Unsupervised Curriculum Domain Adaptation for No-Reference Video Quality Assessment

Pengfei Chen[1,2] Leida Li[1,3,*] Jinjian Wu[1] Weisheng Dong[1] Guangming Shi[1]
[1] Xidian University [2] China University of Mining and Technology [3] Pazhou Lab
cpf00790079@gmail.com, {ldli, gmshi}@xidian.edu.cn, {jinjian.wu, wsdong}@mail.xidian.edu.cn

Abstract

During the last years, convolutional neural networks (CNNs) have triumphed over video quality assessment (VQA) tasks. However, CNN-based approaches heavily rely on annotated data which are typically not available in VQA, leading to the difficulty of model generalization. Recent advances in domain adaptation technique makes it possible to adapt models trained on source data to unlabeled target data. However, due to the distortion diversity and content variation of the collected videos, the intrinsic subjectivity of VQA tasks hampers the adaptation performance. In this work, we propose a curriculum-style unsupervised domain adaptation to handle the cross-domain no-reference VQA problem. The proposed approach could be divided into two stages. In the first stage, we conduct an adaptation between source and target domains to predict the rating distribution for target samples, which can better reveal the subjective nature of VQA. From this adaptation, we split the data in target domain into confident and uncertain subdomains using the proposed uncertainty-based ranking function, through measuring their prediction confidences. In the second stage, by regarding samples in confident subdomain as the easy tasks in the curriculum, a fine-level adaptation is conducted between two subdomains to fine-tune the prediction model. Extensive experimental results on benchmark datasets highlight the superiority of the proposed method over the competing methods in both accuracy and speed. The source code is released at https://github.com/cpf0079/UCDA.

1. Introduction

Benefitting from the evolution of affordable and reliable consumer capture devices, and the tremendous popularity of social media platforms, recent years have witnessed an explosion of user-generated videos shared and streamed over the Internet [8]. Improving the efficiency of video coding, storage, and streaming over communication networks is a principle goal of video sharing and streaming platforms. One relevant research direction is the perceptual optimization of rate-distortion tradeoffs in video encoding and streaming, where distortion (or quality) is usually modeled using video quality assessment (VQA) algorithms that can predict human judgements of perceptual quality. This has motivated years of research on the topics of perceptual image and video quality assessment (IQA/VQA) [48, 36, 13, 22, 23, 12].

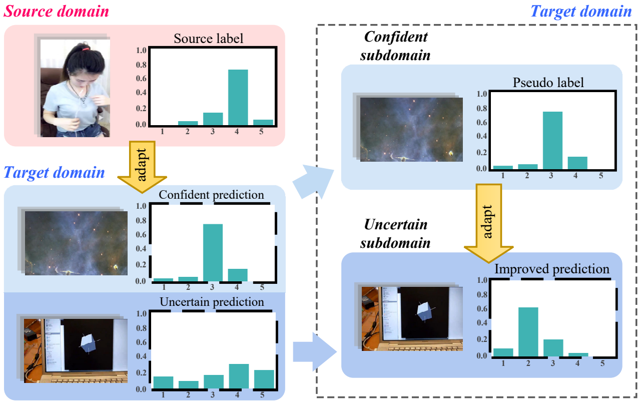

Figure 1. The pipeline of the proposed model could be divided into two stages, i) the model first learns domain-invariant features from labeled source data and unlabeled target data; ii) supervised by the pseudo labels obtained from the first stage, the performances of those uncertain predictions are rectified through the adaptation between two subdomains in the target domain.

Compared to the most reliable subjective VQA study relying on manual annotating, objective approach that automatically predicts the visual quality of distorted videos has long been a popular topic in VQA community. More recently, convolutional neural networks (CNNs) have become a hallmark backbone model to solve vision-related tasks. Despite of the impressive progress in developing objective VQA models, they often encounter challenges in practical applications, where prediction model trained on well-labeled data could be easily crippled given the target sam-

[*] Corresponding author

ples due to the domain disparity [21, 11]. This is because benchmark VQA datasets are usually biased to specific environments, and in practice, it is often hard to acquire new training sets that fully cover the huge variability of real-life test scenarios. Admittedly, generalizing models learned on one visual domain to novel domains has been a major obstacle to the development of VQA techniques.

A natural idea to improve the generalization ability of the trained model is to adopt domain adaptation technique that is invented to handle the cross-domain tasks by learning domain-invariant representations. It has experienced an impressive series of successes in many computer vision tasks such as semantic segmentation, classification and detection [31, 41], but remains a challenge for the particular task of VQA due to the subjective nature of quality assessment. Since videos collected from the testing scenario are bound to have diverse distortions and varied visual contents, part of them tend to be more difficult to be rated than others [25]. This leads to different difficulty degrees for the network to learn transferable knowledge for each target sample, resulting in uncertain predictions and inferior adaptation performance (Figure 1, where we assume that predictions with more concentrated distributions, which show unimodal patterns, have higher prediction confidences as mentioned in [25]).

The conventional wisdom of objective VQA methods is learning a regression model to predict the mean-opinion-score (MOS). However, the subjective nature of quality assessment progress may not be adequately represented by a single scalar number, considering such a scheme ignores the fact that the video to be evaluated would receive divergent opinions from different subjects [44]. This is particularly profound on complex, real-world distorted videos compared to their image counterparts. For instance, the average standard deviation of the subjective scores of the videos in the LIVE-VQC database [34] is 18 on the MOS scale of [0,100]. In this case, we argue that one possible way to better reveal this intrinsic subjectivity may reside in exploring the potentially useful and predictive information contained in the distributions of subjective scores, which has been rarely discussed or utilized in the literature.

In light of the above issues, in this paper, by casting the quality assessment task as a rating distribution prediction problem, we take steps toward a novel domain adaptation approach, dubbed as Unsupervised Curriculum Domain Adaptation (UCDA), to handle the cross-domain no-reference VQA task. The proposed curriculum-style adaptation could be performed in two stages as shown in Figure 1. In the first stage, feature distributions between the labeled source data and unlabeled target data are aligned to produce the prediction for target samples. Then, the target domain data is further split into two subdomains based on the their prediction confidences. In the second stage, a fine-level adaptation is further conducted between two subdomains in a self-supervised manner, aiming to improve the performance of those uncertain predictions in target data by enforcing high prediction confidence. The contributions of this work could be summarized with the following points:

- We propose a novel unsupervised domain adaptation approach for no-reference VQA task, where the rating distribution is used as the predict target to better reveal the intrinsic subjectivity of the quality assessment. To our best knowledge, this is an earlier attempt to explicitly highlight the transferable knowledge for VQA across different domains.
- We develop an uncertainty-based ranking function to sort the samples from target domain into different subdomains based on their prediction confidences, which are used to construct easy/hard tasks in the curriculum.
- We build a two-stage adversarial adaptation to improve the adaptation performance based on the designed curriculum. This is enabled by enforcing high prediction confidence on those uncertain predictions.

2. Related Works

Objective VQA methods can be divided into full reference (FR), reduced reference (RR), and no reference (N-R) in terms of the the availability of reference information. While entire or partial information of reference videos is attainable in FR/RR-VQA metrics [24, 35, 1, 42], NR-VQA metrics exploit distortion-specific or natural video statistical models without the participation of any information from original videos, which is the major advantage in practical applications and also the primary concern in this work.

Early NR-VQA metrics mainly focus on the distortion-specific problems, such as rate adaptation and motion blur [43, 3]. These metrics demonstrate the advantages for the specific distortions, but not for other situations. By contrast, general-purpose NR approaches aim to deal with diversified distortions. Saad *et al.* [32] proposed V-BLIINDS where a model in the discrete cosine transform (DCT) domain and a motion model that quantifies motion coherency were combined to predict video quality. Mittal *et al.* [28] proposed a metric called VIIDEO, which models the intrinsic statistical regularities to quantify disturbances introduced by distortions. Recently, Korhonen [19] selected a comprehensive feature set comprising of empirical motion statistics, specific artifacts, and aesthetics to build the two level video quality model, dubbed TLVQM. In [39], Tu *et al.* proposed a fusion-based VQA model called VIDEVAL, using a feature ensemble and selection procedure on top of existing efficient NR-VQA models.

Performance of NR-VQA models has been significantly boosted by end-to-end optimization of feature engineer-

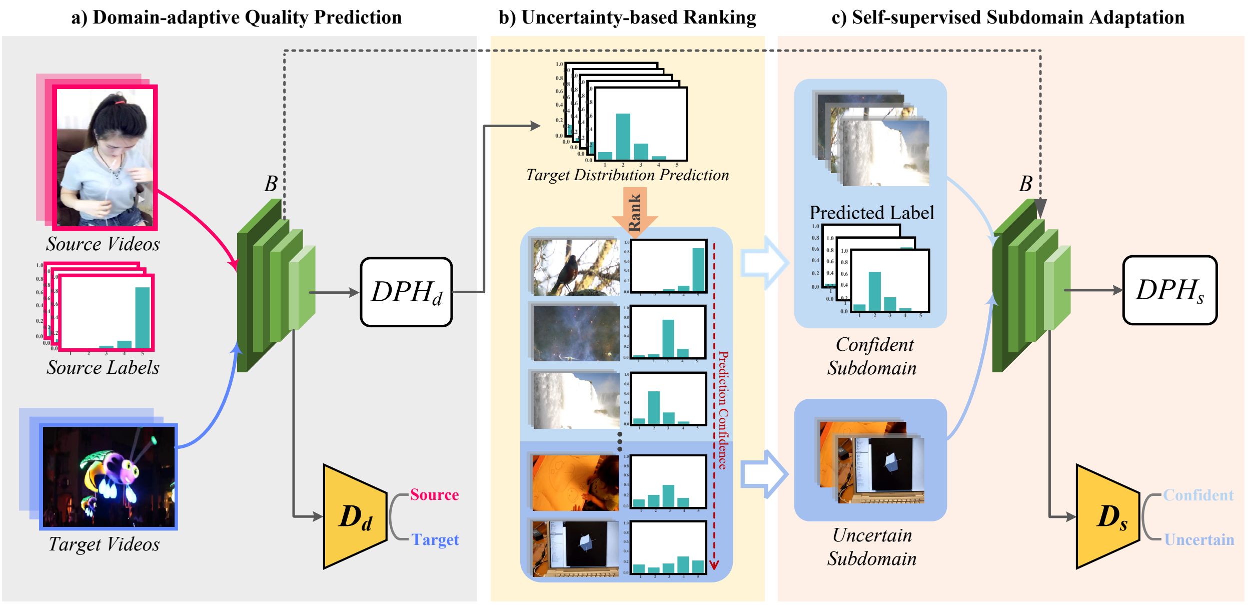

Figure 2. The overall structure of the proposed network. It consists of three parts: **a)** a domain-adaptive quality prediction network, **b)** an uncertainty-based ranking function, **c)** a self-supervised subdomain adaptation network. Given the labeled source data and unlabeled target data, DPH_d is trained to learn domain-invariant features via trying to fool D_d which is optimized to predict a domain label. Then, an uncertainty-based ranking function is introduced to split the target data into confident and uncertain subdomains based on their prediction confidences. Afterwards, DPH_s is further trained to improve the adaptation performance with the help of D_s, which is enabled by conducting adaptation between subdomains self-supervised by the subdomain labels.

ing and quality regression with the help of deep learning technique. Notably, Zhang *et al.* [47] applied weakly supervised learning with CNN and resampling strategy for VQA. Liu *et al.* [26] exploited the 3D-CNN model for codec classification and quality assessment of compressed videos. In [21], a NR-VQA method named VSFA for in-the-wild videos by incorporating content-dependency and temporal-memory effects was validated. A very recent deep learning-based model called RIRNet was developed in [11], where spatio-temporal features from different temporal frequencies were fused to deliver a comprehensive distortion description. In [7], Chen *et al.* obtained generalized spatio-temporal feature representations for NR-VQA task.

While the boundaries of benchmark performance have been pushed to new limits, these models are designed to be domain-specific and tend to encounter challenge in cross-scenario generalization due to the domain gap. In contrast, our work tries to investigate the suitability of representations learned across different domains, a flexibility we believe is necessary for training quality-aware representations intended for different applications.

3. Proposed Approach

In this section, we present the details of the proposed model, where Figure 2 illustrates the entire framework. The whole model could be disentangled into three parts: **1)** a domain-adaptive quality prediction network to produce prediction for target samples through aligning the feature distributions between source and target domains, which consists of the Backbone network (B), the Distribution Prediction Head (DPH_d, where the subscript d indicates which part it belongs to) and the Domain classifier (D_d); **2)** an uncertainty-based ranking function to rank the target samples based on their prediction confidences and sort them into confident and uncertain subdomains; and **3)** a self-supervised subdomain adaptation network to conduct an adaptation between two subdomains, where the backbone network B is further optimized with the help of DPH_s and D_s. By adversarially optimizing the domain classifiers in both stages with the backbone network, the domain gap of learned features could be alleviated effectively.

3.1. Domain-adaptive Quality Prediction

Formally, we have access to the labeled source video x_s with its associated rating distribution y_s^{dis} drawn from the source domain $\{(\mathbf{X}_s, \mathbf{Y}_s^{dis})\}$, as well as the unlabeled target video x_t drawn from the target domain $\{\mathbf{X}_t\}$. Intuitively, the objective of the domain-adaptive quality prediction loss is summarized as:

$$\mathcal{L}_{pre1} = \mathcal{L}_{DPH_d}(\mathbf{X}_s, \mathbf{Y}_s^{dis}), \tag{1}$$

where $\mathcal{L}_{DPH_d}(\mathbf{X}_s, \mathbf{Y}_s^{dis})$ denotes the classification objective for rating distribution prediction from source data.

The concept behind the domain adaptation is to optimize the shared parameters of the backbone network so that the learned features are discriminative for the primary task of video quality prediction, but are uninformative for the task of domain classification [31]. To close the domain gap between the source and target domains, D_d is trained to predict the domain labels for the input feature, where the gradient reverse layer (GRL) [14] is employed to flip the sign of gradients and jointly train all parameters. The optimization of D_d is achieved via the following adversarial loss function:

$$\mathcal{L}_{adv1} = \frac{1}{2}(\mathcal{L}_{adv}(\mathbf{X}_s) + \mathcal{L}_{adv}(\mathbf{X}_t)), \tag{2}$$

$$\mathcal{L}_{adv}(\mathbf{X}_s) = -\frac{1}{N_s}\sum_{i=1}^{N_s} \log(D_d(B(x_s^i))), \tag{3}$$

$$\mathcal{L}_{adv}(\mathbf{X}_t) = -\frac{1}{N_t}\sum_{j=1}^{N_t} \log(1 - D_d(B(x_t^j))), \tag{4}$$

in which N_s and N_t are the numbers of example from source and target domains, respectively. Therefore, the objective for model training in the first stage is given by:

$$\min_{B,D_1} \max_{D_d} \ \mathcal{L}_{pre1} - \lambda_1 \cdot \mathcal{L}_{adv1}, \tag{5}$$

where D_1 refers to the *DPH*$_d$. λ_1 is a hyperparamter controlling influence degree of the prediction loss and adversarial loss in the first stage.

3.2. Uncertainty-based Ranking

Despite the progress achieved by the domain adaptation techniques, they often encounter prediction uncertainty problem in the context of VQA task. While uncertain predictions with scattered distributions hinder the adaptation performance, one straight way to solve the problem is by enforcing high prediction confidence on those uncertain predictions. To achieve this goal, we decide to adopt a curriculum-style learning scheme following the "easy-to-hard" pattern in target domain. By solving easy tasks first which aim to infer some necessary properties about the target domain, the prediction network could be trained in such a way following those inferred properties to encourage confident predictions on hard tasks [46].

However, it remains intractable to define these easy and hard tasks due to the lack of basis of division for target data. To meet this challenge, we propose an information theory-based distance measurement to determine the confidence levels for target predictions from the first stage. In specific, for each target video x_t, based on the rating distribution prediction $\hat{y}_t^{dis}$ generated from *DPH*$_d$ (for brevity, we omit the superscript in the rest part of this section), we define a simple-yet-effective way for measuring the prediction confidence of the target video, $I(\hat{y}_t)$, as:

$$I(\hat{y}_t) = DUD(\hat{y}_t) + \varepsilon \cdot MED(\hat{y}_t), \tag{6}$$

where $DUD(\cdot)$ and $MED(\cdot)$ are calculated to measure the distance from the uniform distribution and the maximum entropy distribution over the quality scale having the same mean value, respectively. A low value of DUD implies that the predicted distribution is more similar to the uniform distribution with a higher degree of subjectivity (low prediction confidence). The introduction of MED aims to overcome the vulnerability that the DUD measurement tends to penalize more skewed distributions having mean values close to extremes of the quality scale [17], while the parameter ε is used to balance their weights. Detailed calculations of the $DUD(\hat{y}_t)$ and $MED(\hat{y}_t)$ are given as:

$$DUD(\hat{y}_t) = d_w(\hat{y}_t, u_t) = \left[\sum_{i=1}^{N}(\hat{Y}_t(i) - U_t(i))^2\right]^{1/2}, \tag{7}$$

$$MED(\hat{y}_t) = d_w(\hat{y}_t, v_t) = \left[\sum_{i=1}^{N}(\hat{Y}_t(i) - V_t(i))^2\right]^{1/2}, \tag{8}$$

where N denotes the quality rankings, $d_w(\cdot)$ refers to the 2-Wasserstein distance [6]. u_t is the discrete uniform distribution while v_t is the maximum entropy distribution derived using the maximum entropy model [16]. U_t and V_t are their corresponding cumulative distribution functions, respectively.

Given a ranking of measurements from $I(\hat{y}_t)$ (the larger the $I(\hat{y}_t)$, the higher the prediction confidence), the hyperparameter η is introduced as a ratio to split the videos from target domain into two subdomains in terms of their prediction confidences, which we denote as confident and uncertain subdomains. That is, the predicted distributions of videos contained in the confident subdomain tend to be more concentrated and unimodal-like than those in the uncertain subdomain.

3.3. Self-supervised Subdomain Adaptation

Let $\mathbf{X}_c$ and $\mathbf{X}_u$ denote the target samples assigned to the confident and uncertain subdomains. On the basis of two divided subdomains, we refer to the predictions of samples from the confident subdomain as the easy tasks in the curriculum considering their high prediction confidences, which have already been solved in the first stage. The second stage training aims to fine-tune the prediction model through confronting the hard tasks, where the data from the uncertain subdomain are enforced high prediction confidence. This could be enabled by conducting adaptation between the two subdomains.

To this end, we opt for the predictions from the first stage as pseudo-ground truth labels for data in the confident subdomain. With the aid of pseudo labels, *DPH*$_s$ could be optimized by minimizing the prediction loss:

$$\mathcal{L}_{pre2} = \mathcal{L}_{DPH_s}(\mathbf{X}_c, \mathbf{Y}_c^{dis}), \tag{9}$$

where $\mathcal{L}_{DPH_s}(\mathbf{X}_c, \mathbf{Y}_c^{dis})$ serves as the classification objective for rating distribution prediction from the confident subdomain. To encourage confident predictions on uncertain subdomain, we adopt the alignment on the latent feature spaces for both subdomains, which is self-supervised by the subdomain labels derived from the ranking function. The adversarial learning loss to optimize D_s is formulated as:

$$\mathcal{L}_{adv2} = \frac{1}{2}(\mathcal{L}_{adv}(\mathbf{X}_c) + \mathcal{L}_{adv}(\mathbf{X}_u)), \tag{10}$$

$$\mathcal{L}_{adv}(\mathbf{X}_c) = -\frac{1}{N_c}\sum_{p=1}^{N_c}\log(D_s(B(x_c^p))), \tag{11}$$

$$\mathcal{L}_{adv}(\mathbf{X}_u) = -\frac{1}{N_u}\sum_{q=1}^{N_u}\log(1 - D_s(B(x_u^q))), \tag{12}$$

where N_c and N_u are the numbers of example from the confident and uncertain subdomains, respectively. Also, the objective for model training in the second stage could be represented by:

$$\min_{B,D_2}\max_{D_s} \mathcal{L}_{pre2} - \lambda_2 \cdot \mathcal{L}_{adv2}, \tag{13}$$

where D_2 stands for the *DPH*$_s$, and λ_2 acts as the trade-off weighting for the prediction loss and adversarial loss in the second stage.

The final output of *DPH*$_s$ is the predicted rating distribution Q^{dis}, which could be further aggregated into the quality score Q^{sco} the same way the MOS computes from:

$$Q^{sco} = \sum_{m=1}^{N} m \cdot Q^{dis}(m), \tag{14}$$

where N represents total rankings of the rating distribution.

4. Experiments and Analysis

4.1. Experimental Protocols

Database. To evaluate the performance of our method, we leverage LBVD database [10] which contains subjective data with rating distributions as the source domain database, and five other popular VQA databases as the target domain databases. They could be further classified into two categories: LIVE VQA [33] and CSIQ VQA [40] are composed of videos with artificial distortion, while the contents in CVD2014 [29], KoNViD-1k [15] and LIVE-VQC [34] suffer from authentic distortion where no reference video is available.

LBVD database [10] (MOS and rating distribution). This database is a large-scale video quality assessment database for distorted live broadcasting videos, where 1013 samples, each lasts 10s, were collected in the database.

LIVE Video Quality database [33] (only MOS). The database contains 160 videos divided into 10 groups with a resolution of 768×432. Each group contains one reference video and its corresponding 15 distorted videos whose length are 10s.

CSIQ Video Quality database [40] (only MOS) This database contains 12 reference videos and 216 distorted videos generated from 6 distortion types, with a resolution of 832×480.

CVD2014 video database [29] (only MOS). This database aims at complex distortions introduced during video acquisition. It contains 234 videos of resolution 640×480 or 1280×720. The videos are 10-25s with 11-31fps.

KoNViD-1k database [15] (only MOS). This database aims at natural distortions. It comprises a total of 1,200 videos of resolution 960×540 that are fairly filtered from a large public video dataset. The videos are 8s long with 24/25/30fps.

LIVE-VQC database [34] (only MOS). This database contains 585 videos of unique content, captured using 101 different devices (43 device models) by 80 different users with wide ranges of levels of complex, authentic distortions.

Evaluation criteria. We adopt two popular performance criteria, the Pearson Linear Correlation Coefficient (PLCC) and the Spearman Rank-order Correlation Coefficient (SRCC) to measure the accuracy and the monotonicity of the results, respectively. A well-performing quality assessment method is expected to deliver PLCC, SRCC values close to 1. Considering the inconsistency of the scale between objective predictions and the subjective scores, we adopt a four-parameter logistic function for mapping the objective score to the subjective score as outlined in [4].

4.2. Implementation Details

We initialize the backbone network B with a C3D [38] network pre-trained on Kinetics [5]. In the first stage, the architecture of the *DPH*$_d$ model could be denoted by FC(128) - GDN - FC(N) - Softmax using shorthand notations, where FC(n) indicates a fully connected layer with n nodes. GDN is a generalized divisive normalization (GDN) joint nonlinearity layer that is inspired biologically, and has proven effective in quality assessment [27]. In the second stage, the *DPH*$_s$ model shares the same architecture with *DPH*$_d$. We adopt the same architecture in [41] to train both D_d and D_s. With respect to the parameters, we empirically adopt $\varepsilon = 0.5$, $\lambda_1 = \lambda_2 = 0.8$ in all experiments.

We employ PyTorch framework in all experimental implementations. To train *DPH*$_d$, the Adam optimizer [18] with a learning rate of $5e-4$ is deployed to minimize the EMD loss [20] for classification task which benefits from taking into account of the relations between ordered ratings. In the second stage for training *DPH*$_s$, a learning rate of $1e-4$ decayed by a factor of 0.2 every 20 epochs is used. To train D_d and D_s, we apply an SGD optimizer [2] with a learning rate of $1e-4$ and momentum 0.9.

Table 1. Quantitative results of different methods on five publicly available target databases. All the results are trained using the LBVD database [10] as the source domain dataset. Larger PLCC, SRCC values indicate better performance. Best and second best performances of both settings (supervised/unsupervised) are **highlighted** and <u>underlined</u>. Note that VMAF [24] and STRRED [35] are FR/RR metrics that could not be evaluated on those authentic distorted databases. (†) indicates the variants of predicting the rating distribution.

	Database		LIVE VQA [33]		CSIQ VQA [40]		CVD2014 [29]		KoNViD-1k [15]		LIVE-VQC [34]	
	Method	Dis.	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑
Supervised	VMAF [24]		0.7124	0.7220	0.7483	0.7697	–	–	–	–	–	–
	STRRED [35]		<u>0.7985</u>	**0.7967**	<u>0.8155</u>	<u>0.8490</u>	–	–	–	–	–	–
	V-BLIINDS [32]		0.7482	0.7244	0.7710	0.7843	0.7222	0.7068	0.6273	0.6158	0.6592	0.6413
	TLVQM [19]		0.7511	0.7338	0.7740	0.7956	0.8215	0.8352	0.7608	0.7692	0.7514	0.7522
	VIDEVAL [39]		0.7781	0.7636	0.7994	0.8067	<u>0.8445</u>	<u>0.8580</u>	**0.7865**	**0.7804**	<u>0.7961</u>	**0.7816**
	VSFA [21]		0.7278	0.7001	0.7816	0.7980	0.8277	0.8431	0.7391	0.7452	0.7807	0.7645
	RIRNet [11]		**0.8091**	<u>0.7828</u>	**0.8426**	**0.8574**	**0.8780**	**0.8891**	<u>0.7812</u>	<u>0.7755</u>	**0.7982**	<u>0.7713</u>
Unsupervised	*NoAdapt*		0.5873	0.5615	0.6051	0.6172	0.7002	0.6597	0.6547	0.6371	0.6335	0.6219
	VIIDEO [28]		0.6518	0.6240	0.5447	0.4906	0.2083	0.1544	0.3058	0.3412	0.1146	0.0734
	TCoN [30]		0.6868	0.6727	0.7231	0.7245	0.7527	0.7479	0.7336	0.7305	0.6741	0.6918
	TCoN† [30]	✓	0.6923	0.6749	0.7260	0.7318	0.7581	0.7405	<u>0.7380</u>	<u>0.7421</u>	0.6805	0.6947
	TA^{3}N [9]		0.6917	0.6993	0.7397	0.7305	0.7707	0.7412	0.7193	0.7030	0.6932	0.6962
	TA^{3}N† [9]	✓	<u>0.7005</u>	<u>0.7016</u>	<u>0.7422</u>	<u>0.7434</u>	<u>0.7743</u>	0.7560	0.7177	0.7104	0.7019	0.7071
	UDA		0.6749	0.6837	0.7005	0.6989	0.7582	0.7604	0.6981	0.7085	0.7082	0.7007
	UDA†	✓	0.6833	0.6870	0.7042	0.7069	0.7693	<u>0.7760</u>	0.7146	0.7278	<u>0.7114</u>	<u>0.7052</u>
	UCDA(Ours)	✓	**0.7797**	**0.7835**	**0.8283**	**0.8167**	**0.8414**	**0.8475**	**0.7909**	**0.7851**	**0.7702**	**0.7622**

4.3. Performance Evaluation

We evaluate the proposed UCDA in two variants (*NoAdapt* indicates the model is trained on the source domain and directly test on the target domain without adaptation; *UDA* has exactly the same architecture as our proposed method in the first stage except that the predicted target is the scalar score, and without further curriculum-style adaptation), and by comparing with several competitors including: **1)** eight VQA methods, among which VIIDEO [28] is the only unsupervised type that is training-free, and the other seven metrics are the supervised type (VMAF [24], STRRED [35], V-BLLINDS [32], TLVQM [19], VIDEVAL [39], VSFA [21] and RIRNet [11]) which are directly trained and test on the target dataset; **2)** two general video domain adaptation methods (TCoN [30] and TA^{3}N [9]) as the baseline models to demonstrate the effectiveness of the designed curriculum-style adaptation considering no existing domain adaptation method are specialized in VQA task (we evaluate their performances with the same settings as our approach).

Experiments on each database are processed by k-fold (k = 10) cross-validation, ensuring the training sets and testing sets are not overlapped in content. This procedure is repeated 10 times and the average values of PLCC and S-RCC results across all repetitions for the mentioned competitors and the proposed algorithm are given in Table 1. We also evaluate whether predicting the rating distribution contributes to the prediction performance by reporting the performances achieved by invariants of *UDA*, TCoN and TA^{3}N, where the regression quality score prediction is replaced by the rating distribution prediction.

From the experimental results, we have several observations. First, the proposed metric convincingly outperforms all other unsupervised metrics with respect to prediction accuracy (PLCC) and monotonicity (SRCC) on all target datasets, providing clear quantitative evidence of the effectiveness of the proposed UCDA. Specifically, it outperforms the variant *UDA* by a large margin (0.1048, 0.1278, 0.0836, 0.0928 and 0.0620 in terms of PLCC on five target datasets, respectively), which highlights the benefit of the designed curriculum-style adaptation on enforcing high prediction confidence on those uncertain predictions. Second, although unsupervised VQA method VIIDEO is designed for arbitrary distortion types, it does not perform well on realistic distortions in CVD2014, KoNViD-1k and LIVE-VQC datasets. Third, in general, better performances are attained by these variants that try to predict the rating distributions. This finding further strengthes our observation that predicting the rating distributions instead of the scalar scores leads to a higher correlation to the intrinsic nature of quality assessment, which can benefit the overall quality prediction. Last but not least, an inspiring discovery is that the performance of the proposed UCDA surpasses most of the comparison supervised methods without any supervision from the target dataset, and even achieves the best performance in KoNViD-1k dataset which has the largest size in all test datasets. We stress that there are very few VQA algorithms in the literature that work well on unsupervised setting, and our UCDA is very competitive in that sense.

4.4. Ablation Study

Impact of adopting different datasets as the source domain. To check the effectiveness of UCDA beyond leveraging the LBVD as the source domain dataset, we go a further step by evaluating the proposed method with respect to choosing source domain from other datasets containing only MOS values. Since our method requires training on labels of quality rating distributions, we follow [37] to approximate them from the available MOS values through maximum entropy optimization. In particular, the rating distribution is calculated as the maximum entropy distribution corresponding to the MOS as the mean value. We then report the prediction performances of the proposed UCDA compared with *UDA* in Table 2. According to the results, it is desirable that the selection of the source domain has an impact on the performance of the adaptive prediction model. Compared to *UDA*, our method consistently achieves much better performances. What is more important, it could maintain well-performing regardless on all transfer tasks. This manifests to some extent that the introduction of the curriculum-style adaptation in our framework facilitates the generalization ability of the learned model.

Table 2. Prediction performances of the proposed method and *UDA* with respect to different source domain datasets measured by PLCC. The column-wise and the row-wise datasets are selected as the source and target domain datasets, respectively.

	Source\Target	LIVE	CSIQ	CVD	KoNViD	VQC
UCDA	LIVE	N/A	0.8291	0.7747	0.7280	0.7314
	CSIQ	0.7850	N/A	0.7531	0.7397	0.6821
	CVD	0.7607	0.7917	N/A	0.7655	0.7447
	KoNViD	0.7582	0.8126	0.8462	N/A	0.7598
	VQC	0.7614	0.8103	0.8215	0.7817	N/A
	Average	0.7663	0.8109	0.7938	0.7537	0.7295
UDA	LIVE	N/A	0.7419	0.6244	0.5990	0.5918
	CSIQ	0.7273	N/A	0.5987	0.5746	0.5676
	CVD	0.6115	0.6842	N/A	0.6676	0.6715
	KoNViD	0.5987	0.6990	0.7505	N/A	0.7132
	VQC	0.6336	0.7063	0.7323	0.7247	N/A
	Average	0.6428	0.7079	0.6765	0.6415	0.6360

Effect of Hyperparameter η. We conduct an ablation study on finding a proper value for the hyperparameter η to split the samples of the target domain into two subdomains. Different values of η are selected for setting up the decision boundary for the separation on the validation set. Figure 3 exhibits the experimental results on all five target domain databases. When the number of η is small, an obvious performance gap could be observed when we incrementally increase η (compared to the configuration where η is 0), indicating that effectiveness of the designed scheme. However, once, the number of η comes up to a certain value (0.6 in this work), increasing the number does not improve the performance further or even tends to witness a declining. These findings suggest that the scale of the partition does affect the prediction performance, and thus 0.6 is chosen as the proportion to split the target domain in our experiments.

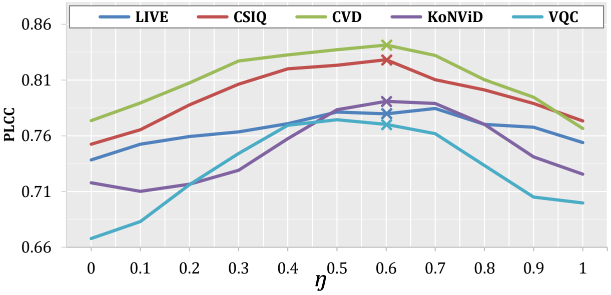

Figure 3. Impact of hyperparameter η on prediction performances, which are measured by PLCC.

4.5. Extension to Semi-supervised Case

To further study the robustness of the proposed algorithm, we extend the proposed approach to a semi-supervised setting, where a part of target labels are available for taking part in the training of DPH_d (labels from target datasets are processed the same way described in Section 4.4). Extensive experiments are conducted on adapting from the LBVD to KoNViD-1k and LIVE-VQC datasets. Results with varying ratios of target labels, ranging from 0.1 to 0.5, are reported in Table 3, where other domain adaptation methods are included for comparison. There we show, by adding the available number of target labels, we consistently observe that the proposed UCDA achieves the best prediction performance on both transfer tasks regardless of the amount of annotated training data for the target domain. At the same time, by comparison with other metrics, the performance of our model is positively correlated with the number of training samples available from the target domain. This shows the benefit of the designed curriculum-style adaptation on taking full advantages of the supervision information.

Table 3. Experimental results measured by PLCC under the semi-supervised setting on two transfer tasks.

	LBVD→KoNViD			LBVD→VQC		
Method	10%	30%	50%	10%	30%	50%
NoAdapt	0.6547	0.6547	0.6547	0.6335	0.6335	0.6335
TCoN [30]	0.7442	0.7517	0.7595	0.6920	0.6985	0.7056
TA^{3}N [9]	0.7257	0.7316	0.7385	0.7091	0.7206	0.7237
UDA	0.7063	0.7134	0.7205	0.7160	0.7251	0.7314
UCDA(Ours)	**0.8029**	**0.8202**	**0.8386**	**0.7893**	**0.8085**	**0.8237**

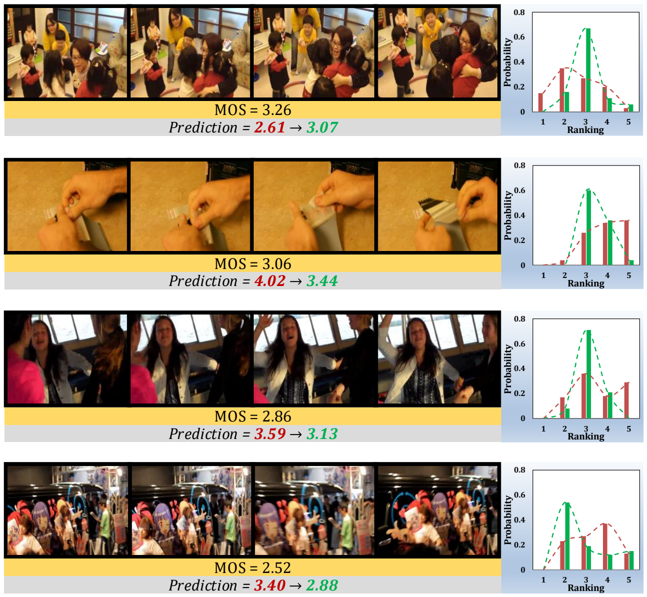

Figure 4. Qualitative comparison by examples from KoNViD-1k [15] database. For each example video, right of the sampled frames are the predicted rating distributions before and after the curriculum-style adaptation, corresponding quality scores calculated from the distributions combined with their MOS labels are provided at the bottom. Note that the predictions with **red** text indicate the results of the DPH_d model, while **green** indicating the predictions of the DPH_s model.

4.6. Qualitative Evaluation of UCDA

We take the KoNViD-1k database to visualize the performance boost brought by UCDA. A representative set of videos belonging to the uncertain subdomain, along with their predicted rating distributions and quality scores, are visible in Figure 4. To further investigate how our curriculum mechanism works, we also include the prediction results made by DPH_d before the self-supervised subdomain adaptation as the baseline model for comparison.

It can be observed that the prediction distributions tend to be more unimodal-like, and the calculated quality scores are closer to the MOS labels than the baseline model, proving that the curriculum-style adaptation could encourage high prediction confidence and further improve the performances. It is interesting to see that the scene and content of the video clips from the uncertain subdomain are generally complex where more than one object are contained. An interpretation is that multiple objects are more likely to distract the attention of the subjects, affecting their judgements to some extent. Besides, the MOS labels corresponding to these samples are mainly concentrated near the mid-quality range, which is consistent with the observation in [45] assuming that humans tend to give more consistent ratings (smaller variances) to videos at the two ends of the quality range than those in the mid-quality range.

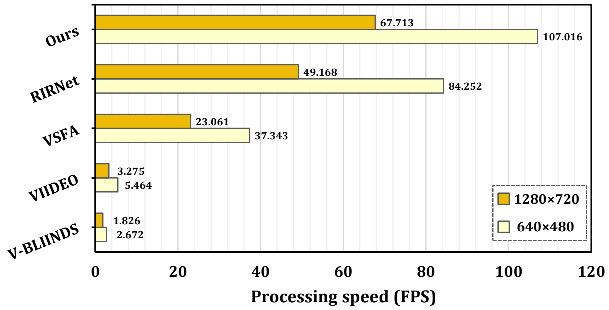

Figure 5. Average processing speed in frames-per-second (FPS) of different NR-VQA models on CVD2014 video database.

4.7. Computational Efficiency

Besides the performance, computational efficiency is also crucial for NR-VQA methods. We compare the average processing speed of V-BLIINDS, VIIDEO, VSFA, RIRNet, and the proposed UCDA on CVD2014 dataset, where all the video samples share the spatial resolution of 640×480 and 1280×720. To provide a fair comparison for the computational efficiency of different methods, all tests are carried out on a computer with a E5-2630 CPU and 64 GB RAM. The default settings of the original codes are used without any modification. We repeat the tests ten times and the FPS for each method is shown in Figure 5. It is worth noting that all deep learning-based models are much faster than the conventional ones, where V-BLIINDS and VIIDEO can only process less than 6 frames per second. Our method could achieve real-time processing speed (over 67 fps) on 720p videos, which is very helpful for practical applications.

5. Conclusion

This paper presented a novel path for unsupervised domain adaptation approach in cross-domain NR-VQA task based on two ideas, *i.e.*, rating distribution prediction and curriculum-style adaptation. The former, compared to simply predicting the scalar quality score, is more informative and with higher correlation to the subjective nature of quality assessment. By measuring the prediction confidence of the predicted rating distributions, the latter helps to develop a two-stage adaptation to improve the adaptation performance by enforcing high prediction confidence on those uncertain predictions in target domain. Results from extensive experiments consistently validated the the effectiveness and efficiency of the proposed method.

Acknowledgements. This work was supported by National Natural Science Foundation of China (62171340, 61771473, 61991451), the Natural Science Foundation of Jiangsu Province (BK20181354, BK20200649), the Key Project of Shaanxi Provincial Department of Education (Collaborative Innovation Center, 20JY024), and the Fundamental Research Funds for the Central Universities (JBF211902).

References

[1] C. G. Bampis, Z. Li, and A. C. Bovik. Spatiotemporal Feature Integration and Model Fusion for Full Reference Video Quality Assessment. *IEEE Trans. Circuits Syst. Video Technol.*, 29(8):2256–2270, 2019.

[2] L. Bottou. Large-Scale Machine Learning with Stochastic Gradient Descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[3] T. Brandão and M. P. Queluz. No-Reference Quality Assessment of H. 264/AVC Encoded Video. *IEEE Trans. Circuits Syst. Video Technol.*, 20(11):1437–1447, 2010.

[4] ITU-R BT.500. Methodology for the Subjective Assessment of the Quality of Television Pictures. *International Telecommunication Union*, 2002.

[5] J. Carreira and A. Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 4724–4733. IEEE, 2017.

[6] J. A. Carrillo, R. J. McCann, and C. Villani. Contractions in the 2-Wasserstein Length Space and Thermalization of Granular Media. *Archive for Rational Mechanics and Analysis*, 179(2):217–263, 2006.

[7] B. Chen, L. Zhu, G. Li, F. Lu, H. Fan, and S. Wang. Learning Generalized Spatial-Temporal Deep Feature Representation for No-Reference Video Quality Assessment. *IEEE Trans. Circuits Syst. Video Technol.*, pages 1–1, 2021.

[8] C. Chen, Y. Lin, S. Benting, and A. Kokaram. Optimized Transcoding for Large Scale Adaptive Streaming Using Playback Statistics. In *Proc. IEEE Int. Conf. Image Process. (ICIP)*, pages 3269–3273. IEEE, 2018.

[9] M. Chen, Z. Kira, G. AlRegib, J. Yoo, R. Chen, and J. Zheng. Temporal Attentive Alignment for Large-Scale Video Domain Adaptation. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 6321–6330, 2019.

[10] P. Chen, L. Li, Y. Huang, F. Tan, and W. Chen. QoE Evaluation for Live Broadcasting Video. In *Proc. IEEE Int. Conf. Image Process. (ICIP)*, pages 454–458. IEEE, 2019.

[11] P. Chen, L. Li, L. Ma, J. Wu, and G. Shi. RIRNet: Recurrent-In-Recurrent Network for Video Quality Assessment. In *Proc. ACM Int. Conf. Multimedia (ACM MM)*, pages 834–842. ACM, 2020.

[12] P. Chen, L. Li, J. Wu, Y. Zhang, and W. Lin. Temporal Reasoning Guided QoE Evaluation for Mobile Live Video Broadcasting. *IEEE Trans. Image Process.*, 30:3279–3292, 2021.

[13] Y. Fang, H. Zhu, Y. Zeng, K. Ma, and Z. Wang. Perceptual Quality Assessment of Smartphone Photography. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 3674–3683. IEEE, 2020.

[14] Y. Ganin and V. Lempitsky. Unsupervised Domain Adaptation by Backpropagation. In *Proc. Int. Conf. Mach. Learn. (ICML)*, pages 1180–1189, 2015.

[15] V. Hosu, F. Hahn, M. Jenadeleh, H. Lin, H. Men, T. Szirnyi, S. Li, and D. Saupe. The Konstanz Natural Video Database (KoNViD-1k). In *Proc. Int. Conf. Quality Multimedia Exper. (QoMEx)*, pages 1–6. IEEE, 2017.

[16] E. T. Jaynes. On the Rationale of Maximum-Entropy Methods. *Proceedings of the IEEE*, 70(9):939–952, 1982.

[17] C. Kang, G. Valenzise, and F. Dufaux. Predicting Subjectivity in Image Aesthetics Assessment. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2019.

[18] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[19] J. Korhonen. Two-Level Approach for No-Reference Consumer Video Quality Assessment. *IEEE Trans. Image Process.*, 28(12):5923–5938, 2019.

[20] E. Levina and P. Bickel. The Earth Mover's Distance is the Mallows Distance: Some Insights from Statistics. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, volume 2, pages 251–256. IEEE, 2001.

[21] D. Li, T. Jiang, and M. Jiang. Quality Assessment of In-the-Wild Videos. In *Proc. ACM Int. Conf. Multimedia (ACM MM)*, pages 2351–2359. ACM, 2019.

[22] L. Li, W. Lin, X. Wang, G. Yang, K. Bahrami, and A. C. Kot. No-Reference Image Blur Assessment Based on Discrete Orthogonal Moments. *IEEE Trans. Cybern.*, 46(1):39–50, 2016.

[23] L. Li, W. Xia, W. Lin, Y. Fang, and S. Wang. No-Reference and Robust Image Sharpness Evaluation Based on Multiscale Spatial and Spectral Features. *IEEE Trans. Multimedia*, 19(5):1030–1040, 2017.

[24] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara. Toward A Practical Perceptual Video Quality Metric. *The Netflix Tech Blog*, 6, 2016.

[25] Z. Li and C. G. Bampis. Recover Subjective Quality Scores from Noisy Measurements. In *2017 Data Compression Conference (DCC)*, pages 52–61. IEEE, 2017.

[26] W. Liu, Z. Duanmu, and Z. Wang. End-to-End Blind Quality Assessment of Compressed Videos Using Deep Neural Networks. In *Proc. ACM Int. Conf. Multimedia (ACM MM)*, pages 546–554. ACM, 2018.

[27] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo. End-to-End Blind Image Quality Assessment Using Deep Neural Networks. *IEEE Trans. Image Process.*, 27(3):1202–1213, 2018.

[28] A. Mittal, M. A. Saad, and A. C. Bovik. A Completely Blind Video Integrity Oracle. *IEEE Trans. Image Process.*, 25(1):289–300, 2016.

[29] M. Nuutinen, T. Virtanen, M. Vaahteranoksa, T. Vuori, P. Oittinen, and J. Häkkinen. CVD2014-A Database for Evaluating No-Reference Video Quality Assessment Algorithms. *IEEE Trans. Image Process.*, 25(7):3073–3086, 2016.

[30] B. Pan, Z. Cao, E. Adeli, and J. C. Niebles. Adversarial Cross-Domain Action Recognition with Co-Attention. In *Proc. AAAI Conf. Artif. Intell. (AAAI)*, volume 34, pages 11815–11822, 2020.

[31] F. Pan, I. Shin, F. Rameau, S. Lee, and I. S. Kweon. Unsupervised Intra-Domain Adaptation for Semantic Segmentation Through Self-Supervision. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 3763–3772. IEEE, 2020.

[32] M. A. Saad, A. C. Bovik, and C. Charrier. Blind Prediction of Natural Video Quality. *IEEE Trans. Image Process.*, 23(3):1352–1365, 2014.

[33] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack. Study of Subjective and Objective Quality Assessment of Video. *IEEE Trans. Image Process.*, 19(6):1427–1441, 2010.

[34] Z. Sinno and A. C. Bovik. Large-Scale Study of Perceptual Video Quality. *IEEE Trans. Image Process.*, 28(2):612–627, 2019.

[35] R. Soundararajan and A. C. Bovik. Video Quality Assessment by Reduced Reference Spatio-Temporal Entropic Differencing. *IEEE Trans. Circuits Syst. Video Technol.*, 23(4):684–694, 2013.

[36] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang. Blindly Assess Image Quality in the Wild Guided by a Self-Adaptive Hyper Network. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 3667–3676. IEEE, 2020.

[37] H. Talebi and P. Milanfar. NIMA: Neural image assessment. *IEEE Trans. Image Process.*, 27(8):3998–4011, 2018.

[38] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning Spatiotemporal Features with 3D Convolutional Networks. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 4489–4497. IEEE, 2015.

[39] 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.*, 30:4449–4464, 2021.

[40] P. V. Vu and D. M. Chandler. ViS3: An Algorithm for Video Quality Assessment via Analysis of Spatial and Spatiotemporal Slices. *Journal of Electronic Imaging*, 23(1):013016, 2014.

[41] T. Vu, H. Jain, M. Bucher, M. Cord, and P. Prez. ADVENT: Adversarial Entropy Minimization for Domain Adaptation in Semantic Segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 2512–2521. IEEE, 2019.

[42] J. Wu, Y. Liu, W. Dong, G. Shi, and W. Lin. Quality Assessment for Video with Degradation Along Salient Trajectories. *IEEE Trans. Multimedia*, 21(11):2738–2749, 2019.

[43] Q. Wu, H. Li, F. Meng, and K. N. Ngan. Toward a Blind Quality Metric for Temporally Distorted Streaming Video. *IEEE Trans. Broadcast.*, 64(2):367–378, 2018.

[44] H. Zeng, L. Zhang, and A. C. Bovik. Blind Image Quality Assessment with a Probabilistic Quality Representation. In *Proc. IEEE Int. Conf. Image Process. (ICIP)*, pages 609–613. IEEE, 2018.

[45] W. Zhang, K. Ma, G. Zhai, and X. Yang. Uncertainty-Aware Blind Image Quality Assessment in the Laboratory and Wild. *IEEE Trans. Image Process.*, 30:3474–3486, 2021.

[46] Y. Zhang, P. David, H. Foroosh, and B. Gong. A Curriculum Domain Adaptation Approach to the Semantic Segmentation of Urban Scenes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(8):1823–1841, 2020.

[47] 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.*, 29(8):2224–2255, 2019.

[48] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi. MetaIQA: Deep Meta-learning for No-Reference Image Quality Assessment. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 14143–14152. IEEE, 2020.