RankDVQA: Deep VQA based on Ranking-inspired Hybrid Training

Chen Feng, Duolikun Danier, Fan Zhang and David Bull
Visual Information Laboratory, University of Bristol, Bristol, UK, BS1 5DD
{chen.feng, duolikun.danier, fan.zhang, dave.bull}@bristol.ac.uk

Abstract

In recent years, deep learning techniques have shown significant potential for improving video quality assessment (VQA), achieving higher correlation with subjective opinions compared to conventional approaches. However, the development of deep VQA methods has been constrained by the limited availability of large-scale training databases and ineffective training methodologies. As a result, it is difficult for deep VQA approaches to achieve consistently superior performance and model generalization. In this context, this paper proposes new VQA methods based on a two-stage training methodology which motivates us to develop a large-scale VQA training database without employing human subjects to provide ground truth labels. This method was used to train a new transformer-based network architecture, exploiting quality ranking of different distorted sequences rather than minimizing the difference from the ground-truth quality labels. The resulting deep VQA methods (for both full reference and no reference scenarios), FR- and NR-RankDVQA, exhibit consistently higher correlation with perceptual quality compared to the state-of-the-art conventional and deep VQA methods, with average SROCC values of 0.8972 (FR) and 0.7791 (NR) over eight test sets without performing cross-validation. The source code of the proposed quality metrics and the large training database are available at https://chenfeng-bristol.github.io/RankDVQA.

1. Introduction

With the explosion of video streaming and conferencing services, there has been a surge in the prevalence of video content on the internet. It is reported that videos account for 82% of the global internet traffic, with 4.8 Zetabytes of video data being transmitted annually [8]. Given such increasing demand, it is important that the quality of the transmitted videos matches the requirements of the service provider and the expectations of the end-user. To this end, video quality assessment (VQA) methods are employed to provide an objective measure of the perceived video quality. They therefore represent a crucial component in the video coding process, enabling operating points exhibiting appropriate trade offs between video quality and bit rate. Other than video compression, VQA methods are also used to assess the performance of various processing tasks such as denoising [33, 56], restoration [31, 64], frame interpolation [9, 15] and super-resolution [36, 37].

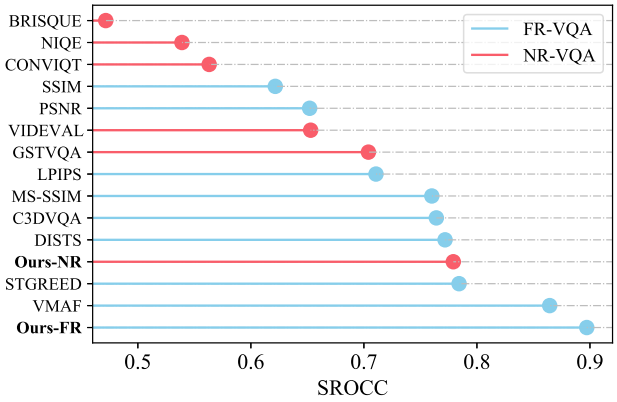

Figure 1. The performance of the proposed FR- and NR-RankDVQA and selected benchmark VQA methods. Here the average SROCC values were used as the statistical parameter to measure their correlation with subjective ground truth. Both FR- and NR-RankDVQA achieve the best performance in each category (FR and NR), and NR-RankDVQA even outperforms some FR quality metrics. The results for all the deep VQA methods shown here are based on the fixed model parameters without performing re-training on each database (intra database cross-validation).

Conventional VQA approaches rely on signal processing techniques to measure distortions. Notable examples include PSNR, SSIM [65] and its variants [50, 66, 67], VIF [55] and VIIDEO [45], VBLIINDS [51]. With the increasing popularity of learning-based techniques, the handcrafted features in these metrics have also been enhanced using machine learning models such as Support Vector Regressors to achieve improved prediction performance, e.g., VMAF [30]. More recently, deep learning has driven the development of video quality assessment methods [22, 71], achieving promising results when compared to classical ap-

proaches. However, deep VQA methods tend to be constrained by the following issues. (i) **The lack of reliable large and diverse training databases**: Most existing methods were trained using a relatively small video database (typically with only a few hundreds subjective labels)[1]. This is widely acknowledged to be insufficient for training a model with a relatively high network capacity. This is why most existing deep VQA methods only perform well when cross-validated on a single database yet show unsatisfactory cross-dataset generalization ability, with inconsistent overall performance compared to conventional or regression-based VQA methods, such as VMAF [30]. (ii) **Ineffective training methods**: The training processes used were designed to consider one distorted content at a time. In contrast, the ability of a metric to differentiate the quality of differently distorted versions of the same (or different) source content is an important characteristic of a VQA method which has not yet been exploited.

To address the above issues, in this paper, we propose a ranking-based hybrid training methodology for deep VQA. The main contributions of this work are summarized below.

- This paper proposes a **new two-stage training methodology**, which combines patch-wise VQA with spatio-temporal pooling. This first uses VMAF-based quality ranking information to train a deep VQA model, and then employs multiple small video databases with subjective ground truth to train an aggregation network for spatio-temporal pooling. By using this new training framework, the proposed deep VQA methods can achieve significantly improved model generalization, and avoid the need to perform intra-database cross-validation which is not a practical evaluation method for real-world applications.
- More importantly, by using VMAF to generate ranking-based labels, we have, for the first time, developed a **large-scale training dataset** (204,800 video dodecuplet groups with VMAF ranking-based annotations) for optimizing patch-level VQA models (Stage 1) without performing costly subjective tests. This solution circumvents problem (i) - the lack of reliable large training databases.
- During the training processes in both stages, we utilize the **quality ranking information associated with different distorted videos** (from the same or different source content). As such, we reformulate the primary aim of VQA as differentiating the quality of two distorted videos, rather than providing an absolute index for an individual sequence. This takes account of issue (ii) - ineffective training methods, and for the first time enables the combination of multiple VQA databases for training (in Stage 2). To the best of our knowledge, we are the first to adopt ranking-based training for VQA and extend it by additionally considering dual-source ranking information.
- Based on the new training methodology, we have trained a full reference (FR) and a no-reference (NR) deep VQA method (RankDVQA), employing a transformer-based architecture (for Stage 1) and a CNN-based aggregation network (for Stage 2). Through comprehensive evaluation on eight commonly used video quality datasets, we show that both proposed metrics (FR-and NR-RankDVQA) outperform their competing approaches including VMAF (for the full reference metric) without re-training or cross-validation on each database. This is shown in Figure 1, where the average SROCC values are 0.8972 (FR) and 0.7791 (NR). To the best of our knowledge, the proposed FR quality metric is the first that consistently outperforms VMAF on various video quality databases without re-training.

[1]Note that different VQA databases cannot be simply combined because of different experimental settings during subjective data collection.

2. Related Work

Depending on the availability of reference content, VQA methods can be classified into three sub-groups: full reference (FR), reduced reference (RR) and no reference (NR). RR and NR models require either partial or no information about the reference video, while FR approaches use both the reference and the distorted video content as inputs. The work mainly focuses on the FR and NR scenarios.

FR VQA methods. Among existing FR VQA methods, PSNR and SSIM [65] are the most widely used approaches, primarily due to their simplicity and low computational complexity. However, to achieve better correlation with subjective opinions, many perceptually-optimized quality metrics have been proposed. These include SSIM variants [50, 66, 67], MAD [62], STMAD [63], MOVIE [53] and PVM [74]. Netflix has developed a machine learning based quality metric, VMAF [30], combining six different video features using a Support Vector Regressor (SVR), which provides consistently better correlation with subjective opinions over a wide range of content and distortion types than most conventional VQA methods.

In recent years, deep neural networks have been increasingly applied to image [1, 4, 21] and video [6, 10, 22, 40, 42, 69, 71] quality assessment. Such methods have demonstrated the potential to compete with conventional quality metrics. However, due to the limited cardinality of available subjective databases, most deep VQA methods have been trained and evaluated on small databases through cross validation. To train Convolutional Neural Network (CNN) models, each video sequence is typically segmented into a number of patches, all of which are labeled with the sequence-level subjective quality scores. This sub-optimal solution often results in overfitting and inconsistent performance across different databases without re-training.

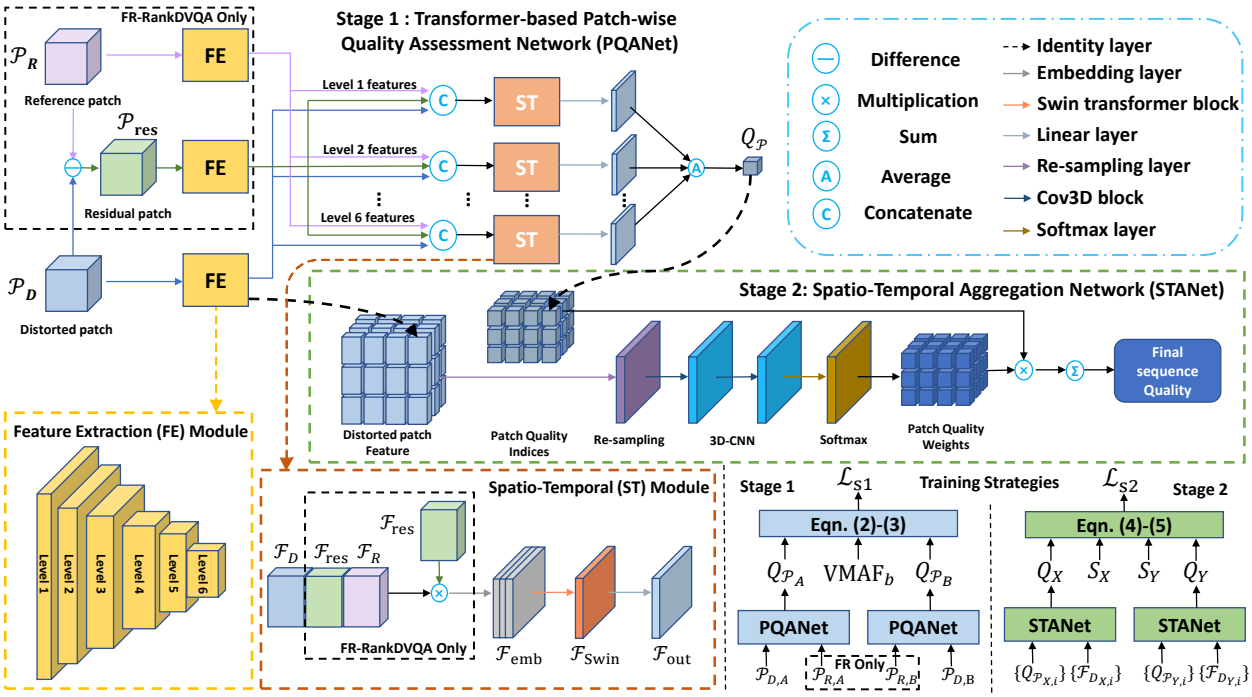

Figure 2. Illustration of the proposed approach including network architectures and training strategies in both stages.

NR VQA methods. NR VQA models estimate the quality of a distorted video without any reference information. Similar to FR methods, conventional NR quality metrics usually extract features from the distorted video and compare them in various (spatial, temporal, spatio-temporal, and frequency and wavelet) domains. Notable examples include BRISQUE [44], NIQE [46], VIIDEO [45], Video BLIINDS [51], BIQI [47] and DIIVINE [48]. Recently, a number of deep learning-based NR VQA methods, including DeepVBQA [2], TLVQM [25], DeepSTQ [83], VSFA [26], RAPIQUE [58], MDTVSFA [27], VIDEVAL [57] and GSTVQA [5] have been proposed, providing enhanced performance over classic methods, although their development is still constrained with the same issue with deep FR methods - lack of sufficient and diverse training material.

Learning to rank. Most deep VQA methods mentioned above were trained by minimizing the $\ell 1/\ell 2$ distance to the subjectively assessed ground truth. However, it is more important for a quality metric to provide accurate ranking results given different distorted sequences. To this end, learning-to-rank based methods have been proposed in the context of no reference image quality assessment [11, 12, 34, 39, 81, 82]. However, these methods do not fully exploit the ranking information associated with distorted versions generated from both single and multiple sources. Furthermore, this approach has not been fully applied for generic VQA - it has only been used within a bespoke video frame interpolation quality metric [15].

3. Proposed Method: RankDVQA

The architecture of RankDVQA is shown in Figure 2. The input distorted video sequence (together with its corresponding original counterpart for the FR case) is first segmented into non-overlapped $H \times W \times C \times N$ video volume (patches). Here H and W correspond to the spatial resolution of the each image, and they are set to 256. $C = 3$ is the number of channels (YCbCr) and N is the number of sampled frames, set to 12 here, following [15]. The input dodecuplet(s) is first processed by a transformer-based network which produces a patch-level quality index (Stage 1). The indices for all the dodecuplets in the same sequence will be passed to an aggregation network that outputs the final sequence level quality index (Stage 2). The network architecture, training material and methodologies used in these two stages are described below.

3.1. Stage 1: Patch Level Quality Assessment

Network architecture. In the first stage, the proposed method employs a transformer-based patch quality assessment network, PQANet, to predict a quality index, given the input distorted patch, $\mathcal{P}_D$, (and its reference counterpart, $\mathcal{P}_R$, in the full reference scenario), with a size of 256×256×3×12. If this network is used for FR quality assessment, a residual patch, $\mathcal{P}_{\text{res}}$ is also calculated [21, 22],

$$\mathcal{P}_{\text{res}} = \frac{\log(1/((\mathcal{P}_{\mathcal{R}} - \mathcal{P}_{\mathcal{D}})^2 + \mathcal{E}/(2^B - 1)^2)}{\log((2^B - 1)^2/\mathcal{E})}, \quad (1)$$

where $\mathcal{E}$ equals 1 [21, 22], and B is the video bit depth.

The network architecture used in the first stage (shown in Figure 2) is similar to that reported in [15], which is based on Swin Transformer [35] blocks. Similar designs have also been used in [18, 35, 68]. This network first employs a Feature Extraction (FE) module (a pyramid network) to extract six levels of feature maps for each input patch. Each level in the pyramid network comprises two 3×3 Conv2D layers, with the second layer having a stride of 2 for downsampling. The number of channels in each level increases in a progressive manner, starting with 16 in the first level and reaching 128 in the final level. For each level, we concatenate the features from all input patches to obtain the input for the spatio-temporal (ST) module.

Based on the concatenated feature maps at each level, we employ the ST module to estimate the spatio-temporal quality at this level. The ST module first normalizes all input features in the channel dimension [79], and concatenates them together. Different from [15], in the full reference scenario, we perform an element-wise multiplication between residual features, $\mathcal{F}_{\text{res}}$, and the concatenated features, $[\mathcal{F}_D \ \mathcal{F}_{\text{res}} \ \mathcal{F}_R]$, (where $\mathcal{F}_D, \mathcal{F}_R$ are extracted from $\mathcal{P}_D$ and $\mathcal{P}_R$), before feeding them into the linear embedding layer that projects features to a fixed dimension of 32. This has been proved to offer evident improvement over absolute position embedding in [35]. The output, $\mathcal{F}_{\text{emb}}$ is then processed through a Swin Transformer block [35] and a linear layer to obtain the quality index at this feature level. Finally the indices for all six feature levels are averaged to calculate the final patch level quality index, $Q_{\mathcal{P}}$.

Training Database Generation. As highlighted in Section 1, there are two primary issues with most current deep VQA methods, (i) a lack of large diverse training databases (ii) ineffective training strategies. To overcome the first issue, instead of using small video databases containing subjective opinion scores, we developed a large-scale training database containing diverse distorted sequences with artifacts commonly encountered in modern video streaming scenarios.

To generate the training content, we selected 230 source sequences from the BVI-DVC dataset [38] and the training database used in the CVPR 2022 CLIC video compression challenge [13]. Each source sequence was segmented into 64 frames and converted to YCbCr 4:2:0 format. These sequences were then compressed using four standard video codecs: H.264/AVC x264 [43], H.265/HEVC HM 16.20 [16], AOM/AV1 1.0.0 [3] and H.266/VVC VTM 7.0 [17], at four quantisation levels to create diverse distortion types. The codec versions and configurations are summarized in *Supplementary Material*. To further augment the training data, we also generated training content through resolution adaptation, which is commonly used in video streaming servers [19]. Specifically, all the source sequences were first down-sampled to three lower spatial resolutions, by a factor of 1.5, 2 or 3, using the Lanczos3 filter [59]. These low resolution sequences were then compressed by the four codecs mentioned above with the same coding configurations (also with four quantisation levels). Each compressed low resolution video was then decoded and up-sampled to its original resolution using the Lanczos3 filter to obtain the distorted sequence. This results in a total number of 14720 (230 sources $\times$ 4 resolutions $\times$ 4 codecs $\times$ 4 quantization levels) distorted sequences. All the distorted videos are converted to YCbCr 4:4:4 format for patch generation.

In order to perform ranking-based training, we produce patch level training material based on two approaches - single-source patch generation and dual-source patch generation. The former is used to improve the ability of the employed network to identify the quality difference between distorted versions of the same source. This is an important property of a VQA method when it is used for comparing the quality of videos generated by different methods, such as for video codec comparison. On the other hand, training on dual-source patches allows the model to produce more reliable quality scores for videos with different content characteristics. We note that these two characteristics have not been previously exploited in the development (or training) of VQA methods.

To produce single-source patches, we first randomly selected two distorted sequences corresponding to the same reference sequence, then randomly segmented each distorted sequence and its reference counterpart using a non-overlapping spatio-temporal sliding window of size $256 \times 256 \times 12$. Each segment generates three dodecuplets (one for each distorted version, and one for the reference). Secondly, to generate dual-source patches, we randomly selected two distorted sequences, and segmented each (and their respective reference videos) using the same sliding window as mentioned above. The resulting patches can come from different source sequences, or from the same source sequence but at different spatio-temporal locations. Doing this for each segment produces four dodecuplets (one for each distorted patch, and one for each reference patch). It is noted that, for training our NR VQA method, only distorted patches are used, while both distorted and reference patches are employed for the training of the full reference quality metric, as shown in Figure 2.

To label all these patches with reliable quality scores which correlate well with subjective ground truth, inspired by the previous contributions in [7, 49, 73], we used the VMAF (Video Multi-method Assessment Fusion) metric to produce training targets for generated patches by comparing each distorted patch with its corresponding reference. It is further noted that, although VMAF provides relatively consistent correlation with perceptual quality on various video quality databases, its average correlation coefficient (SROCC) of 0.85 is far from perfect [75].

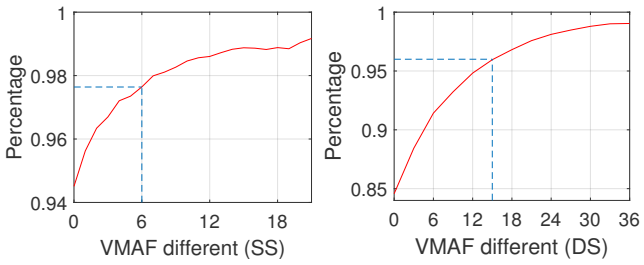

Figure 3. VMAF difference between every two distorted sequences versus the accuracy of VMAF compared to subjective ground truth on the VMAF+ [30] database. (Left) Single sources. (SS) (Right) Dual sources (DS).

To further improve the reliability of the training material, we adopted a similar approach to that in [15] to evaluate the performance of VMAF on an existing database, VMAF+ [30], in terms of the accuracy to differentiate the quality of different distorted sequences of the same (or different) source sequences. The results are summarized in Figure 3. It can be observed that, when the VMAF value difference (between two distorted videos) is larger than 6 (for single source) or 15 (for dual sources), the differentiation accuracy according to the subjective score difference is above 95%[2]. We therefore use these two thresholds to remove potential outliers in the labeled patches. Specifically, for patches in each segment, if the absolute VMAF value difference between two distorted versions is smaller than the threshold (for single source or dual source), we exclude this segment from the final training dataset. In total, this generated approximately 204,800 groups (each of them corresponds to a single segment) of dodecuplet patches associated with the same number of VMAF difference labels, in which there are 30% single sourcing patches and 70% dual sourcing patches.

Ranking-inspired Training Strategy As shown in Figure 2, the employed PQANet is trained in a Siamese method similar to that in [34, 70, 72]. Given two input distorted patches $\mathcal{P}_{D,A}$, $\mathcal{P}_{D,B}$ (and their corresponding reference patches $\mathcal{P}_{R,A}$, $\mathcal{P}_{R,B}$ for FR VQA) from the same patch group, the difference between outputs of the patch quality assessment network, $Q_{\mathcal{P}_A}$ and $Q_{\mathcal{P}_B}$, are compared against their associated VMAF difference (after binarization), VMAF_b ($\mathrm{VMAF}_b = 1$ if the VMAF value of patch A is larger than patch B. Otherwise, $\mathrm{VMAF}_b = 0$). Specifically, we first calculate a probability p with a sigmoid layer,

$$p = \mathrm{sigmoid}(Q_{\mathcal{P}_A} - Q_{\mathcal{P}_B}), \tag{2}$$

and obtain the binary cross entropy loss as:

$$\mathcal{L}_{s1} = -(\mathrm{VMAF}_b \log(p) + (1 - \mathrm{VMAF}_b) \log(1-p)). \tag{3}$$

[2]We have used the same accuracy ratio 97.5% as in [15, 34] for single source cases. A slightly lower ratio (96%) was utilized for dual source cases in order to remove fewer outliers (to obtain more training samples).

This was used as the loss function to train the PQANet.

3.2. Stage 2: Spatio-temporal Pooling

Network Architecture. Most of the existing spatial-temporal pooling approaches in deep VQA only take the obtained patch-level quality indices as inputs [6, 22, 71], but ignore the information within the distorted content. In this work, inspired by [69] in which only temporal information was extracted for pooling, we designed a novel spatial-temporal aggregation network, shown in Figure 2, which accepts both patch-level score tensors and the distorted features maps (extracted in Stage 1) to obtain the final sequence level quality score.

First, the network takes all the patch level quality indices generated in Stage 1, groups them in a 3D tensor and normalizes the tensor size to $W_{\mathcal{P}} \times H_{\mathcal{P}} \times T_{\mathcal{P}}$ (the re-sampling is based on simple local mean). Here we set up $W_{\mathcal{P}} = 16$, $H_{\mathcal{P}} = 9$ and $T_{\mathcal{P}} = 10$. The network also re-uses the feature maps $\mathcal{F}_D$ (Level 6 and Level 3 only [28, 29]) generated by the FE module in Stage 1 based on all the distorted patches, and combines these feature maps in a 3D tensor and re-sample it to the same resolution of the patch quality tensor $16 \times 9 \times 10$ (each component in this tensor corresponds to two feature maps at level 6 and 3).

The distorted feature tensor is then fed into two 3D convolution blocks (each containing two 3D CNN layers with kernel size of $3\times3\times3$, channel number of 8, and stride of 1) followed by a Softmax function to capture spatio-temporal information, which is used to weight the patch level quality indices. The weighted indices are then combined to produce the final quality score for this video sequence.

Training Material. In Stage 2, we use video quality databases containing subjective ground truth labels to train the spatio-temporal aggregation network. As we still adopt a ranking-inspired training strategy (as discussed below), we are able to combine training content from multiple video quality databases for the first time, which significantly improves the training effectiveness. We employed two databases, VMAF+ [30] and IVP [77] in the Stage 2 training. VMAF+ was used to train the original VMAF model, while IVP contains diverse distortion types generated by different compression algorithms.

Each sequence in both databases was first segmented into non-overlapped $256\times256\times3\times12$ patches and fed into the PQANet to obtain patch level quality indices. These quality indices together with their corresponding sequence level subjective opinion score (as the training target) were used as the training material in Stage 2.

Training Strategy. As in Stage 1, a similar ranking-inspired training strategy was also employed to train the spatio-temporal aggregation network (STANet). For each pair of randomly selected distorted sequences (from the same database), denoted as X and Y, given their the input patch quality indices, $\{Q_{\mathcal{P}_{X,i}}\}$ and $\{Q_{\mathcal{P}_{Y,i}}\}$, their patch-

level features $\{\mathcal{F}_{D_{X,i}}\}$ and $\{\mathcal{F}_{D_{Y,i}}\}$, and the corresponding normalized ([0, 100]) subjective opinion scores (e.g., MOS or 100-DMOS), S_X and S_Y, we calculate the following loss to optimize STANet.

$$\delta = \text{STANet}(\{Q_{\mathcal{P}_{X,i}}\}, \{\mathcal{F}_{D_{X,i}}\}) - \text{STANet}(\{Q_{\mathcal{P}_{Y,i}}\}, \{\mathcal{F}_{D_{Y,i}}\}) \tag{4}$$

$$\mathcal{L}_{s2} = \| \delta - (S_X - S_Y) \|_2 . \tag{5}$$

We have used a total number of 16,000 sequences pairs from both databases in the Stage 2 training. It is noted that we take the subjective score difference into account rather than perform the binarization as in Stage 1. This is because we believe that this difference value (from real ground truth) provides more important and accurate information compared to that from VMAF.

4. Experiment Setup

Implementation Details. Pytorch 1.10 was used to implement both networks, with the following training parameters: Adam optimization [23] with β_1=0.9 and β_2=0.999; 60 training epochs; batch size of 4; the initial learning rate is 0.0001 with a weight decay of 0.1 after every 20 epochs. Both training and evaluation were executed on a computer with a 2.4GHz Intel CPU and an NVIDIA P100 GPU.

Evaluation Dataset and Configuration. To evaluate the performance of the proposed methods and their model generalization, eight commonly used video quality datasets[3] were employed including NFLX [30], NFLX-P [30], BVI-HD [78], BVI-CCHD [76], BVI-CCHDDO [20], MCL-V [32], SHVC [14] and VQEGHD3 [61]. In this experiment, we did not re-train our model (or other learning-based methods) on these databases and test their (intra-database) performance through cross-validation, as this often leads to less meaningful results (with correlation coefficients that are close to 1) due to overfitting issues [6, 22, 71]. Instead, we only used the training databases mentioned in Section 3 to optimize our approaches, and fixed the model parameters during evaluation. This experimental design challenges against the generalization of the tested VQA methods.

In the evaluation stage, for FR-RankDVQA, each distorted sequence and its corresponding reference are segmented into non-overlapping 256×256×12 spatio-temporal patches and converted into YCbCr 4:4:4 formats as the input of the PQANet. The output quality indices for all these patches are then employed as the input of the STANet to obtain the final quality index of this sequence. For NR-RankDVQA, only distorted sequence is required and it is processed following the same approach described above.

Evaluation metrics. To assess the performance of each VQA method, four correlation statistics have been calculated: the Spearman Rank Order Correlation Coefficient (SROCC), the Pearson Linear Correlation Coefficient (PLCC), the Outlier ratio (OR) and the Root Mean Squared Error (RMSE) [60] to appraise prediction accuracy (PLCC, RMSE), monotonicity (SROCC) and consistency (OR). Due to the limited space in the main paper, we only present SROCC results here, and include results based on the other three metrics in the *Supplementary Material*. Additionally, a significance test was also performed to differentiate between the proposed methods (FR- and NR-RankDVQA) and their benchmarks in each category (FR and NR) on each test database. The approach in [52, 54] was used in which an F-test was conducted on the residuals between the subjective opinion scores and the predicated quality scores by the tested VQA methods through a non-linear regression employing a logistic function [60].

Benchmark VQA Methods. To benchmark the performance of the proposed methods, we tested ten full reference quality assessment methods, including three conventional ones: PSNR, SSIM [65], and MS-SSIM [67]; and five deep VQA methods:[4] DeepQA [21], DeepVQA [22], C3DVQA [71], DISTS [10] and LPIPS [80]; and two machine learning (SVM regression) based quality metrics: GREED [40] and VMAF [30] We have also evaluated eight NR VQA methods, among which MDTVSFA [27], CONVIQT [41] (unsupervised learning), VIDEVAL [57] GSTVQA [5] are deep learning based methods, while VIIDEO [45], BRISQUE [44], NIQE [46] and TLVQM [24] are classic or machine learning-based approaches. For all the deep learning-based quality metrics, we used their publicly available pre-trained models for result generation[5].

5. Experiments

5.1. Quantitative Evaluation

Full Reference VQA Method Comparison. The results in Table 1 shows that FR-RankDVQA achieves the highest overall SROCC value of 0.8972 among all tested FR quality metrics across eight databases. The second best performer is VMAF, which is consistently inferior to FR-RankDVQA. Based on the F-test results, their performance difference is significant on BVI-HD and MCL-V databases. It is also noted that all the deep VQA methods underperform VMAF, and none of them achieve a SROCC value higher than 0.8.

In terms of complexity, the runtime of FR-RankDVQA is 3.88 times slower than VMAF, similar to other underperforming deep FR-VQA methods, such as DeepVQA [22]

[3] In this paper, our evaluation solely focuses on HD content, as this is the dominant format in modern video streaming.

[4] The selection of benchmark deep VQA methods are based on the reported performance in the original literature and the availability of their pre-trained models.

[5] As the primary contribution of this work is to develop generic deep VQA methods which does not require cross-validation, we did not re-training benchmark methods on different databases. The performance presented here shows the generalization characteristic of each model.

SROCC↑ (F-test)	NFLX	NFLX-P	BVI-HD	BVI-CCHD	BVI-CCHDDO	MCL-V	SHVC	VQEGHD3	Overall	Overall (SS)
Full Reference VQA Methods										
PSNR	0.6218 (-1)	0.6596 (-1)	0.6143 (-1)	0.6166 (-1)	0.7497 (-1)	0.4640 (-1)	0.7380 (-1)	0.7518 (-1)	0.6520	0.9476
SSIM [65]	0.5638 (-1)	0.6054 (-1)	0.5992 (-1)	0.7194 (-1)	0.8026 (-1)	0.4018 (-1)	0.5446 (-1)	0.7361 (-1)	0.6216	0.9451
MS-SSIM [67]	0.7136 (-1)	0.7394 (-1)	0.7652 (-1)	0.7534 (-1)	0.8321 (0)	0.6306 (-1)	0.8007 (0)	0.8457 (0)	0.7601	0.9477
DeepQA [21]	0.7298 (-1)	0.6995 (-1)	0.7106 (-1)	0.6202 (-1)	0.6705 (-1)	0.6832 (-1)	0.7176 (-1)	0.7881 (-1)	0.7024	0.8854
LPIPS [80]	0.6793(-1)	0.7859 (-1)	0.6670 (-1)	0.6838 (-1)	0.7678 (-1)	0.6579 (-1)	0.6360 (-1)	0.8075 (0)	0.7107	0.9041
DeepVQA [22]	0.7352 (-1)	0.7609 (-1)	0.7330 (-1)	0.6924 (-1)	0.8120 (0)	0.6142 (-1)	0.8041 (0)	0.7805 (-1)	0.7540	0.9060
C3DVQA [71]	0.7730 (-1)	0.7714 (-1)	0.7393 (-1)	0.7203 (-1)	0.8137 (0)	0.7126 (-1)	0.8194 (0)	0.7329 (-1)	0.7641	0.9421
DISTS [10]	0.7787 (-1)	0.9325 (0)	0.7030 (-1)	0.6303 (-1)	0.7442 (-1)	0.7792 (-1)	0.7813 (0)	0.8254 (0)	0.7718	0.9235
ST-GREED [40]	0.7470 (-1)	0.7445 (-1)	0.7769 (-1)	0.7738 (-1)	0.8259 (0)	0.7226 (-1)	0.7946 (0)	0.8079 (0)	0.7842	0.9460
VMAF 0.6.1 [30]	0.9254 (0)	0.9104 (0)	0.7962 (-1)	0.8723 (0)	0.8783 (0)	0.7766 (-1)	0.9114 (0)	0.8442 (0)	0.8644	0.9455
FR-RankDVQA	0.9393	0.9184	0.8659	0.8991	0.9037	0.8391	0.9142	0.8979	**0.8972**	**0.9814**
No reference VQA methods										
VIIDEO [45]	0.4550 (-1)	0.5527 (-1)	0.1297 (-1)	0.1308 (-1)	0.2523 (-1)	0.0406 (-1)	0.2033 (-1)	0.1881 (-1)	0.2440	0.3087
TLVQM [24]	0.4652 (-1)	0.4720 (-1)	0.3124 (-1)	0.1622 (-1)	0.3420 (-1)	0.2758 (-1)	0.4983 (0)	0.5382 (0)	0.3469	0.8239
BRISQUE [44]	0.7828 (0)	0.7861 (0)	0.2033 (-1)	0.3738 (-1)	0.3746 (-1)	0.3154 (-1)	0.3601 (-1)	0.5467 (0)	0.4716	0.5894
NIQE [46]	0.7959 (0)	0.8269 (0)	0.1932 (-1)	0.4247 (-1)	0.5225 (-1)	0.3985 (-1)	0.6210 (0)	0.5291 (0)	0.5390	0.7029
MDTVSFA [27]	0.5137 (-1)	0.6024 (-1)	0.3725 (-1)	0.4068 (-1)	0.5547 (-1)	0.5712(0)	0.6165 (0)	0.6422 (0)	0.5311	0.8872
CONVIQT [41]	0.6989 (-1)	0.7962 (0)	0.3489 (-1)	0.3706 (-1)	0.5381(-1)	0.6323 (0)	0.4983 (0)	0.6217 (0)	0.5631	0.6846
VIDEVAL [57]	0.7899 (0)	0.7261 (0)	0.5884 (-1)	0.6974 (0)	0.7620 (0)	0.4836 (-1)	0.6428 (0)	0.5326 (0)	0.6529	0.8621
GSTVQA [5]	0.8109 (0)	0.7858 (0)	0.4132 (-1)	0.7447 (0)	0.7665 (0)	0.7385 (0)	0.6710 (0)	0.7011 (0)	0.7040	0.9014
NR-RankDVQA	0.8346	0.7944	0.7326	0.7628	0.7994	0.7631	0.7118	0.8346	**0.7791**	**0.9266**
Ablation Study Results										
V1 ($\ell 1$)	0.8793(0)	0.8816 (0)	0.7583 (-1)	0.7792 (-1)	0.8523 (0)	0.7678 (-1)	0.8238 (0)	0.8501 (0)	0.8190	0.9417
V2 ($\ell 2$)	0.8812 (0)	0.8883 (0)	0.7612 (-1)	0.7794 (-1)	0.8568 (0)	0.7696 (-1)	0.8234 (0)	0.8507 (0)	0.8263	0.9467
V3 (C3D)	0.9034(0)	0.8964 (0)	0.8233 (0)	0.8763 (0)	0.8961 (0)	0.8054 (0)	0.8692 (0)	0.8465 (0)	0.8653	0.9693
V4 (S1)	0.9201(0)	0.8983 (0)	0.8361 (0)	0.8825 (0)	0.8987 (0)	0.8231 (0)	0.8966 (0)	0.8544 (0)	0.8762	0.9617
FR-RankDVQA	0.9393	0.9184	0.8659	0.8991	0.9037	0.8391	0.9142	0.8979	**0.8972**	**0.9814**

Table 1. Performance of the proposed methods, other benchmark approaches and ablation study variants on eight HD test databases. The values in each cell x(y) correspond to the SROCC value (x) and F-test result (y) at 95% confidence interval. y=1 suggests that the metric is superior to FR-RankDVQA in the full reference track or NF-RankDVQA in the no reference track (y=-1 if the opposite is true), while y=0 indicates that there is no significant difference between them. The figures in red and blue indicate the highest and second highest SROCC values respectively in each column.

(4.05×) and C3DVQA [71] (3.37×). Comprehensive complexity figures can be found in the *Supplementary Material*.

No Reference VQA Method Comparison. In the no reference case, NR-RankDVQA also offers the best overall performance compared to other tested NR VQA methods, with an average SROCC of 0.7791. This figure is much higher than that of the second best quality metric, GSTVQA [5] (SROCC = 0.7040), and than some of the full reference VQA methods including PSNR, SSIM, MS-SSIM, WaDIQA, DeepQA, DeepVQA and C3DVQA.

Assessing Content from Single Sources. As pointed out in Section 3.1, it is important for a VQA method to accurately differentiate the quality between distorted versions from the same source or from different sources. The latter has been effectively evaluated above, when the SROCC values were calculated for the whole database. In order to test the single source case, following the evaluation procedure in [15], we first compute the SROCC value for all the distorted sequences from each individual source, and average them among all sources within each database. The overall SROCC values (for all test databases) for selected VQA methods are summarized in Table 1 (Overall (SS) column), while comprehensive results can be found in *Supplementary Material*. It can be observed that the proposed FR- and NR-RankDVQA also outperform the benchmark algorithms in each track (FR and NR). Both properties of the proposed method (full reference) have further been confirmed in Figure 4 when compared with VMAF.

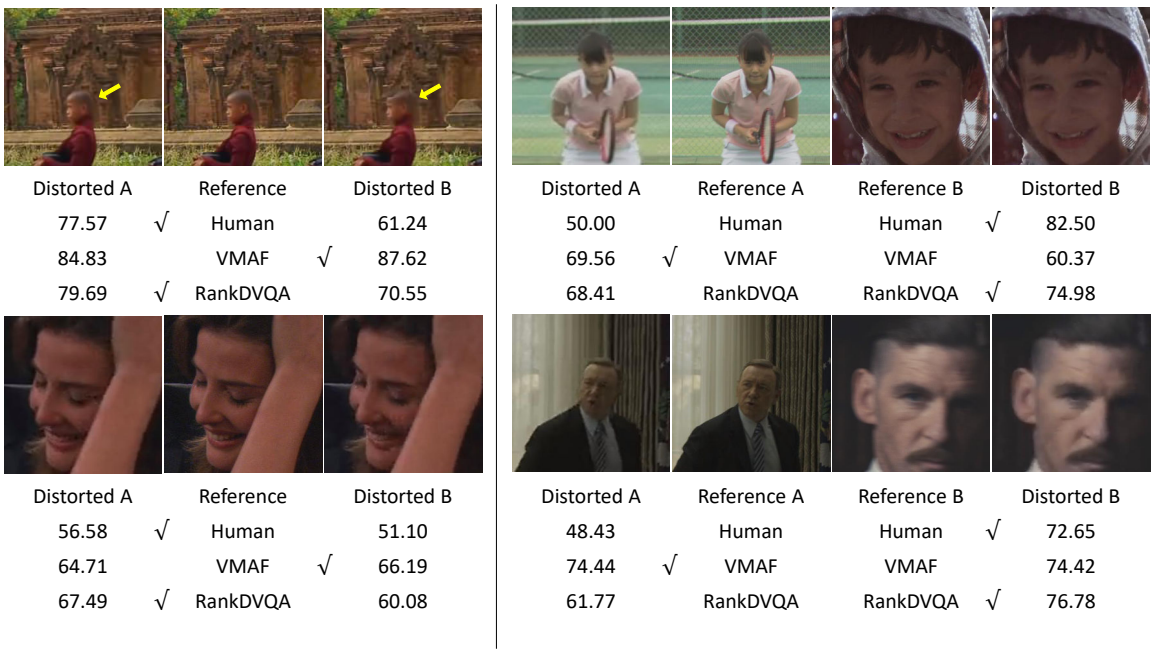

Figure 4. Visual examples demonstrating the superiority of the proposed FR quality metric. For the frames on the left, it can be seen that the distorted video A is given a higher subjective quality score than video B, and FR-RankDVQA correctly predicts this. Similarly for frames on the right, although distorted video B shows less visual distortion from its reference compared to distortion between distorted video A and reference A, VMAF fails to predict the correct quality rank (the proposed method does).

5.2. Ablation Study

To evaluate the primary contributions of this work, we have evaluated the following RankDVQA variants in an ablation study. Here we only focus on the FR scenario.

Effectiveness of the ranking-inspired Losses. One major novelty of this work is the Ranking-inspired training strategy. To evaluate its contribution and compare it with commonly used loss functions, $\ell 1$ and $\ell 2$ (between the output of the network and the training target), in the VQA research community, we have implemented two variants V1($\ell 1$) and V2($\ell 2$) which replaced the Ranking-inspired loss functions with $\ell 1$ or $\ell 2$ in both stages. The results for both variants are also shown in Table 1, which are also outperformed by FR-RanDVQA on all tested databases.

Effectiveness of PQANet. In order to validate the contribution of PQANet to the overall design, we evaluate the training framework on a different deep VQA model, C3DVQA, creating V3(C3D). Except for the exchange of the first-stage network, all other training settings (including the second stage) are unchanged. It can be observed in Table 1 that V3(C3D) is also outperformed by FR-RankDVQA, which implies the effectiveness of the PQANet.

Effectiveness of STANet. To verify the effectiveness of the STANet, we have replaced this aggregation network with a simple arithmetic average operation, which is typically used for spatio-temporal pooling. Its performance is denoted as V4(S1) in Table 1. It can be observed by comparing this variant with FR-RankDVQA, that the latter offers higher SROCC values on all tested video databases.

6. Conclusion

In this paper, we propose new deep VQA methods based on a ranking-inspired hybrid training methodology. We introduce for the first time, the use of a large scale training database (created without the need to perform expensive and time consuming subjective tests) to optimize deep networks with high capacity for quality assessment. This also supports the combination of multiple existing video quality databases (with subjective ground truth labels) to train an aggregation network for spatio-temporal pooling.

The proposed methods, FR-RankDVQA and NR-RankDVQA, were fully tested on eight evaluation databases, and were shown to exhibit higher correlation with opinion scores when compared to other full reference and no reference VQA methods. To the best of our knowledge, FR-RankDVQA is the first deep VQA method that consistently outperforms VMAF on multiple databases without conducting intra-database cross validation. Future work should focus on more sophisticated pooling network architectures and complexity reduction.

Acknowledgement. The authors appreciate the funding from the UKRI MyWorld Strength in Places Programme (SIPF00006/1), the University of Bristol, and the China Scholarship Council.

References

[1] Sewoong Ahn, Yeji Choi, and Kwangjin Yoon. Deep learning-based distortion sensitivity prediction for full-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 344–353, June 2021. 2

[2] Sewoong Ahn and Sanghoon Lee. Deep blind video quality assessment based on temporal human perception. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 619–623. IEEE, 2018. 3

[3] AOM. AOMedia Video 1 (AV1), 2019. 4

[4] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing*, 27(1):206–219, 2018. 2

[5] Baoliang Chen, Lingyu Zhu, Guo Li, Fangbo Lu, Hongfei Fan, and Shiqi Wang. Learning generalized spatial-temporal deep feature representation for no-reference video quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(4):1903–1916, 2022. 3, 6, 7

[6] Junming Chen, Haiqiang Wang, Munan Xu, Ge Li, and Shan Liu. Deep neural networks for end-to-end spatiotemporal video quality prediction and aggregation. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2021. 2, 5, 6

[7] Li-Heng Chen, Christos G. Bampis, Zhi Li, Andrey Norkin, and Alan C. Bovik. Proxiqa: A proxy approach to perceptual optimization of learned image compression. *IEEE Transactions on Image Processing*, 30:360–373, 2021. 4

[8] CISCO. Cisco annual internet report (2018–2023) white paper. Technical report, March 2020. 1

[9] Duolikun Danier, Fan Zhang, and David Bull. FloLPIPS: A bespoke video quality metric for frame interpolation. In *2022 Picture Coding Symposium (PCS)*, pages 283–287, 2022. 1

[10] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2020. 2, 6, 7

[11] Fei Gao, Dacheng Tao, Xinbo Gao, and Xuelong Li. Learning to rank for blind image quality assessment. *IEEE Transactions on Neural Networks and Learning Systems*, 26(10):2275–2290, 2015. 3

[12] S. Alireza Golestaneh, Saba Dadsetan, and Kris M. Kitani. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1220–1230, January 2022. 3

[13] Advanced Encoding Guide. the fifth workshop and challenge on learned image compression (video track). *http://compression.cc/*, 2022. 4

[14] Y. He, Y. Ye, F. Hendry, Y.-K. Wang, and V. Baroncini. SHVC verification test results. In *the JCT-VC meeting, JCTVC-W0095*. ITU-T, ISO/IEC, 2016. 6

[15] Qiqi Hou, Abhijay Ghildyal, and Feng Liu. A perceptual quality metric for video frame interpolation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 234–253. Springer, 2022. 1, 3, 4, 5, 7

[16] ITU-T Rec. H.265. High efficiency video coding. ITU-T Std. 2020. 4

[17] ITU-T Rec. H.266. Versatile video coding. ITU-T Std. 2020. 4

[18] Shaojie Jiang, Qingbing Sang, Zongyao Hu, and Lixiong Liu. Self-supervised representation learning for video quality assessment. *IEEE Transactions on Broadcasting*, 2022. 4

[19] I. Katsavounidis. Dynamic optimizer – a perceptual video encoding optimization framework. *The Netflix Tech Blog*, 2018. 4

[20] A. V. Katsenou, F. Zhang, M. Afonso, and D. R. Bull. A subjective comparison of AV1 and HEVC for adaptive video streaming. In *Proc. IEEE Int Conf. on Image Processing*, 2019. 6

[21] Jongyoo Kim and Sanghoon Lee. Deep learning of human visual sensitivity in image quality assessment framework. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1969–1977, 2017. 2, 3, 4, 6, 7

[22] Woojae Kim, Jongyoo Kim, Sewoong Ahn, Jinwoo Kim, and Sanghoon Lee. Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 219–234, 2018. 1, 2, 3, 4, 5, 6, 7

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[24] Jari Korhonen. Two-level approach for no-reference consumer video quality assessment. *IEEE Transactions on Image Processing*, 28(12):5923–5938, 2019. 6, 7

[25] Jari Korhonen, Yicheng Su, and Junyong You. Blind natural video quality prediction via statistical temporal features and deep spatial features. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3311–3319, 2020. 3

[26] 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. 3

[27] 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. 3, 6, 7

[28] Guoqing Li, Meng Zhang, Jiaojie Li, Feng Lv, and Guodong Tong. Efficient densely connected convolutional neural networks. *Pattern Recognition*, 109:107610, 2021. 5

[29] Wei Li, Kai Liu, Lin Yan, Fei Cheng, YunQiu Lv, and LiZhe Zhang. Frd-cnn: Object detection based on small-scale convolutional neural networks and feature reuse. *Scientific reports*, 9(1):1–12, 2019. 5

[30] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara. Toward a practical perceptual video quality metric. *The Netflix Tech Blog*, 2016. 1, 2, 5, 6, 7

[31] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *arXiv preprint arXiv:2201.12288*, 2022. 1

[32] 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. 6

[33] Ce Liu and William T Freeman. A high-quality video denoising algorithm based on reliable motion estimation. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part III 11*, pages 706–719. Springer, 2010. 1

[34] Xialei Liu, Joost van de Weijer, and Andrew D. Bagdanov. Rankiqa: Learning from rankings for no-reference image quality assessment. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 3, 5

[35] 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. 4

[36] Alice Lucas, Santiago Lopez-Tapia, Rafael Molina, and Aggelos K Katsaggelos. Generative adversarial networks and perceptual losses for video super-resolution. *IEEE Transactions on Image Processing*, 28(7):3312–3327, 2019. 1

[37] Di Ma, Mariana Afonso, Fan Zhang, and David R Bull. Perceptually-inspired super-resolution of compressed videos. In *Applications of Digital Image Processing XLII*, volume 11137, pages 310–318. SPIE, 2019. 1

[38] Di Ma, Fan Zhang, and David Bull. BVI-DVC: a training database for deep video compression. *IEEE Transactions on Multimedia*, 2021. 4

[39] Kede Ma, Wentao Liu, Tongliang Liu, Zhou Wang, and Dacheng Tao. dipiq: Blind image quality assessment by learning-to-rank discriminable image pairs. *IEEE Transactions on Image Processing*, 26(8):3951–3964, 2017. 3

[40] Pavan C. Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C. Bovik. St-greed: Space-time generalized entropic differences for frame rate dependent video quality prediction. *IEEE Transactions on Image Processing*, 30:7446–7457, 2021. 2, 6, 7

[41] Pavan C. Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C. Bovik. Conviqt: Contrastive video quality estimator, 2022. 6, 7

[42] Pavan C Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C Bovik. Image quality assessment using contrastive learning. *IEEE Transactions on Image Processing*, 31:4149–4161, 2022. 2

[43] Loren Merritt and Rahul Vanam. x264: A high performance h. 264/avc encoder. 2006. 4

[44] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. 3, 6, 7

[45] Anish Mittal, Michele A Saad, and Alan C Bovik. A completely blind video integrity oracle. *IEEE Transactions on Image Processing*, 25(1):289–300, 2015. 1, 3, 6, 7

[46] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2012. 3, 6, 7

[47] Anush Krishna Moorthy and Alan Conrad Bovik. A two-step framework for constructing blind image quality indices. *IEEE Signal Processing Letters*, 17(5):513–516, 2010. 3

[48] Anush Krishna Moorthy and Alan Conrad Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE transactions on Image Processing*, 20(12):3350–3364, 2011. 3

[49] Darren Ramsook, Anil Kokaram, Noel O'Connor, Neil Birkbeck, Yeping Su, and Balu Adsumilli. A differentiable vmaf proxy as a loss function for video noise reduction. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 11842, page 118420X, 2021. 4

[50] Abdul Rehman, Kai Zeng, and Zhou Wang. Display device-adapted video quality-of-experience assessment. In *Human Vision and Electronic Imaging XX*, volume 9394, page 939406. International Society for Optics and Photonics, 2015. 1, 2

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

[52] K. Seshadrinathan and A. C. Bovik. Motion tuned spatio-temporal quailty assessment of natural videos. *IEEE Transactions on Image Processing*, 19:335–350, 2010. 6

[53] Kalpana Seshadrinathan and Alan Conrad Bovik. Motion tuned spatio-temporal quality assessment of natural videos. *IEEE Transactions on Image Processing*, 19(2):335–350, 2010. 2

[54] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack. Study of subjective and objective quality assessment of video. *IEEE Transactions on Image Processing*, 19:335–350, 2010. 6

[55] H. R. Sheikh and A. C. Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15:430–444, 2006. 1

[56] Matias Tassano, Julie Delon, and Thomas Veit. Fastdvdnet: Towards real-time deep video denoising without flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1354–1363, 2020. 1

[57] 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. 3, 6, 7

[58] 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. 3

[59] Ken Turkowski. *Filters for Common Resampling Tasks*, page 147–165. Academic Press Professional, Inc., USA, 1990. 4

[60] Video Quality Experts Group. Final report from the video quality experts group on the validation of objective quailty metrics for video quality assessment. Technical report, 2000. 6

[61] Video Quality Experts Group. Report on the validation of video quality models for high definition video content. Technical report, VQEG, 2010. 6

[62] Phong V. Vu, Cuong T. Vu, and Damon M. Chandler. A spatiotemporal most-apparent-distortion model for video quality assessment. In *2011 18th IEEE International Conference on Image Processing*, pages 2505–2508, 2011. 2

[63] Phong V. Vu, Cuong T. Vu, and Damon M. Chandler. A spatiotemporal most-apparent-distortion model for video quality assessment. In *2011 18th IEEE International Conference on Image Processing*, pages 2505–2508, 2011. 2

[64] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1

[65] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 1, 2, 6, 7

[66] Zhou Wang, Ligang Lu, and Alan C Bovik. Video quality assessment based on structural distortion measurement. *Signal processing: Image communication*, 19(2):121–132, 2004. 1, 2

[67] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multi-scale structural similarity for image quality assessment. In *Proc. Asilomar Conference on Signals, Systems and Computers*, volume 2, page 1398. IEEE, 2003. 1, 2, 6, 7

[68] 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 European Conference on Computer Vision (ECCV)*, pages 538–554. Springer, 2022. 4

[69] Haoning Wu, Chaofeng Chen, Liang Liao, Jingwen Hou, Wenxiu Sun, Qiong Yan, and Weisi Lin. Discovqa: temporal distortion-content transformers for video quality assessment. *arXiv preprint arXiv:2206.09853*, 2022. 2, 5

[70] Lin Wu, Yang Wang, Junbin Gao, and Xue Li. Where-and-when to look: Deep siamese attention networks for video-based person re-identification. *IEEE Transactions on Multimedia*, 21(6):1412–1424, 2018. 5

[71] Munan Xu, Junming Chen, Haiqiang Wang, Shan Liu, Ge Li, and Zhiqiang Bai. C3DVQA: Full-reference video quality assessment with 3d convolutional neural network. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4447–4451, 2020. 1, 2, 5, 6, 7

[72] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Deep metric learning for person re-identification. In *2014 22nd International Conference on Pattern Recognition*, pages 34–39. IEEE, 2014. 5

[73] Saman Zadtootaghaj, Nabajeet Barman, Rakesh Rao Ramachandra Rao, Steve Göring, Maria G. Martini, Alexander Raake, and Sebastian Möller. Demi: Deep video quality estimation model using perceptual video quality dimensions. In *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6, 2020. 4

[74] Fan Zhang and David R. Bull. A perception-based hybrid model for video quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(6):1017–1028, 2016. 2

[75] Fan Zhang, Angeliki Katsenou, Christos Bampis, Lukas Krasula, Zhi Li, and David Bull. Enhancing vmaf through new feature integration and model combination. *2021 Picture Coding Symposium (PCS)*, Jun 2021. 4

[76] Fan Zhang, Angeliki V Katsenou, Mariana Afonso, Goce Dimitrov, and David R Bull. Comparing VVC, HEVC and AV1 using objective and subjective assessments. *arXiv preprint arXiv:2003.10282*, 2020. 6

[77] F. Zhang, S. Li, L. Ma, Y. C. Wong, and K. N. Ngan. IVP subjective quality video database, 2009. 5

[78] F. Zhang, F. Mercer Moss, R. Baddeley, and D. R. Bull. BVI-HD: A video quality database for HEVC compressed and texture synthesised content. *IEEE Transactions on Multimedia*, 20(10):2620–2630, October 2018. 6

[79] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 4

[80] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 6, 7

[81] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1):36–47, 2020. 3

[82] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Transactions on Image Processing*, 30:3474–3486, 2021. 3

[83] Wei Zhou and Zhibo Chen. Deep local and global spatiotemporal feature aggregation for blind video quality assessment. In *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 338–341. IEEE, 2020. 3