2019), the other compared methods use the naïve linear re-scaling strategy. Our model trained with the naïve linear re-scaling strategy, denoted as LS-VSFA, does not learn the dataset-specific perceptual scale alignment and uses all three losses after linear re-scaling the subjective quality scores to the same range. We denote our VQA model trained with the proposed mixed datasets training strategy as MDTVSFA. Table 5 reports the overall performance over CVD2014, KoNViD-1k, and LIVE-Qualcomm, where the overall performance is measured by the dataset-size weighted performance values over the three datasets. We can see that our VQA model achieves the best performance in terms of prediction monotonicity (SROCC) and prediction accuracy (PLCC). The last two rows show that our proposed mixed datasets training strategy can achieve better performance than the naïve linear re-scaling strategy. We further carry out the statistical significance test to see whether these comparison results are statistical significant or not. On each dataset, the paired t-test is conducted at 1‰ significance level using the performance values (in 10 runs) of our method MDTVSFA and that of the compared one. The *p*-values are shown in Table 5. All *p*-values are far smaller than 0.001 and it proves that our method is significantly better than all the other methods.

Scatter Plot and Qualitative Examples To have an intuitive feeling, in Fig. 8, we visualize the scatter plots between the subjective scores and predicted scores for the five best-performed methods (excluding TLVQM, since we do not have its raw predictions) in the 10th run. Each row shows the scatter plots for a method. From top to down, the methods are BRISQUE, CORNIA, VBLIINDS, LS-VSFA, and MDTVSFA. The first, second, and third column show the scatter plots on CVD2014, KoNViD-1k, and LIVE-Qualcomm, respectively. In each sub-figure, the x-axis indicates the predicted score by the method while y-axis indicates the MOS. The scatter points are expected to be located at the diagonal line. We can see that the scatter plots for BRISQUE and CORNIA are more dispersive than the ones for VBLIINDS and our method. The scatter points for our method are more densely clustered around and centered at the diagonal line than the others.

In Figs. 9, 10 and 11 , we show several success and failure cases of our method. Figures 9 and 10 show the success cases of MDTVSFA, which means the predictions of MDTVSFA model is consistent with MOS. LS-VSFA has more failure cases than MDTVSFA since the linear re-scaling strategy disturbs the training process. We also show two failure cases of LS-VSFA in Figs. 9 and 10. Besides, there is still a large room for improving the performance of MDTVSFA, and we show a failure case of both MDTVSFA and LS-VSFA in Fig. 11. Such failure may be due to the fact that our models extract frame-level features and do not fully exploit the motion and spatial-temporal information. For example, our methods do

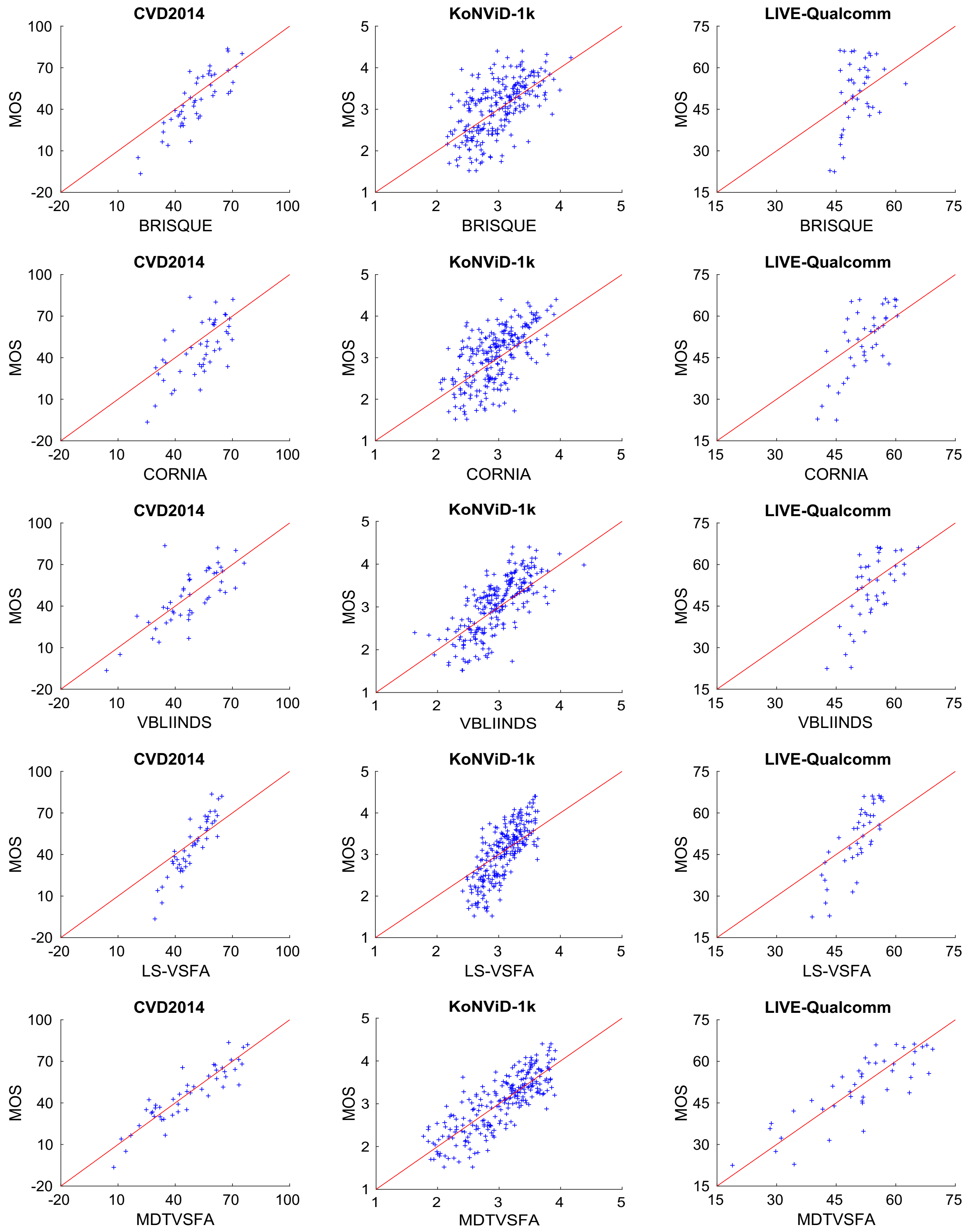
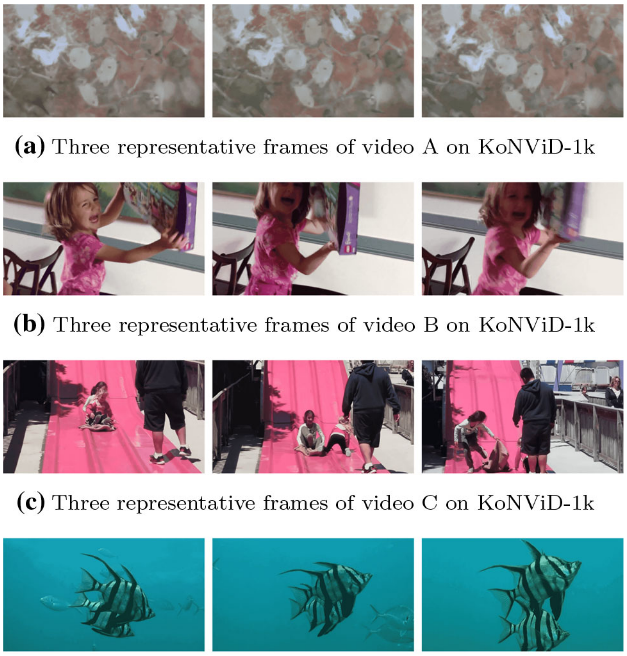


Fig. 8 Scatter plots for BRISQUE, CORNIA, VBLIINDS, LS-VSFA, and MDTVSFA on CVD2014, KoNViD-1k, and LIVE-Qualcomm datasets. The predictions of MDTVSFA shows the best correlation with the mean opinion scores (MOSs) across the datasets



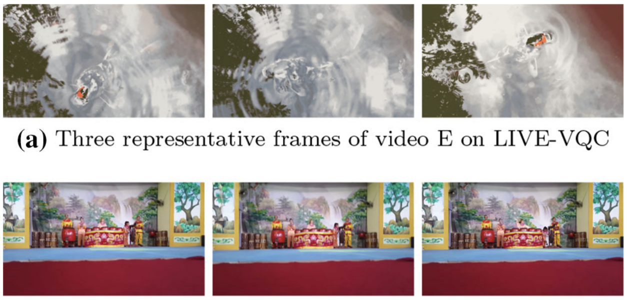
(a) Three representative frames of video A on KoNViD-1k

(b) Three representative frames of video B on KoNViD-1k

(c) Three representative frames of video C on KoNViD-1k

(d) Three representative frames of video D on KoNViD-1k

Fig. 9 Qualitative example on KoNViD-1k test set. The quality rankings provided by MOS and MDTVSFA are both A<B<C<D, but LS-VSFA gives a quality ranking of A<C<B<D. Full-resolution videos are provided in https://bit.ly/3csmHYk



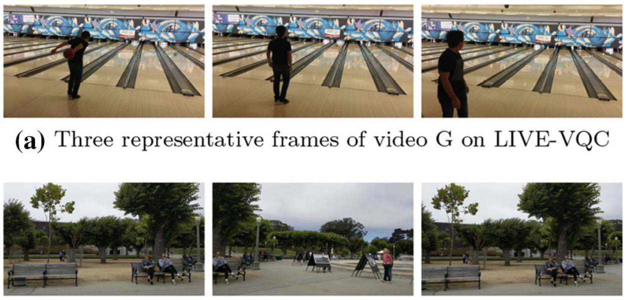
(a) Three representative frames of video E on LIVE-VQC

(b) Three representative frames of video F on LIVE-VQC

Fig. 10 Qualitative example on LIVE-VQC. The quality rankings provided by MOS and MDTVSFA are both E>F, but LS-VSFA gives a quality ranking of E<F. Full-resolution videos are provided in https://bit.ly/3csmHYk

not account for the discomfort caused by suddenly and fast scene change.

Performance on Individual Datasets Besides the overall performance reported in last part, we report performance on individual datasets in this part. Our method is trained by mixing the four datasets while other methods are trained on individual datasets. Table 6 summarizes the performance values on the four datasets individually. The results provided by our method are based only on a single unified model while the results provided by other methods are based on different models trained for different datasets. The natural scene statistics (NSS)-based NR-IQA methods (such as BRISQUE) outperform VIIDEO. This may be owing to the fact that VIIDEO is based only on temporal scene statistics and cannot model the complex distortions. VBLIINDS and TLVQM rely on a lot of carefully-designed handcrafted features that capture the spatial and temporal distortions, and thus they achieve a better performance than the NR-IQA methods and VIIDEO. Our method achieves the best performance in terms of prediction monotonicity (SROCC) and prediction accuracy (PLCC) on the three datasets (LIVE-VQC, LIVE-Qualcomm, and KoNViD-1k). On CVD2014, MDTVSFA slightly outperforms TLVQM in terms of SROCC, while it slightly underperforms TLVQM in terms of PLCC. However, we should note that the results of our method is based only on one single model, which indicates our unified model performs well across datasets.



(a) Three representative frames of video G on LIVE-VQC

(b) Three representative frames of video H on LIVE-VQC

Fig. 11 Another qualitative example on LIVE-VQC. The quality rankings provided by LS-VSFA and MDTVSFA are both G<H, but MOS gives a quality ranking of G>H. Note that the scenes change fast in video H, where full-resolution videos are provided in https://bit.ly/3csmHYk

We further prove the above statement by conducting experiments to compare the models trained by mixing CVD2014, KoNViD-1k, and LIVE-Qualcomm datasets with the models trained on one of the datasets. Table 7 shows the median SROCC of different models on the three datasets. We can see that, no matter which model it is, the unified model trained by mixing all datasets achieves better overall performance than the model trained on one of the datasets. And our model trained with our proposed strategy achieves better overall performance across the datasets than the other models (VBLIINDS, BRISQUE, and TLVQM) trained with the linear re-scaling strategy. Among these datasets, the size of LIVE-Qualcomm dataset is the smallest one. And our model trained only on LIVE-Qualcomm dataset suffered from over-fitting problem. In such situation, mixed datasets training helps alleviating the problem to some extent. So a performance improvement of the proposed model with mixed dataset training is found on LIVE-Qualcomm dataset. This verifies the necessity of mixed datasets training and the effectiveness of our mixed datasets training strategy.

Table 6 Performance comparison on the four VQA datasets individually

Method	LIVE-VQC (Sinno and Bovik 2019a)		LIVE-Qualcomm (Ghadiyaram et al. 2018)	
	SROCC↑	PLCC↑	SROCC↑	PLCC↑
BRISQUE (Mittal et al. 2012)	0.5687 (± 0.0729)	0.5868 (± 0.0642)	0.5036 (± 0.1470)	0.5158 (± 0.1274)
NIQE (Mittal et al. 2013)	0.5892 (± 0.0538)	0.6112 (± 0.0554)	0.4628 (± 0.1052)	0.4638 (± 0.1362)
CORNIA (Ye et al. 2012)	0.5953 (± 0.0170)	0.5926 (± 0.0230)	0.4598 (± 0.1299)	0.4941 (± 0.1327)
VIIDEO (Mittal et al. 2016)	0.1498 (± 0.0995)	0.2454 (± 0.0740)	0.1267 (± 0.1368)	−0.0012 (± 0.1062)
VBLIINDS (Saad et al. 2014)	0.7015 (± 0.0483)	0.7120 (± 0.0501)	0.5659 (± 0.0780)	0.5676 (± 0.0885)
ST-Naturalness (Sinno and Bovik 2019b)	0.5994*	0.6069*	–	–
3D-CNN+LSTM (You and Korhonen 2019)	–	–	0.687*	0.792*
FRIQUEE (Ghadiyaram and Bovik 2017)	–	–	0.6795*	0.7349*
TLVQM (Korhonen 2019)	–	–	0.78 (± 0.07)*	0.81 (± 0.06)*
MDTVSFA	**0.7382** (± 0.0357)	**0.7728** (± 0.0351)	**0.8019** (± 0.0295)	**0.8218** (± 0.0374)

Method	KoNViD-1k (Hosu et al. 2017)		CVD2014 (Nuutinen et al. 2016)	
	SROCC↑	PLCC↑	SROCC↑	PLCC↑
BRISQUE (Mittal et al. 2012)	0.6540 (± 0.0418)	0.6256 (± 0.0407)	0.7086 (± 0.0666)	0.7154 (± 0.0476)
NIQE (Mittal et al. 2013)	0.5435 (± 0.0396)	0.5456 (± 0.0376)	0.4890 (± 0.0908)	0.5931 (± 0.0650)
CORNIA (Ye et al. 2012)	0.6096 (± 0.0343)	0.6075 (± 0.0318)	0.6140 (± 0.0754)	0.6178 (± 0.0792)
VIIDEO (Mittal et al. 2016)	0.2976 (± 0.0522)	0.3026 (± 0.0486)	0.0228 (± 0.1216)	−0.0249 (± 0.1439)
VBLIINDS (Saad et al. 2014)	0.6947 (± 0.0239)	0.6576 (± 0.0254)	0.7458 (± 0.0564)	0.7525 (± 0.0528)
FC Model (Men et al. 2017)	0.572*	0.565*	–	–
STFC Model (Men et al. 2018)	0.606*	0.639*	–	–
STS-CNN200 (Yan and Mou 2019)	0.735*	–	–	–
TLVQM (Korhonen 2019)	**0.78** (± 0.02)*	0.77 (± 0.02)*	**0.83** (± 0.04)*	**0.85** (± 0.04)*
MDTVSFA	**0.7812** (± 0.0278)	**0.7856** (± 0.0240)	**0.8314** (± 0.0416)	0.8407 (± 0.0296)

Mean and standard deviation (std) of the performance values in 10 runs are reported, *i.e.*, mean (± std). In each column, the best mean SROCC and PLCC values are marked in boldface, and the second-best performance values are underlined

*The reported results in their original papers are shown here for reference

4.4 Computational Efficiency

Besides the performance, computational efficiency is also crucial for NR-VQA methods. To provide a fair comparison for the computational efficiency of different methods, all tests are carried out on the same desktop computer with Intel Core i7-6700K CPU@4.00 GHz, 12G NVIDIA TITAN Xp GPU, and 64 GB RAM. The operating system is Ubuntu 14.04. The compared methods are implemented with MATLAB R2016b while our method is implemented with Python 3.6. We use the default settings of the original codes without any modification. We select two videos with different lengths and different resolutions for testing. The tests are run in a separate environment and repeated ten times. The logarithm (with base 10) of the average computation time (seconds) for each method is shown in Fig. 12. The point near the left is the fast one, and the point near the top is the good-performed one. Our method (CPU version) is faster than VBLIINDS—the method with the third-best performance. TLVQM, the second-best performed method, considers two-level features, *i.e.*, low-complexity features for all frames and high-level features for only selected representative frames. It achieves a good trade-off between the performance and computational efficiency. It is worth mentioning that our method can be accelerated to **30x faster or more** (The larger resolution and length the video has, the faster acceleration is) by simply switching the CPU mode to the GPU mode. With the GPU available, our method (GPU version) is at the upper-left, and thus it is the fastest one as well as the best-performed one. To further improve the computational efficiency, we may resort to the light-weight networks.

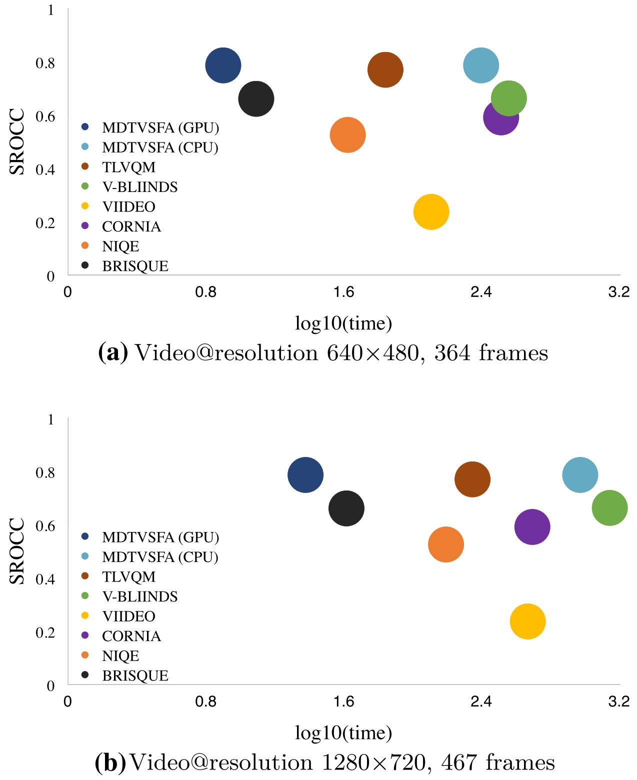
5 Conclusion and Future Work

In this work, we propose a novel unified NR-VQA framework with a mixed datasets training strategy for in-the-wild videos.

Table 7 Performance comparison in terms of median SROCC between the single models trained by mixing all three datasets (CVD2014, KoNViD-1k, and LIVE-Qualcomm) and the models trained on one of the datasets

Model	Train data	Mixed datasets training	Test dataset			Overall
			CVD2014	KoNViD-1k	LIVE-Qualcomm	Performance
BRISQUE	CVD2014	No	0.7582	0.5574	0.4632	0.5794
	KoNViD-1k	No	0.5388	0.6191	0.3019	0.5621
	LIVE-Qualcomm	No	0.3930	0.2341	0.5023	0.2973
	All three datasets	Linear re-scaling	0.7356	0.6300	0.3809	0.6107
VBLIINDS	CVD2014	No	0.7892	0.5787	0.4170	0.5864
	KoNViD-1k	No	0.5681	0.7078	0.4583	0.6544
	LIVE-Qualcomm	No	0.5027	0.5432	0.6018	0.5544
	All three datasets	Linear re-scaling	0.6749	0.6890	0.4684	0.6640
TLVQM	CVD2014	No	0.83*	0.54*	0.38*	–
	KoNViD-1k	No	<0.62*	**0.78***	<0.49*	–
	LIVE-Qualcomm	No	<0.36*	<0.38*	0.788*	–
	All three datasets	Linear re-scaling	–	–	–	0.77*
Our model	CVD2014	No	**0.8747**	0.6051	0.3919	0.6165
	KoNViD-1k	No	0.6474	**0.7809**	0.6732	0.7483
	LIVE-Qualcomm	No	0.5879	0.6128	0.7538	0.6271
	All three datasets	Our strategy	0.8412	0.7659	**0.8157**	**0.7829**

Overall performance indicates the dataset-size weighted median SROCC values in 10 runs. For each column, the largest value is marked in boldface
*The reported SROCC results in the original paper (Korhonen 2019) are shown here for reference
The "<" relation is inferred from the Table VII of Korhonen (2019). "–" indicates that the results are not reported



(a) Video@resolution 640×480, 364 frames

(b) Video@resolution 1280×720, 467 frames

Fig. 12 Bubble charts with the overall performance (mean SROCC values in Table 5) and the logarithm of average computation time (s) on videos with different resolutions and different lengths

The backbone model is a deep neural network designed for characterizing the two eminent effects of HVS, *i.e.*, content-dependency and temporal-memory effects. We enable mixed datasets training by designing two losses (monotonicity-induced loss, linearity-induced loss) for predicting relative quality and perceptual quality, and assigning dataset-specific perceptual scale alignment layers for predicting subjective quality. Our proposed method is compared with the state-of-the-art methods on four publicly available in-the-wild VQA datasets (CVD2014, KoNViD-1k, LIVE-Qualcomm, and LIVE-VQC). Experiments show the superior performance of our method and also verify the effectiveness of our unified VQA model with the mixed datasets training strategy.

However, our mixed datasets training strategy needs to re-train the unified VQA model every time when a new dataset is added to the training data. This will increase the burden of training. In the further study, we will explore lifelong learning for this task. Also, besides video capture, we intend to provide a unified and efficient VQA framework that can handle the whole chain-flow of video production. Moreover, some meta information that is crucial for the video quality, like video resolution, can be used as extra features for improving the model performance. Finally, we intend to apply our unified VQA model for practical computer vision applications such as video enhancement.

Acknowledgements This work was partially supported by the Natural Science Foundation of China under contracts 61572042, 61520106004, and 61527804. This work was also supported in part by National Key R&D Program of China (2018YFB1403900). We acknowledge the High-Performance Computing Platform of Peking University for providing computational resources.

Compliance with Ethical Standards

Conflict of interest The authors declare no conflict of interest.

References

Bampis, C. G., Li, Z., Moorthy, A. K., Katsavounidis, I., Aaron, A., & Bovik, A. C. (2017). Study of temporal effects on subjective video quality of experience. *IEEE Transactions on Image Processing*, *26*(11), 5217–5231.

Barron, J. T. (2019). A general and adaptive robust loss function. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 4331–4339.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

Choi, L. K., & Bovik, A. C. (2018). Video quality assessment accounting for temporal visual masking of local flicker. *Signal Processing: Image Communication*, *67*, 182–198.

Deng, J., Dong, W., Socher, R., Li, LJ., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 248–255.

Dodge, S., & Karam, L. (2016). Understanding how image quality affects deep neural networks. In *International conference on quality of multimedia experience (QoMEX)*, pp. 1–6.

Freitas, P. G., Akamine, W. Y., & Farias, M. C. (2018). Using multiple spatio-temporal features to estimate video quality. *Signal Processing: Image Communication*, *64*, 1–10.

Ghadiyaram, D., & Bovik, A. C. (2017). Perceptual quality prediction on authentically distorted images using a bag of features approach. *Journal of Vision*, *17*(1), 32–32.

Ghadiyaram, D., Pan, J., Bovik, A. C., Moorthy, A. K., Panda, P., & Yang, K. C. (2018). In-capture mobile video distortions: A study of subjective behavior and objective algorithms. *IEEE Transactions on Circuits and Systems for Video Technology*, *28*(9), 2061–2077.

He, H., Zhang, J., Zhang, Q., & Tao, D. (2019). Grapy-ML: Graph pyramid mutual learning for cross-dataset human parsing. arXiv preprint arXiv:1911.12053.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 770–778.

Hosu, V., Hahn, F., Jenadeleh, M., Lin, H., Men, H., Szirányi, T., Li, S., & Saupe, D. (2017). The Konstanz natural video database (KoNViD-1k). In *International conference on quality of multimedia experience (QoMEX)*, pp. 1–6.

Isogawa, M., Mikami, D., Takahashi, K., Iwai, D., Sato, K., & Kimata, H. (2019). Which is the better inpainted image? Training data generation without any manual operations. *International Journal of Computer Vision*, *127*(11–12), 1751–1766.

Juluri, P., Tamarapalli, V., & Medhi, D. (2015). Measurement of quality of experience of video-on-demand services: A survey. *IEEE Communications Surveys and Tutorials*, *18*(1), 401–418.

Kang, L., Ye, P., Li, Y., & Doermann, D. (2014). Convolutional neural networks for no-reference image quality assessment. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 1733–1740.

Kim, W., Kim, J., Ahn, S., Kim, J., & Lee, S. (2018). Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network. In *European conference on computer vision (ECCV)*, pp. 219–234.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

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

Krasula, L., Yoann, B., & Le Callet, P. (2020). Training objective image and video quality estimators using multiple databases. *IEEE Transactions on Multimedia*, *22*(4), 961–969.

Lasinger, K., Ranftl, R., Schindler, K., & Koltun, V. (2019). Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. arXiv preprint arXiv:1907.01341.

Li, D., Jiang, T., & Jiang, M. (2019a). Quality assessment of in-the-wild videos. In *ACM international conference on multimedia (MM)*, pp. 2351–2359.

Li, D., Jiang, T., Lin, W., & Jiang, M. (2019b). Which has better visual quality: The clear blue sky or a blurry animal? *IEEE Transactions on Multimedia*, *21*(5), 1221–1234.

Li, D., Jiang, T., & Jiang, M. (2020). Norm-in-norm loss with faster convergence and better performance for image quality assessment. In *ACM International conference on multimedia (MM)*, pp. 789–797.

Li, X., Guo, Q., & Lu, X. (2016a). Spatiotemporal statistics for video quality assessment. *IEEE Transactions on Image Processing*, *25*(7), 3329–3342.

Li, Y., Po, L. M., Cheung, C. H., Xu, X., Feng, L., Yuan, F., et al. (2016b). No-reference video quality assessment with 3D shearlet transform and convolutional neural networks. *IEEE Transactions on Circuits and Systems for Video Technology*, *26*(6), 1044–1057.

Li, YJ., Lin, CS., Lin, YB., & Wang, YCF. (2019c). Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation. In *IEEE international conference on computer vision (ICCV)*, pp. 7919–7929.

Lin, KY., & Wang, G. (2018). Hallucinated-IQA: No-reference image quality assessment via adversarial learning. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 732–741.

Liu, W., Duanmu, Z., & Wang, Z. (2018). End-to-end blind quality assessment of compressed videos using deep neural networks. In *ACM international conference on multimedia (MM)*, pp. 546–554.

Liu, X., van de Weijer, J., & Bagdanov, A. D. (2017). RankIQA: Learning from rankings for no-reference image quality assessment. In *IEEE international conference on computer vision (ICCV)*, pp. 1040–1049.

Lu, W., He, R., Yang, J., Jia, C., & Gao, X. (2019). A spatiotemporal model of video quality assessment via 3D gradient differencing. *Information Sciences*, *478*, 141–151.

Lv, J., Chen, W., Li, Q., & Yang, C. (2018). Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 7948–7956.

Ma, K., Wu, Q., Wang, Z., Duanmu, Z., Yong, H., Li, H., & Zhang, L. (2016). Group MAD competition—a new methodology to compare objective image quality models. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 1664–1673.

Ma, K., Duanmu, Z., & Wang, Z. (2018). Geometric transformation invariant image quality assessment using convolutional neural networks. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 6732–6736.

Manasa, K., & Channappayya, S. S. (2016). An optical flow-based no-reference video quality assessment algorithm. In *IEEE international conference on image processing (ICIP)*, 1 pp. 2400–2404.

Men, H., Lin, H., & Saupe, D. (2017). Empirical evaluation of no-reference VQA methods on a natural video quality database. In *International conference on quality of multimedia experience (QoMEX)*, pp. 1–3.

Men, H., Lin, H., & Saupe, D. (2018). Spatiotemporal feature combination model for no-reference video quality assessment. In *International conference on quality of multimedia experience (QoMEX)*, pp. 1–3.

Mittal, A., Moorthy, A. K., & Bovik, A. C. (2012). No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, *21*(12), 4695–4708.

Mittal, A., Soundararajan, R., & Bovik, A. C. (2013). Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, *20*(3), 209–212.

Mittal, A., Saad, M. A., & Bovik, A. C. (2016). A completely blind video integrity oracle. *IEEE Transactions on Image Processing*, *25*(1), 289–300.

Moorthy, A. K., Choi, L. K., Bovik, A. C., & De Veciana, G. (2012). Video quality assessment on mobile devices: Subjective, behavioral and objective studies. *IEEE Journal of Selected Topics in Signal Processing*, *6*(6), 652–671.

Nieto, RG., Restrepo, HDB., & Cabezas, I. (2019). How video object tracking is affected by in-capture distortions? In *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 2227–2231.

Nuutinen, M., Virtanen, T., Vaahteranoksa, M., Vuori, T., Oittinen, P., & Häkkinen, J. (2016). CVD2014–a database for evaluating no-reference video quality assessment algorithms. *IEEE Transactions on Image Processing*, *25*(7), 3073–3086.

Park, J., Seshadrinathan, K., Lee, S., & Bovik, A. C. (2013). Video quality pooling adaptive to perceptual distortion severity. *IEEE Transactions on Image Processing*, *22*(2), 610–620.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems (NeurIPS)*, pp. 8024–8035.

Rippel, O., Nair, S., Lew, C., Branson, S., Anderson, AG., & Bourdev, L. (2019). Learned video compression. In *IEEE international conference on computer vision (ICCV)*, pp. 3454–3463.

Saad, M. A., Bovik, A. C., & Charrier, C. (2014). Blind prediction of natural video quality. *IEEE Transactions on Image Processing*, *23*(3), 1352–1365.

Seshadrinathan, K., & Bovik, A. C. (2010). Motion tuned spatio-temporal quality assessment of natural videos. *IEEE Transactions on Image Processing*, *19*(2), 335–350.

Seshadrinathan, K., & Bovik, AC. (2011). Temporal hysteresis model of time varying subjective video quality. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 1153–1156.

Seshadrinathan, K., Soundararajan, R., Bovik, A. C., & Cormack, L. K. (2010). Study of subjective and objective quality assessment of video. *IEEE Transactions on Image Processing*, *19*(6), 1427–1441.

Seufert, M., Egger, S., Slanina, M., Zinner, T., Hoßfeld, T., & Tran-Gia, P. (2014). A survey on quality of experience of HTTP adaptive streaming. *IEEE Communications Surveys and Tutorials*, *17*(1), 469–492.

Siahaan, E., Hanjalic, A., & Redi, J. A. (2018). Semantic-aware blind image quality assessment. *Signal Processing: Image Communication*, *60*, 237–252.

Sinno, Z., & Bovik, A. C. (2019a). Large scale study of perceptual video quality. *IEEE Transactions on Image Processing*, *28*(2), 612–627.

Sinno, Z., & Bovik, AC. (2019b). Spatio-temporal measures of naturalness. In *IEEE international conference on image processing (ICIP)*, pp. 1750–1754.

Triantaphillidou, S., Allen, E., & Jacobson, R. (2007). Image quality comparison between JPEG and JPEG2000. II. Scene dependency, scene analysis, and classification. *Journal of Imaging Science and Technology*, *51*(3), 259–270.

Varga, D. (2019). No-reference video quality assessment based on the temporal pooling of deep features. *Neural Processing Letters*, *50*, 2595–2608.

Varga, D., & Szirányi, T. (2019). No-reference video quality assessment via pretrained CNN and LSTM networks. *Signal, Image and Video Processing*, *13*, 1569–1576.

VQEG. (2000). Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment. https://www.its.bldrdoc.gov/media/8212/frtv_phase1_final_report.doc.

Wang, H., Katsavounidis, I., Zhou, J., Park, J., Lei, S., Zhou, X., et al. (2017). VideoSet: A large-scale compressed video quality dataset based on JND measurement. *Journal of Visual Communication and Image Representation*, *46*, 292–302.

Wang, Y., Jiang, T., Ma, S., & Gao, W. (2012). Novel spatio-temporal structural information based video quality metric. *IEEE Transactions on Circuits and Systems for Video Technology*, *22*(7), 989–998.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004a). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612.

Wang, Z., Lu, L., & Bovik, A. C. (2004b). Video quality assessment based on structural distortion measurement. *Signal Processing: Image Communication*, *19*(2), 121–132.

Xu, J., Ye, P., Liu, Y., & Doermann, D. (2014). No-reference video quality assessment via feature learning. In *IEEE international conference on image processing (ICIP)*, pp. 491–495.

Yan, P., & Mou, X. (2019). No-reference video quality assessment based on spatiotemporal slice images and deep convolutional neural networks. In *Proc. SPIE 11187, Optoelectronic Imaging and Multimedia Technology* VI, pp. 74–83.

Yang D, Peltoketo VT, Kamarainen JK (2019) CNN-based cross-dataset no-reference image quality assessment. In *ieee international conference on computer vision workshop (ICCVW)*, pp. 3913–3921

Ye, P., Kumar, J., Kang, L., & Doermann, D. (2012). Unsupervised feature learning framework for no-reference image quality assessment. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 1098–1105.

You, J., & Korhonen, J. (2019). Deep neural networks for no-reference video quality assessment. In *IEEE international conference on image processing (ICIP)*, pp. 2349–2353.

You, J., Ebrahimi, T., & Perkis, A. (2014). Attention driven foveated video quality assessment. *IEEE Transactions on Image Processing*, *23*(1), 200–213.

Zhang, L., Shen, Y., & Li, H. (2014). VSI: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image Processing*, *23*(10), 4270–4281.

Zhang, R., Isola, P., Efros, AA., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 586–595.

Zhang, W., & Liu, H. (2017). Study of saliency in objective video quality assessment. *IEEE Transactions on Image Processing*, *26*(3), 1275–1288.

Zhang, W., Liu, Y., Dong, C., & Qiao, Y. (2019a). RankSRGAN: Generative adversarial networks with ranker for image super-resolution. In *IEEE international conference on computer vision (ICCV)*, pp. 3096–3105.

Zhang, W., Ma, K., & Yang, X. (2019b). Learning to blindly assess image quality in the laboratory and wild. arXiv preprint arXiv:1907.00516.

Zhang, Y., Gao, X., He, L., Lu, W., & He, R. (2019c). Blind video quality assessment with weakly supervised learning and resampling strategy. *IEEE Transactions on Circuits and Systems for Video Technology*, *29*(8), 2244–2255.

Zhang, Y., Gao, X., He, L., Lu, W., & He, R. (2020). Objective video quality assessment combining transfer learning with CNN. *IEEE Transactions on Neural Networks and Learning Systems*, *31*(8), 2716–2730.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.