Semi-Supervised Learning of Perceptual Video Quality by Generating Consistent Pairwise Pseudo-Ranks

Shankhanil Mitra, Graduate Student Member, IEEE, Saiyam Jogani, and Rajiv Soundararajan, Senior Member, IEEE

Abstract—Designing learning-based no-reference (NR) video quality assessment (VQA) algorithms for camera-captured videos is cumbersome due to the large number of human annotations of quality. In this work, we propose a semi-supervised learning (SSL) framework exploiting many unlabelled and very limited numbers of authentically distorted labelled videos. Our main contributions are twofold. Leveraging the benefits of consistency regularization and pseudo-labelling, our SSL model generates pairwise pseudo-ranks for the unlabelled videos using a student-teacher model on strong-weak augmented videos. We design the strong-weak augmentations to be quality invariant to use the unlabelled videos effectively in SSL. The generated pseudo-ranks are used along with the limited labels to train our SSL model. Our primary focus in SSL for NR VQA is to learn mapping from video feature representations to quality scores. We compare various feature extraction methods and show that our SSL framework can lead to improved performance on these features. We present a spatial and temporal feature extraction method based on predicting spatial and temporal entropic differences. We show that these features help achieve robust performance when trained with limited data, providing a better baseline to apply SSL. Extensive experiments on three popular VQA datasets demonstrate that the proposed semi-supervised VQA method improves on the performance of existing methods in terms of correlation with human opinion by approximately 15—20%.

Index Terms—No reference video quality assessment, quality feature learning, semi-supervised learning, pairwise ranks.

I. INTRODUCTION

T

HE ubiquitous availability of mobile cameras has led to a proliferation in the generation of video content. Numerous videos are captured by individuals worldwide every day and uploaded and shared through video service providers. With the rapid growth in such content, it has become essential to monitor and control the quality of videos for efficient storage, transmission, and retrieval. This requirement motivates the study of perceptual video quality assessment (VQA) algorithms to generate video quality scores according to human judgements.

VQA algorithms can be broadly classified into full reference (FR), reduced reference (RR) and no reference (NR) methods. FR methods require a reference video for comparison to evaluate the quality of a distorted video [1], [2], [3]. RR methods require a small amount of information from the reference for quality assessment of the distorted video [4]. NR methods operate only on the distorted video and do not require a reference video for comparison [5]. In this work, we focus primarily on VQA of camera-captured videos that are authentically distorted during the capture process and where a reference video is usually not available. In such scenarios, the NR VQA setting is thus most relevant.

NR VQA algorithms are popularly designed in a machine learning framework. The algorithms are developed by extracting video features and regressing them against human opinion scores obtained through subjective studies. Indeed, deep networks are increasingly being studied [6], [7], [8] to support the design of NR VQA algorithms. Nevertheless, such approaches require a large number of human annotated videos with quality scores. As the conduct of large-scale human studies to collect such scores can be quite cumbersome, such an approach does not scale when newer studies need to be conducted as increasingly diversely distorted videos are generated. The human annotation of video quality is also much more time-consuming compared with image quality or other annotation tasks such as image classification. This fact motivates the study of NR VQA algorithms with few labelled videos. Thus, we focus on the problem of designing learning-based NR VQA algorithms with limited labels. In addition to the limited labels, we assume access to unlabelled videos to design our VQA models. Thus, one could view our problem as a semi-supervised NR VQA problem.

To the best of our knowledge, the problem of semi-supervised NR video quality assessment has rarely been studied in the literature. Although there is some work on semi-supervised image quality assessment [9], [10], [11], the extension to video is nontrivial due to need for modeling temporal distortions. Furthermore, no studies have investigated the training of video quality models with limited labelled data. While several
strategies for semi-supervised learning have been explored in image and video classification, many of those strategies are not directly applicable to NR VQA. For example, consistency regularization methods [12] are usually based on quality degradations of the image/video and are thus not appropriate for VQA. Many pseudo-labelling strategies have been designed for classification [13]. Additionally, their direct application to VQA is not obvious.

We present a novel and reliable semi-supervised learning (SSL) strategy for NR VQA. We perform semi-supervised learning by enforcing consistency regularization on the unlabelled examples. A popular approach to achieve consistency regularization involves student-teacher networks where the student and teacher models are leveraged to make consistent predictions on augmented unlabelled data. However, the augmentations studied in the literature tend to distort video quality, thereby rendering such approaches irrelevant for VQA. One of our main contributions is the design of quality invariant strong-weak augmentations that enable us to apply consistency regularization to semi-supervised NR VQA. The teacher model predictions are usually considered pseudo-labels for the unlabelled data with which the student model can be updated. In earlier works on consistency regularization [14], the unreliability of the pseudo-labels limits the learning of the student network, which is also termed confirmation bias. To account for the unreliability of pseudo-labels, we hypothesize that when the quality predictions of a pair of unlabelled videos differ beyond a threshold, their pairwise ranking is likely correct. Thus, we train the student and teacher networks to be consistent in their pairwise quality rank predictions of video pairs. We refer to our entire learning framework consisting of student-teacher models, strong-weak augmentations and pseudo-rank generation as Learning with consistent Pseudo-Ranks (LPR).

Although we can apply our SSL strategy to different feature extraction methods to yield improvements, learning good video quality features upfront before performing SSL on the target dataset can yield more reliable pseudo-labels. In recent years, successful CNN-based methods [15], [16] on synthetically distorted videos were designed by learning on weak quality labels such as GMSD [17], MS-SSIM [3], and ST-RRED [4]. Thus, we design a video quality feature learning method on videos suffering from synthetic distortions such as compression and transmission losses. In particular, we learn deep spatial and temporal quality features from a subset of frames to effectively model distortions in consumer generated content. The VIDEVAL model [26] adopts an approach of feature selection from features of existing different image and video quality assessment models.

**CNN-based NR VQA:** One broad set of approaches employing CNNs involves the use of CNNs in conjunction with other heuristics. The 3D shearlet transform output was processed using CNNs to predict video quality in one of the first attempts to leverage this approach [27]. A combination of spatial features from CNNs with handcrafted features for temporal cues was utilized for NR VQA [28]. CNN features have also been combined with heuristic feature-based methods to achieve state-of-the-art NR VQA performance [29], [30].
On the other hand, a few methods design a fully CNN-based approach for NR VQA. An end-to-end deep learning framework was developed to predict compressed video quality for specific codecs [6]. Motion representations have also been derived in an end-to-end manner for NR VQA [31]. The use of 3D CNNs was explored along with long short-term memory units [7]. Pretrained ResNet50 [32] features trained for image classification are passed through gated recurrent units for successful NR VQA [8]. PVQ [33] extracts both 2D and 3D features from pretrained Paq-2-Piq [34] and 3D ResNet-18, respectively, to predict global video quality. MLSP-VQA-FF [35] extracts features at multiple levels from the pretrained network and regresses against ground truth quality. Shen et al. [36] designed NR VQA by adopting a hierarchical fusion of features at each scale and mapping these features to ground truth quality. FAST-VQA [37] employs efficient sampling techniques and finetunes a pretrained Swin-T transformer [38] on video fragments. TSCVT-BVQA [39] learns a VQA model by transferring spatial knowledge from pretrained image quality assessment (IQA) and temporal knowledge from a pretrained action recognition model.

Weakly supervised NR VQA: Zhang et al. [40] consider weakly supervised NR VQA, where they learn features by predicting a full-reference image quality measure on the frames. The learned features are then regressed against all available human opinion scores through a quality score histogram feature. Furthermore, a resampling strategy is adopted to select appropriate samples. While our approach also uses a full-reference objective model to learn features, we learn both spatial and temporal features using the perceptually relevant ST-RRED model. Zhang et al. does not consider how such features can be applied to learn quality on authentically distorted videos with limited labels. UCDA [41] explores unsupervised domain adaptation from synthetic video distortions to authentic distortions. However, this method requires a large number of human labels in the source domain to ensure effective performance in the target domain.

Self-supervised feature learning for NR VQA: The Video CORNIA model [42] adopts a dictionary learning approach to learn frame-level quality features. However, because it relies on training with full-reference video quality measures, it cannot be used for authentically distorted videos. CSPT [43] is a pretrained-based self-supervised learning method that learns quality-aware features through the video frame prediction task. The resulting features are used to predict quality using full supervision. VISION [44] learns spatiotemporal quality-aware features using multiview contrastive learning from unlabelled videos.

Unsupervised NR VQA: The VIIDEO [45] model represents a completely blind NR VQA model that does not involve training of any kind [45]. The model identifies intrinsic statistical regularities in natural videos and measures deviations in such properties when distortions are introduced. STEM [46], VIQE [47], and TPQI [48] adopt a similar approach by combining NIQE [49] with the perceptual strengthening hypothesis on temporal information. Nevertheless, it may be possible to perform better than unsupervised models by utilizing labelled and unlabelled distorted videos through semi-supervised learning.

B. Semi-Supervised Learning

The problem of semi-supervised VQA has not been studied much in the literature to the best of our knowledge. Thus, there are no standard benchmarks for comparison. SSL methods relevant to VQA can mostly be divided into three categories [50]: pseudo-labelling, consistency regularization and hybrid models. Pseudo-label [13], [51]-based methods in SSL for classification use the label with the maximum confidence as the pseudo-label for unlabelled data. These types of methods are unsuitable for regression tasks such as VQA. Hybrid methods such as FixMatch [53] use a photometric transformation-based strong-weak augmentation strategy on the student-teacher-based model. None of the existing SSL methods can be directly applied to the VQA task. Moreover, these consistency regularization-based methods suffer from a confirmation bias in the target pseudo-labels due to noisy predictions.

In summary, our proposed work addresses the problem of learning with limited labelled videos by designing a VQA-oriented SSL method. Furthermore, we design our work using a robust video quality feature representation instead of learning an end-to-end model, as deep models in general are not robust in the very limited labelled data setting.

III. METHODOLOGY

A. Problem Formulation

Given a set $\mathcal{V}$ of labelled (or annotated with human opinion subjective scores) videos and a set $\mathcal{U}$ of unlabelled videos, the goal is to learn an NR VQA model that can predict video quality without a reference. The model is evaluated on a test set $\mathcal{W}$. The sets $\mathcal{V}$, $\mathcal{U}$ and $\mathcal{W}$ are similar in the nature of distortions that are evident. This work assumes that all these are obtained by splitting a given VQA database containing authentically distorted videos into non-overlapping sets.

Since the set $\mathcal{V}$ of labelled videos is very small, it may be challenging to learn a model from scratch on such limited data. Thus, we focus on semi-supervised learning to learn a relevant mapping from quality-aware feature representations to the video quality score. Robust features can provide a better baseline when regressed against ground truth than learning a model from scratch with very limited labelled data.

B. Overview

We first describe the configuration of various quality-aware feature models that we use in our SSL framework. We then provide details of our STED-TLQVM features in the context of such feature models. Finally, we introduce our SSL framework. As mentioned in Section II-B, existing SSL methods in the literature are unsuitable for VQA tasks due to quality variant augmentations and confirmation bias. In this work, we adopt a novel pseudo-rank-based consistency regularization approach.
Fig. 1. Overall structure of STED feature learning on synthetically distorted videos. Spatial quality features are learned by training a network to predict SRRED indices from video frames. Similarly, temporal features are learned from frame differences by regressing against TRRED indices. Additionally, there is the layer at which spatial and temporal features are extracted.

C. Quality Learning Model

We now discuss typical architectures of a quality-aware model to estimate video quality. We apply our semi-supervised learning framework to such architectures. Video quality features can be at the frame-level or video-level or a hybrid of both types of features. We focus mainly on how pretrained features can be mapped to perceptual quality using semi-supervised learning. We describe these frameworks as follows.

1) Frame-Level Feature Model: Let the spatiotemporal features extracted at the frame level of a video \( v \) be \( x_n \), where \( n \) is the frame index, \( n \in \{1, 2, \ldots, N\} \), and \( N \) is the total number of frames in a given video. As shown in Fig. 2, the spatiotemporal features are passed through two fully connected layers with ReLU and sigmoid non-linearity, represented as function \( f(\cdot) \), to obtain an output \( f(x_n) \).

\[
z = \frac{1}{N} \sum_{n=1}^{N} f(x_n).
\]  

(1)

The temporal averaging of features to obtain a global video feature is more beneficial than passing the features through recurrent layers, as observed in the literature [29]. \( z \) can be further passed through fully connected layers to output a scalar that represents video quality. Hence, we learn a mapping from \( z \) to perceptual quality using fully connected layers \( g(\cdot) \) as follows:

\[
\hat{q}(v) = g(z).
\]  

(2)

The goal of semi-supervised learning of frame-level features is to learn the mappings \( f(\cdot) \) and \( g(\cdot) \).

Examples of such frame-level features include VSFA [8] and HEKE [15]. In this work, we build on the robust performance of the spatiotemporal entropic differences (ST-RRED) [54] index for compression and transmission distortions. The robust performance is attributed to the use of highly regular statistical models of frames and frame differences. ST-RRED is an NSS-based approach that computes localized entropic differences between the
reference and distorted video frames and frame differences. The spatiotemporal entropic differences were recently predicted in a no-reference manner, and their utility in achieving robust generalization performance in measuring compression and transmission distortions was shown in [16]. In this work, we use a dual CNN to learn features that can predict the spatial reduced reference entropic differences (SRRED) and temporal reduced reference entropic differences (TRRED) from frames and frame differences. The dual CNN setup for SRRED and TRRED is trained on synthetically distorted (compression, transmission and noise) videos, as shown in Fig. 1. We concatenate the intermediate layer output of the two CNNs and refer to the resulting features as STED features. Although STED features are trained to predict SRRED and TRRED for synthetic distortions, they contain some latent representations of video quality that can be leveraged for predicting the quality of authentically distorted videos. Thus, one can interpret our use of STED features as a form of transfer learning.

2) Video-Level Feature Model: Handcrafted or CNN-based video level features denoted as t are passed through a fully connected network \( g(\cdot) \) as above to estimate video quality:

\[
\hat{q}(v) = g(t).
\]

(3)

Examples of such video-level features include TTVQM [18] and VIDEVAL [26]. Note that these features are obtained by considering the video as a whole and not by averaging frame-level features. TTVQM [18] addresses the potential inefficiencies of NSS and designs heuristic features to capture blockiness, sharpness extremes, and camera shake. In our framework, we learn only the mapping \( g(\cdot) \) through semi-supervised learning.

3) Hybrid Feature Model: When both frame- and video-level features are present, we combine the above two models. In general, we concatenate video-level features t with frame-level features \( z \) to estimate overall video quality. The pair \( (z, t) \) is then fed to \( g(\cdot) \) to predict the perceptual quality of the video as follows:

\[
\hat{q}(v) = g(z, t).
\]

(4)

While hybrid features such as CNN-TLVQM [29] and PatchVQ [33] exist in the literature, we propose a combination of NSS-based frame-level and non-NSS-based video-level features. The features in the TTVQM [18] approach are complementary to NSS-based STED features. We refer to STED features as NSS-based because they are designed to predict SRRED and TRRED, which in turn are based on NSS models. In this setup, both \( f(\cdot) \) and \( g(\cdot) \) are learned during semi-supervised learning.

In the remainder of this section, we describe our SSL formulation using this hybrid model, which contains both the frame-level and video-level features. We compare the performance of our STED-TLVQM hybrid feature model with other frame-level and video-level features in Section IV-C.

D. Semi-Supervised Learning by Generating Consistent Pseudo-Rank Pairs

To leverage the unlabelled videos for effective SSL, we consider a student-teacher-based consistency regularization approach, where two models are maintained, a student model and a teacher model. In the case of generic consistency regularization [50], a consistency constraint is applied between the student and teacher model predictions on augmented unlabelled samples. The above learning strategy has two major drawbacks: 1) existing augmentations in the literature, such as noisy input or photometric transformed data [50], may not be appropriate for the VQA task because they can alter video quality, and 2) using the teacher model’s prediction as weak supervision for the student model may produce a confirmation bias due to the teacher’s noisy prediction. In this work, we propose a novel quality-invariant strong-weak augmentation strategy and provide pseudo-rank-based supervision for students to counter confirmation bias.

We design our augmentations based on video subsampling of the frames. Recent VQA works such as TTVQM [18], VISION [44], and VIDEVAL [26] have shown that the quality estimated from video frames subsampled to as low as 1 frame per second (fps) is approximately equivalent to the video quality at the full frame rate. Thus, videos subsampled at different frame rates are quality invariant. The frame-level features temporally subsampled at 1 frame per second (referred to as strong augmentation) are input to the student model. The target pseudo-label is provided by the teacher model, which makes a prediction based on the frame-level features, temporally subsampled at half the original frame rate (referred to as weak augmentation).

To counter confirmation bias in the teacher model’s prediction on the weakly augmented videos, we generate pseudo-ranks instead of pseudo-labels for a pair of unlabelled videos. Our hypothesis is that if the teacher’s quality predictions of two unlabelled videos \( u_1 \) and \( u_2 \) differ by greater than a threshold \( \tau \), then the pairwise quality ranking of the videos inferred from their predicted qualities is likely to be correct. Thus, we generate pairwise pseudo-ranks of the unlabelled videos in terms of their qualities predicted by the teacher model and use these ranks to supervise the student model.

Mathematically, we create two models \( q_s \) and \( q_t \) corresponding to the student and teacher models similar to (4), where \( \theta_s \) and \( \theta_t \) are their corresponding parameters. Initially, both models are identical. For any video \( u \), its quality prediction using the student and teacher models are obtained as \( q_s(u) \) and \( q_t(u) \). If video \( u \) has a frame rate \( r \), let \( T_s(u) \) and \( T_w(u) \) be the strong and weak augmentation functions that select frames at frame rates of 1 fps and \( r/2 \) fps, respectively. For a pair of videos \( u_1 \) and \( u_2 \), with \( u_1, u_2 \in U \) and \( |q_s(T_w(u_1)) - q_t(T_w(u_2))| > \tau \), we define the pairwise pseudo-ranking as

\[
r(u_1, u_2) = \begin{cases} 1 & \text{if } q_t(T_w(u_1)) \geq q_t(T_w(u_2)) \\ 0 & \text{otherwise} \end{cases}.
\]

(5)

We update the student model’s parameters using the available labels and the pseudo-rank pairs. In particular, the student model...
is trained to ensure that it satisfies the video rankings according to the pseudo-rank pairs generated by the teacher model. Enforcing the student model prediction to match the pairwise pseudo-ranks of videos generated by the teacher model has two benefits. First, this approach achieves consistency regularization by being invariant to the different augmentations that are applied as input to student-teacher models. Second, the student learning of hard binary quality ranks of unlabelled video pairs provided by the teacher model achieves entropy minimization through confidence maximization in pseudo-labeling-based SSL approaches [55]. Thus, our model combines the benefits of both consistency regularization and pseudo-labeling.

Using the generated pairwise pseudo-ranks, we deploy rank-based learning that has been widely studied in the literature [56], [57], where a Siamese network is used to predict the target from a pair of data samples. We use the quality predictions $q_u(T_u(u_1))$ and $q_u(T_u(u_2))$ of videos $u_1$ and $u_2$ satisfying $q_t(T_u(u_1)) - q_t(T_u(u_2)) > \tau$ to compute the probability of $q_u(T_u(u_1)) > q_u(T_u(u_2))$ as

$$\tilde{r}(u_1, u_2) = \sigma(q_u(T_u(u_1)) - q_u(T_u(u_2)))$$

$$= \frac{\exp(q_u(T_u(u_1)) - q_u(T_u(u_2)))}{1 + \exp(q_u(T_u(u_1)) - q_u(T_u(u_2)))}. \quad (6)$$

Thus, the unsupervised loss for training the student model on the unlabelled set $\mathcal{U}$ is given as

$$\mathcal{L}_u = \sum_{(u_1, u_2) \in \mathcal{U}} \mathcal{L}_{cross}(\tilde{r}(u_1, u_2), \tilde{r}(u_1, u_2))_{|q_u(T_u(u_1)) - q_u(T_u(u_2)) > \tau} \quad (7)$$

where $\mathcal{L}_{cross}(p_1, p_2)$ is the binary cross entropy defined as

$$\mathcal{L}_{cross}(p_1, p_2) = -p_1 \log p_2 - (1 - p_1) \log(1 - p_2). \quad (8)$$

For every video $v$ in the labelled set $\mathcal{V}$, its quality prediction using the student model is obtained as $q_s(T_u(v))$, and the corresponding ground truth is denoted as $q(v)$. The supervised loss on the labelled set is given as

$$\mathcal{L}_s = \sum_{v \in \mathcal{V}} |q(v) - q_s(T_u(v))|. \quad (9)$$

The optimization problem for learning the student network is given as

$$\min_{\theta_s} (\mathcal{L}_s + \lambda \mathcal{L}_u). \quad (10)$$

where $\lambda$ represents the relative weight between the two losses. $\lambda$ is chosen such that the order of magnitude of the supervised and unsupervised loss terms are similar so that the unsupervised loss cannot overpower the effect of supervised loss.

Suppose the parameters of the teacher model and student model at iteration $n$ are given by $\theta_t^{(n)}$ and $\theta_s^{(n)}$. The student model parameters are updated as in (10), while the teacher model is then updated as a moving average of successive student models similar to Mean Teacher [19] as

$$\theta_t^{(n)} = \alpha \theta_t^{(n-1)} + (1 - \alpha) \theta_s^{(n)}. \quad (11)$$

In principle, because the teacher model is updated at every iteration, the pseudo-rank pairs must be updated every iteration. To limit the computational overhead of generating the rank pairs of all the unlabelled videos at every iteration, we update the pseudo-ranks after every $K$ training iterations. However, the teacher model is updated after every iteration.

IV. EXPERIMENTS AND RESULTS

A. Databases

We evaluate our semi-supervised video quality learning method on three popular authentically distorted VQA datasets described as follows:

1) KoNViD-1K [58]: This dataset contains 1200 videos with a wide variety of content, distortion types and subjective quality variations. The videos have a resolution of $960 \times 540$, a frame rate of 24, 25 or 30 frames per second and are 8 seconds in duration.

2) LIVE Video Quality Challenge (VQC) Database [59]: The LIVE VQC database consists of 585 videos of unique content available at 18 different spatial resolutions ranging between $1980 \times 1080$ and $320 \times 240$ across landscape and portrait modes. All the videos are 10 seconds long.

3) LIVE Qualcomm Database [60]: This database consists of 208 videos accounting for distortions generated during the camera capture process using eight mobile devices. The videos have a spatial resolution of $1920 \times 1080$ and are 15 seconds long when played at 30 fps.

Similar to [29], because our focus is authentically distorted videos through camera capture, we omit the YouTube UGC dataset [26] as it contains a large proportion of artificially generated content in the form of animations and computer graphics.

For learning STED features as described in Section III-C1, we use several synthetic databases such as the LIVE Mobile VQA dataset [61], LIVE VQA dataset [62], EPFL-Polimi dataset [63], ECVQ and EVVQ datasets [64] and the CSIQ database [65]. In particular, we used only the videos from these synthetic datasets and no subjective scores. The features are learned to predict the SRRED and TRRED on these videos because a reference video is available in all these synthetic datasets.

B. Experimental Setting

Semi-supervised methods are typically evaluated by treating most of the dataset as unlabelled and a small part of the dataset as labelled. We first divide the dataset into training and testing at ratios of 80% and 20%. We evaluate the performance when only 30, 60 and 120 videos belonging to the training set are labelled in the form of mean opinion scores. Furthermore, the videos with labels are randomly sampled from the training set. We conduct our experiments on ten different splits of the dataset into training and testing and report the median performance.

We evaluate the performance of VQA methods using conventional measures such as Spearman’s rank order correlation coefficient (SROCC) and the Pearson linear correlation coefficient (PLCC) between the predicted quality scores and the ground truth quality scores.
C. Performance Analysis of Quality Features in Limited Labeled Data Regime

We first conduct an experiment where we compare different video quality features using the limited labeled data and supervised learning. Thus, unlabeled data are not used during training in this experiment. The goal of this analysis is to identify features that perform best under the limited labeled data regime. We believe that features that work well in this regime can be bootstrapped to most improve performance with semi-supervised learning.

1) Benchmarking Quality Aware Features: We find that learning a CNN from scratch with limited data for VQA yields poor performance. Thus, we compare our method with recent heuristic and pretrained CNN-based methods for VQA with limited labels. In particular, we compare the STED-TLVQM features described in Section III-C with heuristic feature-based methods such as Video BLIINDS [5], VIDEVAL [26], and TLVQM [18]. While HEKE [15], VSFA [8], and TCSVT-BVQA [39] are frame-level feature models, FAST-VQA [37] and RAPIQUE [30] are video-level feature models. We also perform a comparison with recent hybrid feature-based methods such as CNN-TLVQM [29] and PatchVQ [33].

2) Training Details: While STED is trained on synthetically distorted videos, the learned features are then used along with handcrafted TLVQM features in STED-TLVQM. In STED, we train the spatial feature extraction network using SRRED for 20 epochs with a batch size of 16 and the Adam [66] optimizer. Because we train the temporal network from scratch on the synthetic videos and the video frames at the original resolution are used as input, a batch size of 8 is chosen to train this network using TRRED for 30000 iterations. Note that our spatial and temporal feature learning framework allows us to train with videos of any resolution. The trained STED model is used to train this model. Neither the augmentations nor the new teacher to generate pseudo-labels. FixMatch [53] uses a photometric transformation-based strong-weak augmentation strategy on a student-teacher-based model. Any photometric...
augmentations used in the above methods for the VQA task were removed.

2) Semi-Supervised Training Details: Initially, the network parameters corresponding to \( f(\cdot) \) and \( g(\cdot) \) are trained for 1000 iterations using just the supervised loss as in (9) and SGD with an initial learning rate of \( 10^{-1} \), decay rate of \( 10^{-2} \) and momentum of 0.9. We then incorporate the pairwise pseudo-rank-based loss to fine-tune the parameters of the student model corresponding to \( f(\cdot) \) and \( g(\cdot) \) with \( \lambda = 0.1 \). These parameters are trained for 1000 iterations with a pseudo-rank update for the unlabelled data after every \( K = 50 \) iterations. The teacher model used to update the pseudo-ranks has a smoothing coefficient \( \alpha = 0.9 \) referred to in (11). As in the previous stage, we employ a batch size of 32 for the 60- and 120-label cases and 16 for the 30-label case. Because the predicted quality score lies in the 0 to 1 scale, we choose a threshold of \( \tau = 0.1 \) to select pairs of videos with pseudo-rank labels.

3) Performance Comparisons: We compare the performance of our method LPR against PL [13], MT [19], NS [52], and FM [53] on the KoNVid-1 K, LIVE VQC, and LIVE Qualcomm datasets in Tables III and IV. We find that LPR not only outperforms other semi-supervised methods on the three authentically distorted databases but also shows considerable improvement over the baseline supervised model trained on low data. We also conduct statistical significance tests to validate the importance of the correlation coefficient differences observed in Tables III and IV. These results are given in the supplementary material.

4) Cross-Database Performance Analysis: To analyse the generalization performance of our semi-supervised models, we
Fig. 3. Framework of our semi-supervised learning approach LPR on authentically distorted camera captured videos. v is a video belonging to the labelled set \( V \), and \( u_i \) and \( u_j \) are a pair of videos belonging to the unlabelled set \( U \). We extract frame-level STED representations of a strongly and weakly augmented version of a video belonging to an unlabelled set \( U \). Similarly, we extract video-level TLVQM representations from the video. While the teacher model generates a pseudo-rank from this pair of unlabelled features, the student model learns to predict the rank of a video pair provided by the teacher. Note that the weights of the student models are shared.

TABLE V
GENERALISED SROCC PERFORMANCE ANALYSIS OF LPR WITH VARIOUS QUALITY AWARE FEATURE BASED ALGORITHM ON KoNVid-1K, LIVE VQC, AND LIVE QUALCOMM DATASETS

| Features (\( \cdot \)) | KoNVid-1K | LIVE VQC | LIVE Qualcomm |
|------------------------|----------|----------|---------------|
| \( f(\cdot) \) | 30 labels | 60 labels | 120 labels |
| | 30 labels | 60 labels | 120 labels |
| TVQM | 0.524 (0.033) | 0.599 (0.023) | 0.663 (0.027) | 0.588 (0.047) | 0.641 (0.054) | 0.663 (0.053) | 0.659 (0.042) | 0.577 (0.033) | 0.752 (0.039) |
| VIDEVAL | 0.513 (0.050) | 0.563 (0.043) | 0.617 (0.024) | 0.558 (0.025) | 0.634 (0.061) | 0.677 (0.063) | 0.688 (0.086) | 0.577 (0.085) | 0.663 (0.092) |
| RAPIQUE | 0.549 (0.051) | 0.633 (0.063) | 0.694 (0.059) | 0.573 (0.032) | 0.659 (0.048) | 0.708 (0.071) | 0.452 (0.081) | 0.560 (0.073) | 0.661 (0.077) |
| HEIKE | 0.516 (0.053) | 0.550 (0.046) | 0.623 (0.057) | 0.486 (0.051) | 0.540 (0.050) | 0.613 (0.044) | 0.442 (0.067) | 0.537 (0.041) | 0.623 (0.094) |
| CNN-TLVQM | 0.580 (0.041) | 0.670 (0.038) | 0.693 (0.040) | 0.591 (0.058) | 0.667 (0.071) | 0.686 (0.068) | 0.394 (0.047) | 0.578 (0.035) | 0.693 (0.038) |
| PatchVQ | 0.543 (0.019) | 0.591 (0.025) | 0.632 (0.018) | 0.541 (0.019) | 0.592 (0.030) | 0.616 (0.034) | 0.434 (0.047) | 0.538 (0.043) | 0.607 (0.026) |
| FASTVQA | 0.564 (0.022) | 0.640 (0.024) | 0.673 (0.033) | 0.556 (0.020) | 0.619 (0.022) | 0.661 (0.023) | 0.468 (0.075) | 0.552 (0.052) | 0.667 (0.059) |
| TCSVT-BVQA | 0.583 (0.064) | 0.656 (0.064) | 0.704 (0.063) | 0.594 (0.058) | 0.643 (0.066) | 0.669 (0.041) | 0.487 (0.077) | 0.637 (1.130) | 0.731 (0.077) |
| STED-TLVQM | 0.675 (0.059) | 0.708 (0.043) | 0.750 (0.051) | 0.667 (0.064) | 0.709 (0.051) | 0.751 (0.073) | 0.577 (0.083) | 0.664 (0.073) | 0.794 (0.073) |

The numbers in brackets indicate the increment in performance by learning on unlabelled data using LPR on various feature based algorithms.

conduct cross-database experiments and compare them with other SSL methods such as MT [19], FM [53], NS [52], and PL [13]. We take the models trained using a few labelled samples of one database and test them on a different database. In total, we have six train-test settings across the KoNVid-1K [58], LIVE VQC [59], and LIVE Qualcomm [60] databases. For each of these settings, we train the models for 30, 60, and 120 label cases and report the results in Fig. 4. Our LPR model achieves superior performance over other semi-supervised algorithms in all settings. We provide the baseline performance in Fig. 4 to show the relative gain in performance with unlabelled videos for different SSL methods. We also note an improvement in SROCC values as the number of labels for supervision increases. This trend is consistent with the improvement observed in the same database testing scenarios.

E. Learning Pseudo-Ranks With Different Quality Representations

We now study the relevance of our LPR SSL model on various quality-aware features described in Section IV-C. Both \( f(\cdot) \) and \( g(\cdot) \) are trained for frame-level or hybrid model features, while for video-level features, \( f(\cdot) \) does not exist. In Table V, we report
the SROCC performance of LPR using the quality-aware features under different settings of the number of labelled videos. We also show the incremental performance gain due to SSL by comparing it with the corresponding supervised learning with limited labels. Our SSL approach consistently improves performance on all feature representations, showing its stability.

V. ABLATION STUDIES

A. Impact of Various Components in LPR

We first evaluate the need for each of the main contributions of our work in SSL, particularly the need for augmentations, learning on pseudo-ranks and the threshold on the difference in the predicted scores of the teacher model to obtain reliable pseudo-ranks. In Table VI, we report the results of an ablation experiment on the KoNVid-1K, LIVE VQC, and LIVE Qualcomm databases for the 30 labelled video scenarios.

W/O Augmentation: We train LPR on the unlabelled samples without augmenting the input to either the student or teacher model. The student model is trained to predict the pseudo-ranks generated by the teacher. Note that the teacher is obtained as an exponential moving average of the student model parameters.

W/O Pseudo-rank: In this setup, we modify the student-teacher model to learn pseudo-labels of the unlabelled videos rather than the pseudo-ranks of a video pair. Thresholding does not apply in this case, as pseudo-label of each unlabelled example are directly learned here. This experiment proves the need for our generation of pseudo-ranks.

W/O Thresholding: Here, the threshold $\tau$ is taken to be 0. Thus, the model is trained with all the unlabelled video pairs. This experiment studies the need for obtaining reliable pseudo-ranks by thresholding the difference in quality predictions.

W/O Moving Average: In this experiment, we update the teacher weights with current student weights rather than the exponential moving average of past and present student weights. Here, the student and teacher models are identical, and their outputs in the respective augmentations must be consistent. The gradients pertaining to the student model prediction loss are not propagated back to the teacher.

Table VI shows that all the components of our model are important. However, the generation of reliable pseudo-ranks is extremely important, as performance drops significantly without this component. We observe that these trends are fairly consistent across all datasets.

B. Analysis of Hyperparameters

We now analyse our model with respect to the threshold $\tau$ and how the performance on the unlabelled data improves with training. In Fig. 5(a), we first analyse how the performance (SROCC) of our model varies with respect to the choice of the threshold $\tau$. We present the results when 30 labels are available in the respective datasets. While a threshold $\tau = 0$ implies that all the pseudo-rank pairs are selected, a very high choice of threshold implies that very few pairs are selected. We observe that a
choice of $\tau = 0.1$ yields optimum performance across different datasets. While the drop in performance is steady for $\tau > 0.1$ on the LIVE VQC and KoNViD-1 K datasets, there is a steeper drop in performance on the LIVE Qualcomm database. The LIVE Qualcomm database is a smaller database, and with larger values of $\tau$, very few unlabelled videos that satisfy the threshold criterion are selected, which can bias training and decrease performance.

In Fig. 5(b), we track the performance of our model on the unlabelled data as training proceeds. This experiment aims to understand how the model improves with training, ultimately leading to its superior performance on the testing dataset. We evaluate the accuracy of the pseudo-ranks for unlabelled video pairs with respect to the true ranks for these pairs. In all three datasets, we see that as learning proceeds, the fraction of unlabelled video pairs with the correct pseudo-ranks keeps improving. Nevertheless, some saturation is seen as the training proceeds beyond a point.

The student network is fed with strongly augmented features with severe subsampling of the STED features as input. In Fig. 5(c), we vary these frame rates from 0.25 to 4 frames per second and record the performance. We see that subsampling the input of less than 1 frame per second (fps) impacts performance due to a significant reduction in the frame level feature information. As performance is fairly consistent over the range of 1–4 fps across all the datasets, we choose a subsampling rate of 1 fps to strongly input the frame-level quality features. Note that as the frame rate increases, the effect of the strong-weak augmentation decreases.

C. Complementarity of Quality Aware Features

We perform an error-based complementarity study on the STED and TLVQM features in predicting video quality scores. We regress the TLVQM [18] features and STED-based features against the ground truth MOS on 80% data for each of the three authentically distorted video databases. We then compute the absolute error between the predicted quality and the MOS of the remaining 20% test videos. Fig. 6 shows the scatter plot between the absolute error in predicting the quality using the TLVQM [18] and STED features. The plot is divided into four quadrants based on whether the individual error is greater or less than a 20% of the MOS range of that particular dataset. The two sets of features yield complementary predictions for certain samples, and thus, combining them can improve overall model performance, as evident from Tables I and II.

D. Comparison With Unsupervised VQA

We also compare our SSL-based LPR method against various unsupervised VQA algorithms in the literature. This analysis shows that the inclusion of a few labelled videos in SSL can bridge the gap between supervised and unsupervised VQA methods. In Table VII, we compare our method STED-TLVQM + LPR with a classical unsupervised VQA method such as NIQE [49] and more recent methods such as STEM [46], VlIQE [47], TPOQI [48], and VISION [44]. Despite training with only 30 labels, our method outperforms these unsupervised methods on most authentically distorted databases. Additionally, an increase in the number of labelled videos results in a
performance gain. While unsupervised methods largely capture distortion-specific quality variations, human ratings are also influenced by other aspects, such as aesthetics and scene understanding. Thus, we conclude that LPR can bridge the gap in performance between supervised and unsupervised methods by learning with limited human opinion.

VI. CONCLUSION

We designed a framework for NR VQA of authentically distorted videos when only limited labels are available for training a video quality model. We showed the effective use of unlabelled videos by generating pairwise pseudo-ranks with student-teacher models on strong-weak augmented videos and using such ranks to improve the model. While we showed the utility of our learning approach on different features, we also presented a particular feature model for spatial and temporal features learned with spatiotemporal entropic differences. Our framework shows that the performance on authentically distorted videos can be significantly improved in terms of correlation with human perception, even when only a few videos are labelled with human opinion scores.

REFERENCES

[1] K. Seshadrinathan and A. C. Bovik, “Motion tuned spatio-temporal quality assessment of natural videos,” IEEE Trans. Image Process., vol. 19, no. 2, pp. 335–350, Feb. 2010.
[2] P. Y. Yu, C. T. Vu, and D. M. Chandler, “A spatiotemporal most-apparent-distortion model for video quality assessment,” in Proc. 18th IEEE Int. Conf. Image Process., 2011, pp. 2505–2508.
[3] Z. Wang, E. Simoncelli, and A. Bovik, “Multiscale structural similarity for image quality assessment,” in Proc. 37th Asilomar Conf. Signals Syst. Comput., 2003, pp. 1398–1402.
[4] R. Soundararajan and A. C. Bovik, “Video quality assessment by reduced reference spatio-temporal entropic differencing,” IEEE Trans. Circuits Syst. Video Technol., vol. 23, no. 4, pp. 684–694, Apr. 2013.
[5] M. A. Saad, A. C. Bovik, and C. Charrier, “Blind prediction of natural video quality,” IEEE Trans. Image Process., vol. 23, no. 3, pp. 1352–1365, Mar. 2014.
[6] W. Liu, Z. Duamna, and Z. Wang, “End-to-end blind quality assessment of compressed videos using deep neural networks,” in Proc. 26th ACM Int. Conf. Multimedia, 2018, pp. 546–554, doi: 10.1145/3240508.3240643.
[7] J. You and J. Korhonen, “Deep neural networks for no-reference video quality assessment,” in Proc. IEEE Int. Conf. Image Process., 2019, pp. 2349–2353.
[8] D. Li, T. Jiang, and M. Jiang, “Quality assessment of in-the-wild videos,” in Proc. 27th ACM Int. Conf. Multimedia, 2019, pp. 2351–2359, doi: 10.1145/3343031.3351028.
[9] H. Tang, N. Joshi, and A. Kapoor, “Blind image quality assessment using semi-supervised rectifier networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014, pp. 2877–2884.
[10] Z. Wang, D. Li, and K. Ma, “Semi-supervised deep ensembles for blind video quality assessment,” 2021, arXiv:2106.1408.
[11] G. Yue et al., “Semi-supervised authentically distorted image quality assessment with consistency-preserving dual-branch convolutional neural network,” IEEE Trans. Multimedia, vol. 25, pp. 6499–6511, 2023.
[12] Q. Xie et al., “Unsupervised data augmentation for consistency training,” in Proc. Adv. Neural Inf. Process. Syst., 2020, pp. 6256–6268.
[13] D.-H. Lee, “Pseudo-label: The simple and efficient semisupervised learning method for deep neural networks,” in Proc. ICML Workshop Challenges Representation Learn., 2013, Art. no. 896.
[14] A. Rasmus, H. Valpola, M. Honkala, M. Berglund, and T. Raiko, “Semi-supervised learning with ladder networks,” in Proc. 28th Int. Conf. Neural Inf. Process. Syst., 2015, pp. 5546–5554.
[15] Y. Liu et al., “Spatiotemporal representation learning for blind video quality assessment,” IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 6, pp. 3500–3513, Jun. 2022.
[16] S. Mitra, R. Soundararajan, and S. S. Channappayya, “Predicting spatio-temporal entropic differences for robust no reference video quality assessment,” IEEE Signal Process. Lett., vol. 28, pp. 170–174, 2021.
[17] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, “Gradient magnitude similarity deviation: A highly efficient perceptual image quality index,” IEEE Trans. Image Process., vol. 23, no. 2, pp. 684–695, Feb. 2014.
[18] J. Korhonen, “Two-level approach for no-reference consumer video quality assessment,” IEEE Trans. Image Process., vol. 28, no. 12, pp. 5923–5938, Dec. 2019.
[19] A. Tarvainen and H. Valpola, “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results,” in Proc. 31st Int. Conf. Neural Inf. Process. Syst., 2017, pp. 1195–1204.
[20] X. Li, Q. Guo, and X. Lu, “Spatiotemporal statistics for video quality assessment,” IEEE Trans. Image Process., vol. 25, no. 7, pp. 3329–3342, Jul. 2016.
[21] S. V. Reddy Dendi and S. S. Channappayya, “No-reference video quality assessment using natural spatiotemporal scene statistics,” IEEE Trans. Image Process., vol. 29, pp. 5612–5624, 2020.
[22] K. Manasa and S. S. Channappayya, “An optical flow-based no-reference video quality assessment algorithm,” in Proc. IEEE Int. Conf. Image Process., 2016, pp. 2400–2404.
[23] J. E. Caviedes and F. Oberti, “No-reference quality metric for degraded and enhanced video,” in Visual Communications and Image Processing, T. Ebrahimi and T. Sikora, Eds., vol. 5150. Bellingham, WA, USA: SPIE, 2003, pp. 621–632, doi: 10.1117/12.510112.
[24] M. Farias and S. Mitra, “No-reference video quality metric based on artifact measurements,” in Proc. IEEE Int. Conf. Image Process. 2005, pp. III–141.
[25] F. Yang, S. Wan, Y. Chang, and H. R. Wu, “A novel objective no-reference metric for digital video quality assessment,” IEEE Signal Process. Lett., vol. 12, no. 10, pp. 685–688, Oct. 2005.
[26] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, “UGC-VQA: Benchmarking blind video quality assessment for user generated content,” IEEE Trans. Image Process., vol. 30, pp. 4449–4464, 2021.
[27] Y. Li et al., “No-reference video quality assessment with 3D shearlet transform and convolutional neural networks,” IEEE Trans. Circuits Syst. Video Technol., vol. 26, no. 6, pp. 1044–1057, Jun. 2016.
[28] S. Ahn and S. Lee, “Deep blind video quality assessment based on temporal human perception,” in Proc. IEEE 25th Int. Conf. Image Process., 2018, pp. 619–623.
[29] J. Korhonen, Y. Su, and J. You, “Blind natural video quality prediction via statistical temporal features and deep spatial features,” in Proc. 28th ACM Int. Conf. Multimedia, 2020, pp. 3311–3319, doi: 10.1145/3394171.3413845.
[30] Z. Tu et al., “RAPIQUE: Rapid and accurate video quality prediction of user generated content,” IEEE Open J. Signal Process., vol. 2, pp. 425–440, 2021.
[31] P. Chen, L. Li, L. Ma, J. Wu, and G. Shi, “RIRNet: Recurrent-in-recurrent network for video quality assessment,” in Proc. 28th ACM Int. Conf. Multimedia, 2020, pp. 834–842, doi: 10.1145/3394171.3413717.
[32] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.
[33] Z. Ying, M. Mandal, D. Ghadiyaram, and A. Bovik, “Patch-VQ: Patching up the video quality problem,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 14019–14029.
[34] Z. Ying et al., “From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 3572–3582, doi: 10.1109/CVPR42600.2020.00363.

[35] F. Göktekin, B. A. A. H. Lin, and D. Saake, “KonVid-150K: A dataset for no-reference video quality assessment of videos in-the-wild,” IEEE Access, vol. 9, pp. 72193–72160, 2021.

[36] W. Shen et al., “An end-to-end no-reference video quality assessment method with hierarchical spatiotemporal feature representation,” IEEE Trans. Broadcast., vol. 68, no. 3, pp. 651–660, Sep. 2022.

[37] H. Wu et al., “FAST-VQA: Efficient end-to-end video quality assessment with fragment sampling,” in Proc. Eur. Conf. Comput. Vis., 2022, pp. 538–554.

[38] Z. Liu et al., “Swin transformer: Hierarchical vision transformer using shifted windows,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 9992–10002, doi: 10.1109/ICCV48922.2021.00986.

[39] B. Li, W. Zhang, M. Tian, G. Zhai, and X. Wang, “Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception,” IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 9, pp. 5944–5958, Sep. 2022.

[40] Y. Zhang, X. Gao, L. He, W. Lu, and R. He, “Blind video quality assessment with weakly supervised learning and resampling strategy,” IEEE Trans. Circuits Syst. Video Technol., vol. 29, no. 8, pp. 2244–2255, Aug. 2019.

[41] P. Chen, L. Li, J. Wu, W. Dong, and G. Shi, “Unsupervised curriculum domain adaptation for no-reference video quality assessment,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 5178–5187.

[42] J. Xu, P. Ye, Y. Liu, and D. Doermann, “No-reference video quality assessment via feature learning,” in Proc. IEEE Int. Conf. Image Process., 2019, pp. 491–495.

[43] P. Chen, L. Li, J. Wu, W. Dong, and G. Shi, “Contrastive self-supervised pre-training for video quality assessment,” IEEE Trans. Image Process., vol. 31, pp. 458–471, 2022.

[44] S. Mitra and R. Soundararajan, “Multiview contrastive learning for completely blind video quality assessment of user generated content,” in Proc. 30th ACM Int. Conf. Multimedia, 2022, pp. 1914–1924, doi: 10.1145/3503161.3548064.

[45] A. Mittal, M. A. Saad, and A. C. Bovik, “A completely blind video integrity oracle,” IEEE Trans. Image Process., vol. 25, no. 1, pp. 289–300, Jan. 2016.

[46] P. Kancharla and S. S. Channappayya, “Completely blind quality assessment of user generated video content,” IEEE Trans. Image Process., vol. 31, pp. 263–274, 2022.

[47] Q. Zheng, Z. Tu, X. Zeng, A. C. Bovik, and Y. Fan, “A completely blind video quality evaluator,” IEEE Signal Process. Lett., vol. 29, pp. 2228–2232, 2022.

[48] L. Liao et al., “Exploring the effectiveness of perceptual representation in blind video quality assessment,” in Proc. 30th ACM Int. Conf. Multimedia, 2022, pp. 837–846, doi: 10.1145/3503161.3548749.

[49] R. Soundararajan and A. C. Bovik, “Making a "completely blind" image quality analyzer,” IEEE Signal Process. Lett., vol. 20, no. 3, pp. 209–212, Mar. 2013.

[50] Y. Yang, Z. Song, I. King, and Z. Xu, “A survey on double supervision learning,” IEEE Trans. Knowl. Data Eng., vol. 35, no. 9, pp. 8934–8954, Sep. 2023.

[51] H. Pham, Z. Dai, Q. Xie, and Q. V. Le, “Meta pseudo labels,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 11557–11568.

[52] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, “Self-training with noisy student improves ImageNet classification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 10684–10695.

[53] K. Sohn et al., “FixMatch: Simplifying semi-supervised learning with consistency and confidence,” in Proc. 34th Int. Conf. Neural Inf. Process. Syst., 2020, pp. 596–608.

[54] A. C. Bovik, R. Soundararajan, and C. G. Bampis, “On the robust performance of the ST-RRED video quality predictor,” 2017, Online. Available: https://live.ece.utexas.edu/research/Quality/ST-RRED

[55] M. Sajjadi, M. Javanmardi, and T. Tasdizen, “Regularization with stochastic transformations and perturbations for deep semi-supervised learning,” in Proc. 30th Int. Conf. Neural Inf. Process. Syst., 2016, pp. 1171–1179.

[56] C. Burges et al., “Learning to rank using gradient descent,” Proc. 22nd Int. Conf. Mach. Learn., 2005, pp. 89–96, doi: 10.1145/1102351.1102363.

[57] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, “dipQ: Blind image quality assessment by learning-to-rank discriminative image pairs,” IEEE Trans. Image Process., vol. 28, no. 8, pp. 3951–3964, Aug. 2017.

[58] V. Hou et al., “The Konstanz natural video database (KonViDV-1K),” in Proc. IEEE 9th Int. Conf. Qual. Multimedia Experience, 2017, pp. 1–6.