Knowledge Guided Semi-supervised Learning for Quality Assessment of User Generated Videos

Shankhanil Mitra, Rajiv Soundararajan

Visual Information Processing Lab, Indian Institute of Science, Bengaluru
{shankhanilm, rajivs}@iisc.ac.in

Abstract

Perceptual quality assessment of user generated content (UGC) videos is challenging due to the requirement of large scale human annotated videos for training. In this work, we address this challenge by first designing a self-supervised Spatio-Temporal Visual Quality Representation Learning (ST-VQRL) framework to generate robust quality aware features for videos. Then, we propose a dual-model based Semi Supervised Learning (SSL) method specifically designed for the Video Quality Assessment (SSL-VQA) task, through a novel knowledge transfer of quality predictions between the two models. Our SSL-VQA method uses the ST-VQRL backbone to produce robust performances across various VQA datasets including cross-database settings, despite being learned with limited human annotated videos. Our model improves the state-of-the-art performance when trained only with limited data by around 10%, and by around 15% when unlabelled data is also used in SSL. Source codes and checkpoints are available at https://github.com/Shankhanil006/SSL-VQA.

Introduction

The emergence of video capturing devices such as smartphones, DSLRs, and GoPro has led to millions of users uploading or accessing videos via various sharing platforms such as YouTube, Instagram, Facebook and so on. This necessitates the quality assessment (QA) of videos to monitor and control the user experience. However, a reference video is often not available for user generated content (UGC), motivating the study of no reference (NR) video QA (VQA). Further, the videos also suffer from complex camera captured distortions which makes the task of NR VQA extremely challenging.

The recent decade has seen significant progress in NR VQA, based on classical or handcrafted features (Saad, Bovik, and Charrier 2014; Xu et al. 2014; Ghadiyaram and Bovik 2017; Tu et al. 2021a,b) and deep learning based approaches (Li, Jiang, and Jiang 2019, 2021; Wu et al. 2022; Chen et al. 2020a; Shen et al. 2022). The deep learning based approaches particularly require training on large amount of labelled data, which is cumbersome and expensive to acquire. This leads to us to the question of how we can design NR VQA models which can be trained with very limited labelled training data, yet achieve excellent generalisation performance on multiple datasets in terms of correlation with human perception.

Our focus in this work is on designing semi-supervised NR VQA method with limited labelled along with unlabelled data. Since UGC videos have diverse quality characteristics, we believe that pretraining a robust video quality feature backbone is extremely important to transfer knowledge during semi-supervised learning. With this motivation, we approach the problem using a combination of contrastive self-supervised pretraining followed by semi-supervised finetuning. A few self-supervised contrastive learning based methods have been designed for NR VQA recently (Mitra and Soundararajan 2022; Madhusudana et al. 2022; Chen et al. 2022) to learn rich video quality features. However, none of these methods yet exploit the performance benefit offered by the attention mechanism in transformer based models. One of the major challenges in training such transformer based architectures for VQA is the difficulty in training such networks end-to-end. We leverage recent literature on end-to-end training of Swin-transformers for supervised VQA (Wu et al. 2022) to overcome this difficulty in self-supervised video quality representation learning. Further, we employ a novel statistical contrastive learning loss instead of a point-wise similarity loss to make the learning more robust. Thus, in the first stage of our approach, we learn rich quality aware spatio-temporal features without requiring any human annotations.

In the second stage of our approach, we leverage the limited number of quality labels in a semi-supervised learning (SSL) framework. While several SSL based methods on pseudo-labelling and consistency regularisation have been explored in video action recognition (Xu et al. 2022; Singh et al. 2021; Kumar and Rawat 2022), they need to adapted for the specific task of VQA. In this direction, we employ knowledge transfer between two measures of video quality evaluated on the unlabelled videos. The first measure is based only on human annotations, while the second measure uses a distance between features of the distorted video and a corpus of pristine videos along with human labels. Such knowledge transfer helps overcome the drawback of limited human annotations, while simultaneously trying to help determine a perceptually relevant distance to a corpus of pristine videos. We show that the above semi-supervised learning helps design an effective VQA method with limited labels.

We conduct several experiments on multiple cross and intra datasets to validate the performance of our proposed framework. To summarize, our main contributions consist of:

1. Self-supervised statistical contrastive learning of spatio-temporal video quality representations with a transformer based architecture.
2. Semi-supervised learning of video quality by knowledge transfer between models based on limited human labels and feature distances to a corpus of pristine videos.
3. Impressive cross-database performance despite the model being trained with very few human annotated videos.

Related Work

Classical Feature based VQA. Historically, handcrafted heuristics based features have been shown to produce robust performance across various VQA datasets. Among them, VBLIINDS (Saad, Bovik, and Charrier 2014) and VCORNIA (Xu et al. 2014) learn natural scene statistics of video frames by modelling the discrete cosine transform (DCT) or 3D-DCT. In recent years, TLVQM (Korhonen 2019) has shown considerable improvement in VQA performance by modelling temporal low complexity features with spatial high complexity features. VIDEVAL (Tu et al. 2021a) is an ensemble of various handcrafted features designed to capture diverse quality attributes in a video. Nevertheless, authentically distorted videos in UGC have mixed distortions, which are very hard to model using the above statistical methods.

Supervised pretraining based VQA. Existing deep learning methods mostly regress fixed quality aware features against human opinion scores due to the computational complexity of training large models. VSFA (Li, Jiang, and Jiang 2019) and MDTVSFA (Li, Jiang, and Jiang 2021) learn a gated recurrent unit on top of features generated by ResNet50 (He et al. 2016). PVQ (Ying et al. 2021) extracts 2D and 3D pretrained features from image quality assessment (IQA) and action recognition tasks. Recently, FAST-VQA (Wu et al. 2022) learns an end-to-end model by spatially fragmenting the video clips thus reducing the complexity. TCSVT-BVQA (Li et al. 2022) on the other hand learns a VQA model by transferring spatial knowledge from pretrained IQA, and temporal knowledge from a pretrained action recognition model.

Unsupervised pretraining based VQA. VISION (Mitra and Soundararajan 2022), and CONVIQT (Madhusudana et al. 2022) present self-supervised learning based quality aware feature extractors. The fixed features from these self-supervised models can be further regressed against opinion scores to develop an end-to-end quality model. In our work, we first train a self-supervised quality feature extractor and use it to build an end-to-end SSL framework.

Unsupervised VQA. VQA methods such as STEM (Kancharla and Channappayya 2022), VISION (Mitra and Soundararajan 2022), and NVQE (Liao et al. 2022) do not require any human labelled videos in their design and give reasonable quality estimates for UGC videos. Nevertheless, their performance with respect to the methods trained with human opinion scores is under par.

Semi-Supervised Learning. To the best of our knowledge, there exist no end-to-end SSL algorithms designed for the VQA task. SSL methods for classification can be broadly classified into pseudo-labelling, consistency regularisation, and hybrid methods. While pseudo-labelling as such is unsuitable for regression, consistency regularisation and its hybrid versions such as Mean Teacher (Tarvainen and Valpola 2017), FixMatch (Sohn et al. 2020), MixMatch (Berthelot et al. 2019), and Meta Pseudo-Label (Pham et al. 2021) are better suited for regression tasks. In the case of QA, these algorithms can not be directly applied as augmentations for image/video classification are quality variant.

We remark that SSL has not been explored much even in the IQA literature. While Conde *et al.*(Conde, Burchi, and Timofte 2022) study SSL methods for full reference IQA, some other methods (Wang, Li, and Ma 2021; Yue et al. 2022) train an NR IQA model with a large number of labelled images and generate pseudo-labels on the unlabelled data.

Spatio-Temporal VQ Representation Learning

Overview. First, we learn a self-supervised spatio-temporal backbone to capture Video Quality (VQ) aware features from unlabelled videos. The VQ representation based feature extractor is used as a backbone in our semi-supervised model to get robust performance despite learning on limited data.

We embark on solving multiple key challenges in learning a 3D self-supervised representation learning for VQA. It is computationally hard and even infeasible to train a video transformer using contrastive learning with videos of high resolution. Inspired by recent works on VQA (Wu et al. 2022, 2023), we propose a quality invariant sampling strategy that preserves the global context and local quality of videos to overcome the computational challenges while training such models. In addition, a 3D vision transformer such as video swin transformer (Liu et al. 2022) captures both short and long duration temporal distortions such as shakiness, motion blur, and flicker in videos on account of its design. Finally, in traditional contrastive learning (Chen et al. 2020b; Tian, Krishnan, and Isola 2020; Tao, Wang, and Yamasaki 2020), a point-wise similarity between the global representation of features is optimized, which ignores the local variations in video space-time. To address this problem, we propose a statistical contrastive loss where both global and local information are shared between a contrasting pair of video features.

Quality Consistent Sampling and Video Augmentation. To capture quality-aware representations, we choose contrastive pairs of video clips from synthetically distorted UGC videos having similar content but different levels and types of distortions. We synthetically distort UGC videos similar to VISION (Mitra and Soundararajan 2022) to model mixed camera captured and synthetic distortions from which

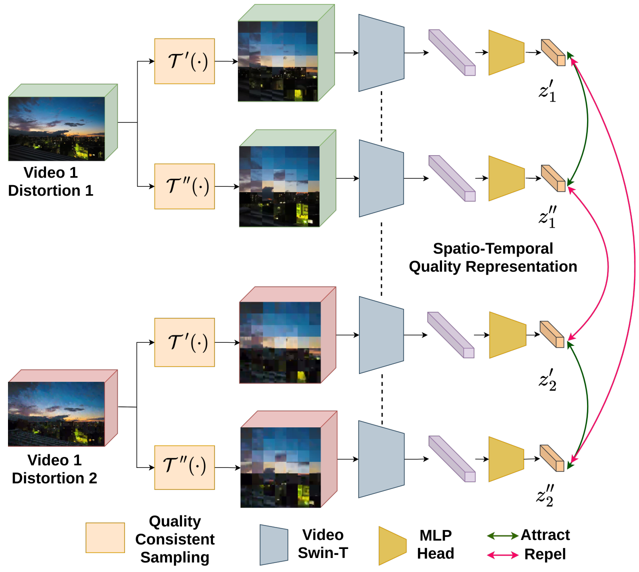

Figure 1: Framework of Spatio-Temporal VQ Representation Learning (ST-VQRL). For two distorted versions of the same video, Video 1, we sample an augmented pair of clips from each distorted version using QCS. The quality aware contrastive loss is used to train the network to attract augmented clips from same video and repel clips from other distorted versions in the embedding space.

quality aware features can be learned. We then employ spatial fragment sampling (crop multiple patches at original resolution and splice them together) of continuous video frames to capture both local and global video distortions in frames (Wu et al. 2022). As shown in literature (Wu et al. 2022, 2023), random fragments sampled from a video clip, share similar quality representations. We refer to the process of obtaining fragments from video clips as quality consistent sampling (**QCS**).

Let $\mathcal{T}'(\cdot)$ and $\mathcal{T}''(\cdot)$ denote the generators of two instances of random QCS. We apply $\mathcal{T}'(\cdot)$ and $\mathcal{T}''(\cdot)$ on the same clip as shown in Figure 1 thus generating augmented clips having similar local and global quality. Augmented versions of a clip constitute a positive pair. Fragments sampled from different distorted versions of the same video clip using $\mathcal{T}'(\cdot)$ and $\mathcal{T}''(\cdot)$ constitute a negative pair.

Statistical Contrastive Loss for Representation Learning. We capture spatio-temporal representations from a video clip using a Video Swin-T (Liu et al. 2022) backbone given as $f_\theta(\cdot)$ with model parameters θ. Consider a set of K video clips $\{V_1, V_2, \dots, V_K\}$ of the same scene content with different distortions. Let, $\mathcal{T}'(V_i)$ and $\mathcal{T}''(V_i)$ denote the pair of augmented clips of every video V_i. Let $z_i' = f_\theta(\mathcal{T}'(V_i))$, and $z_i'' = f_\theta(\mathcal{T}''(V_i))$ be the feature representations of the augmented pair of clips of V_i, where z is of dimension $N \times C$. We propose a statistical constrastive loss between the spatio-temporal feature representation of the pairs to capture both the global description and local variations in the representations of fragments.

Our contrastive loss minimises a distance between the augmented pair of representations z_i', and z_i'' and maximises the distance between z_i', and z_k', where $k \neq j$ and $k \in \{1, 2, \dots, K\}$. In particular, we are inspired by the work in NIQE (Mittal, Soundararajan, and Bovik 2013), where it is shown that a statistical distance between image features is relevant to perceptual quality. We treat the $N \times C$ feature vector as a set of N spatio-temporal samples of dimension C drawn from a multivariate Gaussian (MVG) model. Let the MVG model parameters be (μ', Σ'), and (μ'', Σ'') for feature representations z', and z''. Thus, we obtain a quality aware distance between any z', and z'' as

$$d(z', z'') = \sqrt{(\mu' - \mu'')^T \left(\frac{\Sigma' + \Sigma''}{2}\right)^{-1} (\mu' - \mu'')}. \quad (1)$$

Therefore, the quality aware contrastive loss with $\mathcal{T}'(V_i), i \in \{1, 2, \dots, K\}$ taken as an anchor view is $\mathcal{L}' = \frac{1}{K}\sum_{i=1}^{K} l_i'$, where

$$l_i' = -\log \frac{\exp(-d(z_i', z_i'')/\tau)}{\sum_{j=1}^{K} \exp(-d(z_i', z_j'')/\tau)}. \quad (2)$$

Similarly, taking $\mathcal{T}''(V_i), i \in \{1, 2, \dots, K\}$ as anchor, we obtain a loss $\mathcal{L}''$, and the overall loss is given as,

$$\mathcal{L}_c = \mathcal{L}' + \mathcal{L}''. \quad (3)$$

Knowledge Transfer based SSL-VQA

Given a set of labelled UGC videos $V = \{(v_1, y_1), \cdots (v_{N_l}, y_{N_l})\}$ (annotated with human opinion scores), and a set of unlabelled videos $U = \{u_1, \cdots u_{N_u}\}$, our proposed approach learns quality assessment of UGC videos by utilizing both sets. As shown in Figure 2, we design a dual-model learning setup, where one model directly maps the video features to a scalar video quality score while the other model maps the distance between the representations of a distorted video and corpus of pristine videos to video quality. The two models differ in the use of a corpus of pristine videos to predict quality. While the use of distance to a corpus imposes more structure in the quality prediction, it may also limit the quality modelling capability. Our goal is to transfer quality aware knowledge learned by the individual models to each other. While the backbone in both the models is initialised using pretrained ST-VQRL (with parameter θ), we update their parameter separately as θ', and θ'' respectively during finetuning.

Regressor based Quality Model. We attach a regressor head on top of the spatio-temporal feature encoder viz. ST-VQRL and train this model end-to-end as shown in Figure 2. Let $g_\phi(\cdot)$ be a non-linear regressor head with parameter ϕ applied on top of self-supervised ST-VQRL feature extractor $f_{\theta'}(\cdot)$ to predict a scalar quality estimate of videos. $g_\phi(\cdot)$ comprises of two 3d convolutional layers with filter size $1 \times 1 \times 1$ to preserve the local quality characteristics of the generated features of $f_{\theta'}(\cdot)$. Therefore, the predicted quality for any video $x \in (V_v \cup U)$, where $V_v = \{v_i\}_{i=1}^{N_l}$ is given as,

$$Q_R(x) = g_\phi(f_{\theta'}(x)).$$

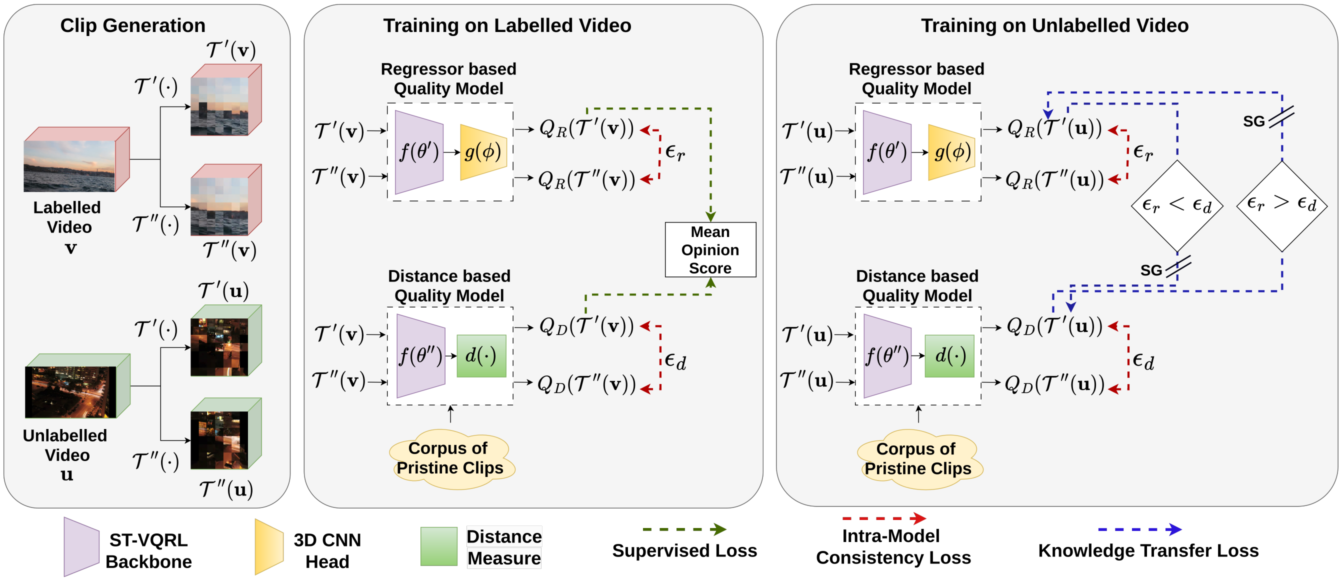
Clip Generation
$\mathcal{T}'(\mathbf{v})$
$\mathcal{T}'(\cdot)$
Labelled Video $\mathbf{v}$
$\mathcal{T}''(\cdot)$
$\mathcal{T}''(\mathbf{v})$
$\mathcal{T}'(\mathbf{u})$
$\mathcal{T}'(\cdot)$
Unlabelled Video $\mathbf{u}$
$\mathcal{T}''(\cdot)$
$\mathcal{T}''(\mathbf{u})$
Training on Labelled Video
Regressor based Quality Model
$\mathcal{T}'(\mathbf{v})$
$\mathcal{T}''(\mathbf{v})$
$f(\theta')$
$g(\phi)$
$Q_R(\mathcal{T}'(\mathbf{v}))$
$Q_R(\mathcal{T}''(\mathbf{v}))$
ϵ_r
Mean Opinion Score
Distance based Quality Model
$\mathcal{T}'(\mathbf{v})$
$\mathcal{T}''(\mathbf{v})$
$f(\theta'')$
$d(\cdot)$
$Q_D(\mathcal{T}'(\mathbf{v}))$
$Q_D(\mathcal{T}''(\mathbf{v}))$
ϵ_d
Corpus of Pristine Clips
Training on Unlabelled Video
Regressor based Quality Model
$\mathcal{T}'(\mathbf{u})$
$\mathcal{T}''(\mathbf{u})$
$f(\theta')$
$g(\phi)$
$Q_R(\mathcal{T}'(\mathbf{u}))$
$Q_R(\mathcal{T}''(\mathbf{u}))$
ϵ_r
SG
$\epsilon_r < \epsilon_d$
$\epsilon_r > \epsilon_d$
SG
Distance based Quality Model
$\mathcal{T}'(\mathbf{u})$
$\mathcal{T}''(\mathbf{u})$
$f(\theta'')$
$d(\cdot)$
$Q_D(\mathcal{T}'(\mathbf{u}))$
$Q_D(\mathcal{T}''(\mathbf{u}))$
ϵ_d
Corpus of Pristine Clips
ST-VQRL Backbone
3D CNN Head
Distance Measure
Supervised Loss
Intra-Model Consistency Loss
Knowledge Transfer Loss

Figure 2: Overview of Semi-Supervised Learning for VQA (SSL-VQA) method. Every batch consists of labelled ($v \in V_v$) and unlabelled ($u \in U$) samples. First, we generate an augmented pair of clips from v and u using two instances of quality consistent sampling (QCS), $\mathcal{T}'(\cdot)$ and $\mathcal{T}''(\cdot)$. For labelled video v, SSL-VQA optimises the supervised and intra-model consistency loss. For unlabelled video u, SSL-VQA enforces intra-model consistency loss and knowledge transfer loss. Based on the consistency criteria ($\epsilon_r > \epsilon_d$), SSL-VQA transfers knowledge from one model to another.

Distance based Quality Model. The second model $f_{\theta''}(\cdot)$ described in Figure 2 is necessarily a feature encoder like ST-VQRL. $f_{\theta''}(\cdot)$ measures the distance between the feature representation of a corpus of pristine videos and any distorted video $x \in (V_v \cup U)$. Let, $z^x = f_{\theta''}(x)$ denote the feature representation of x. Similarly, for N_p pristine or clean videos, we get feature embedding z^r using $f_{\theta''}(\cdot)$. Thereafter, we fit a multivariate Gaussian (MVG) model on the corpus of pristine feature set z^r to get (μ^r, Σ^r). Let the model parameters for the distorted video representation z^x be (μ^x, Σ^x). The quality estimate of video x is given as $Q_D(x) = \exp(-d(z^r, z^x)/\tau)$, where

$$d(z^r, z^x) = \sqrt{(\mu^r - \mu^x)^T \left(\frac{\Sigma^r + \Sigma^x}{2}\right)^{-1} (\mu^r - \mu^x)}.$$

We predict the overall quality of a video during inference stage as $(Q_R(y) + Q_D(y))/2$, where y is any test video.

Supervised Learning Loss

We train both the models with the labelled videos using the ground truth opinion scores. Both models are trained separately on a mini-batch of size B_l by minimising the batch-wise Pearson's linear correlation coefficient (PLCC) between predicted quality and ground truth opinion scores. Note that this loss is differentiable and allows for back-propagation. Let $\mathbf{v} = \{v_i\}_{i=1}^{B_l}$ and $\mathbf{y} = \{y_i\}_{i=1}^{B_l}$, where $\{(v_i, y_i)\}_{i=1}^{B_l} \in V$. Thus, the supervised loss is formulated as

$$\mathcal{L}_s = \mathcal{L}_{plcc}(Q_R(\mathcal{T}'(\mathbf{v})), \mathbf{y}) + \mathcal{L}_{plcc}(Q_D(\mathcal{T}'(\mathbf{v})), \mathbf{y}), \quad (4)$$

where $\mathcal{L}_{plcc}(a, b) = \frac{1 - PLCC(a,b)}{2}$ and $\mathcal{T}'(\mathbf{v}) = \{\mathcal{T}'(v_i)\}_{i=1}^{B_l}$. Since learning end-to-end on the original video resolution is computationally intensive, we apply **QCS**.

Intra-Model Consistency Loss

The goal of the intra-model consistency loss is to enable consistent quality predictions for the augmented video clips obtained through QCS. Let the two quality consistent augmented clips for a mini-batch of videos $\mathbf{x} = \{x_i\}_{i=1}^{B}$, where $\{x_i\}_{i=1}^{B} \in (V_v \cup U)$ be $\mathcal{T}'(\mathbf{x})$, and $\mathcal{T}''(\mathbf{x})$. Note that $B = B_l + B_u$, where B_u is the mini-batch length of unlabelled videos. The intra-model consistency loss is given as

$$\begin{aligned}\mathcal{L}_c =& \mathcal{L}_{plcc}(Q_R(\mathcal{T}'(\mathbf{x})), Q_R(\mathcal{T}''(\mathbf{x})))) \\ &+ \mathcal{L}_{plcc}(Q_D(\mathcal{T}'(\mathbf{x})), Q_D(\mathcal{T}''(\mathbf{x})))). \end{aligned} \quad (5)$$

Knowledge Transfer based Loss

The goal of the knowledge transfer loss is to utilise both the models effectively for SSL. If a consistency between the prediction of both the models on unlabelled data is enforced, an erroneous prediction of one model may drive the other model to wrong knowledge. So the challenge here is to transfer knowledge from one model to another only when its prediction is reliable. We hypothesize that we can rely on the prediction of the model if it is stable with respect to augmented versions of a sample. To determine model stability with respect to a batch, we evaluate the intra-model consistency for a mini-batch of unlabelled samples $\mathbf{u} = \{u_i\}_{i=1}^{B_u}$,

Setting		Intra Test Database				Cross Database			
Test Dataset		$\text{LSVQ}_{\text{test}}$		$\text{LSVQ}_{\text{1080p}}$		KoNVid-1K		LIVE VQC	
Method	Model Type	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
VBLIIND	Classical Features	0.473	0.456	0.382	0.430	0.545	0.539	0.398	0.434
TLVQM		0.599	0.582	0.441	0.473	0.592	0.597	0.531	0.551
VIDEVAL		0.607	0.597	0.491	0.547	0.545	0.543	0.416	0.459
VSFA	Supervised Pretraining	0.663	0.645	0.536	0.543	0.664	0.669	0.651	0.668
FAST-VQA		0.682	0.677	0.552	0.558	0.679	0.666	0.652	0.672
TCSVT-BVQA		0.687	0.679	0.449	0.465	0.682	0.680	0.665	0.682
VISION	Self-Supervised Pretraining	0.523	0.478	0.427	0.446	0.606	0.612	0.615	0.636
CONVIQT		0.636	0.624	0.468	0.464	0.662	0.673	0.600	0.627
SSL-VQA^{-} w/o consistency loss		0.704	0.693	0.562	0.577	0.715	0.713	0.669	0.678
SSL-VQA^{-}		**0.719**	**0.717**	**0.587**	**0.601**	**0.736**	**0.735**	**0.683**	**0.688**

Table 1: Performance analysis of SSL-VQA^{-} with ST-VQRL backbone compared against other popular VQA methods with classical features, supervised pretrained, and self-supervised pretrained backbones when trained only with limited human annotated videos without any unlabelled data.

where each $u_i \in U$. The model whose consistency loss between the augmented pair of videos in Equation (5) is less, is considered more stable. Let ϵ_r denote the regressor based quality model's consistency error and ϵ_d denote the distance based quality model's consistency error. Then,

$$\begin{aligned}\epsilon_r &= \mathcal{L}_{plcc}(Q_R(\mathcal{T}'(\mathbf{u})), Q_R(\mathcal{T}''(\mathbf{u})))\\ \epsilon_d &= \mathcal{L}_{plcc}(Q_D(\mathcal{T}'(\mathbf{u})), Q_D(\mathcal{T}''(\mathbf{u}))).\end{aligned}$$

The prediction of the more stable model is used as a pseudo-label for the other model. The knowledge transferable loss function for a mini-batch is given as

$$\begin{aligned}\mathcal{L}_u &= m\mathcal{L}_{plcc}(Q_R(\mathcal{T}'(\mathbf{u})), sg(Q_D(\mathcal{T}'(\mathbf{u}))))\\ &\quad + (1-m)\mathcal{L}_{plcc}(sg(Q_R(\mathcal{T}'(\mathbf{u}))), Q_D(\mathcal{T}'(\mathbf{u}))),\end{aligned} \tag{6}$$

where $m = \mathbb{1}(\epsilon_r > \epsilon_d)$ is the indicator *mask* and $sg(.)$ denotes the stop gradient operation. This ensures that the stable model provides a pseudo-label or guidance for the other model.

The overall loss to train our SSL-VQA model end-to-end is a combination of the supervised loss, intra-model consistency loss, and knowledge transferable loss as

$$\mathcal{L} = \mathcal{L}_s + \lambda_c\mathcal{L}_c + \lambda_u\mathcal{L}_u, \tag{7}$$

where λ_c, λ_u are hyper parameters to balance the loss terms.

Experiments

In this section, we describe the implementation details, experimental setup, and comparisons with other methods.

Implementation Details of ST-VQRL

Data Generation. We learn our self-supervised ST-VQRL model on a set of synthetically distorted UGC videos. We randomly sample 200 videos out of 28056 training videos of LIVE-FB Large-Scale Social Video Quality (LSVQ) (Ying et al. 2021) database. LSVQ database videos have unique scenes with camera captured distortions, so we augment each of the 200 videos with 12 different synthetic distortion types and levels in the same manner as in VISION (Mitra and Soundararajan 2022). In the statistical contrastive loss and distance based quality model, we use a set of 60 pristine videos from LIVE-VQA (Seshadrinathan et al. 2010), LIVE Mobile (Moorthy et al. 2012), CSIQ VQD (Vu and Chandler 2014), EPFL-PoLiMI (De Simone et al. 2010) and ECCV-EVVQ databases (Rimac-Drıje, Vranje, and Žagar 2010).

Training Details. We encode the distorted video sequence using a Video Swin-T (Liu et al. 2022) architecture modified with gated relative positional bias (Wu et al. 2022) to take into account the discontinuity in a sampled clip due to QCS. QCS is applied by dividing each of the 32 continuous video frames into a 7×7 grid, sampling 32×32 patches from each grid and stitching them together maintaining temporal consistency. The patches within each grid are extracted from the same location for every distorted version of a scene, thus the distorted clips are content consistent with respect to sampling. We train ST-VQRL using AdamW (Loshchilov and Hutter 2019) with a learning rate of 10^{-4} and a weight decay of 0.05 for 30 epochs. The temperature co-efficient τ mentioned in Equation (2) is 10.

Experimental Setup

We conduct two types of experiments to evaluate VQA under limited labelled data. In the first experiment (**Experimental Setting 1**) in Table 1, we define SSL-VQA^{-} as our model learnt with only limited labelled samples and without using any unlabelled data. Thus, we only use the losses in Equations (4) and (5) to understand how the framework can learn with just the limited labelled data available for training. In the second experiment (**Experimental Setting 2**) in Table 2, we perform SSL by also utilizing the unlabelled data. In both the cases we train the model for 30 epochs using AdamW (Loshchilov and Hutter 2019) with a learning rate of 10^{-4} and a weight decay of 0.05. λ_c, and λ_u are chosen to be 1 based on training loss convergence.

Setting	Intra Test Database				Cross Database			
Test Database	$\text{LSVQ}_{\text{test}}$		$\text{LSVQ}_{\text{1080p}}$		KoNVid-1K		LIVE VQC	
Method	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
Mean Teacher	0.716	0.703	0.594	0.603	0.716	0.715	0.679	0.681
Meta PseudoLabel	0.714	0.713	0.586	0.587	0.719	0.716	0.676	0.673
FixMatch	0.722	0.725	0.582	0.596	0.727	0.732	0.685	0.687
SSL-VQA	**0.731**	**0.736**	**0.616**	**0.645**	**0.765**	**0.770**	**0.711**	**0.734**

Table 2: Performance comparison of SSL-VQA with other SSL benchmarks on intra and inter database test settings. All methods are initialised with ST-VQRL backbone for fair comparison and are trained with both labelled and unlabelled samples.

Training Database. LSVQ (Ying et al. 2021) has an official training set of 28056 videos. We extract 2000 videos from it randomly. Out of the 2000 videos, we only use around 500 videos or 1.78% of the training set with human opinion scores and the remaining 1500 unlabelled videos.

Evaluation Database. We employ two different test settings. In the first setting, we test on the official test database $\text{LSVQ}_{\text{test}}$, containing 7400 videos of varying resolution between 240p and 720p, and $\text{LSVQ}_{\text{1080p}}$ containing around 3600 videos of 1080p for intra database performance evaluation. We further test the robustness of SSL-VQA in cross database settings. Particularly, we test on KoNVid-1K (Hosu et al. 2017), and LIVE VQC (Sinno and Bovik 2019), each comprising of 1200 and 585 camera captured authentically distorted videos with varying resolutions. We use the Spearman Rank-Order Correlation Coefficient (SROCC), and Pearson Linear Correlation Coefficient (PLCC) as performance measures. In both Table 1 and 2, we report the median performance over 3 random choices of 500 labelled and 1500 unlabelled videos out of the 2000 chosen videos.

Experimental Setting 1 Analysis

We compare SSL-VQA^{-} with three popular categories of NR VQA models under Experimental Setting 1. Among the models based on classical or heuristic based feature designs, we compare with Video BLIINDS (Saad, Bovik, and Charrier 2014), TLVQM (Korhonen 2019), and VIDEVAL (Tu et al. 2021a). Among recent deep VQA models, we compare with the state-of-the-art FAST-VQA model (Wu et al. 2022) and the popular VSFA (Li, Jiang, and Jiang 2019) and TCSVT-BVQA (Li et al. 2022) methods. Finally, we also compare with recent self-supervised feature learning models such as VISION (Mitra and Soundararajan 2022) and CONVIQT (Madhusudana et al. 2022). The comparison with the supervised pre-trained FAST-VQA model is particularly interesting since the experiment reveals how our self-supervised ST-VQRL backbone of SSL-VQA^{-} enables learning with limited labelled data.

In Table 1, we see that our method outperforms both classical feature based and recent deep learning methods. This increment in performance can be attributed to the use of a robust quality aware feature backbone viz. ST-VQRL. The ST-VQRL encoder not only provides robust performance in the limited data regime but also is independent of any supervision like FAST-VQA, VSFA, and TCSVT-BVQA.

Experimental Setting 2 Analysis

Since there exist no direct end-to-end semi-supervised VQA models in the literature for comparison, we adapt popular semi-supervised methods for the VQA task. Semi-supervised approaches can be broadly divided into pseudo-labelling and consistency regularisation. Since direct pseudo-labelling is not applicable for regression tasks, we rely upon consistency regularisation based methods such as MeanTeacher (Tarvainen and Valpola 2017) and FixMatch (Sohn et al. 2020). We also modify the meta learning based SSL method, Meta Pseudo-Label (Pham et al. 2021) for comparison. We use our self-supervised feature encoder ST-VQRL as the backbone for fair comparison with SSL-VQA approaches.

In Table 2, we provide a quantitative comparison between SSL-VQA and other SSL methods modified for VQA. In both intra-database and cross database settings, we see a considerable improvement of SSL-VQA over other methods. Thus a smart knowledge transfer between the two quality models enriches our SSL framework leading to superior performance.

Ablation Studies

Finetuning on UGC Data: In earlier section, we show SSL-VQA's robustness across intra and inter database test settings. Now we evaluate SSL-VQA by finetuning it for specific VQA tasks. In general, we adapt our model on four smaller VQA datasets viz. KoNVid-1k (Hosu et al. 2017), LIVE VQC (Sinno and Bovik 2019), YouTube-UGC (Wang, Inguva, and Adsumilli 2019), and LIVE Qualcomm (Ghadiyaram et al. 2018). KoNVid-1K and LIVE VQC predominantly have videos captured in the wild with cameras. YouTube-UGC comprises of videos from various domains such as real-world, animation, and gaming, spanning a spatial resolution from 240p to 4K. LIVE Qualcomm on other hand, comprises of authentic videos with specific categories of distortions such as shakiness, stabilisation, and so on. In this experiment, we randomly sample **20%** of the videos with labels in each of the UGC databases above for finetuning and use another non-overlapping 20% for testing. We report the median performance across 10 such splits in Table 3. We observe that such fine-tuning with limited labelled data substantially improves the performance.

Impact of Statistical Contrastive Loss: In VQ representation learning, we noted that the use of a statistical measure between spatio-temporal video features in Equation (1)

Method	KoNVid-1K		LIVE VQC		LIVE QCOMM		YouTube-UGC	
	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
Mean Teacher	0.783	0.786	0.702	0.693	0.722	0.724	0.725	0.730
Meta PseudoLabel	0.792	0.795	0.705	0.710	0.734	0.735	0.729	0.719
FixMatch	0.798	0.799	0.716	0.729	0.740	0.730	0.734	0.730
SSL-VQA	**0.826**	**0.828**	**0.733**	**0.743**	**0.747**	**0.751**	**0.750**	**0.757**

Table 3: Performance on KoNViD, LIVE-VQC, LIVE Qualcomm and YouTube-UGC when SSL-VQA is finetuned on 20% of annotated videos from each database. We provide comparison with other SSL benchmarks which are also finetuned on 20% of labelled videos for each of these databases.

Backbone	KoNVid-1K	LIVE VQC
Pre-trained Video Swin-T	0.728	0.686
ST-VQRL w/ Similarity loss	0.733	0.684
ST-VQRL w/ Statistical loss	**0.765**	**0.711**

Table 4: SROCC performance analysis of different backbones on SSL-VQA performance.

Approach	KoNVid-1K	LIVE VQC
SSL-VQA w/o consistency loss	0.756	0.697
SSL-VQA w/o knowledge loss	0.742	0.686
SSL-VQA	**0.765**	**0.711**

Table 5: SROCC performance analysis of different unsupervised constraints in Equation (5) and (6) respectively.

is more relevant to perceptual quality. To validate our hypothesis, we show in Table 4, the superiority of our model optimised with Equation (3) over the cosine similarity loss as in generic contrastive learning. We also provide the performance when using a supervised Video Swin-T backbone (Liu et al. 2022) pretrained for action recognition over our self-supervised backbone. We infer that ST-VQRL learned from scratch, specifically to capture quality aware features gives better performance than supervised pretrained Video Swin-T. Moreover, ST-VQRL learned using a statistical distance measure captures quality representations better.

Role of Intra-model Consistency and Knowledge Transfer: SSL-VQA model is optimised using an objective function comprising of supervised loss, intra-model consistency loss, and knowledge transfer loss. When we evaluate the model without the consistency loss, note that there is also no *mask* in Equation (6) and unrestricted knowledge transfer happens between the two models. Without the knowledge transfer loss, the unlabelled data is only used to impose intra-model consistency. In Table 5, we see that the absence of either loss hinders the learning performance showing the benefit of our contributions in SSL for VQA.

Impact of Number of Labelled and Unlabelled Videos: As mentioned in experimental setting analysis, we train SSL-VQA on 500 labelled and 1500 unlabelled videos from the LSVQ (Ying et al. 2021) database. Here, we present an analysis of our model's performance with varying numbers of labelled and unlabelled videos. In Figure 3a, and 3b, we present the performance of SSL-VQA trained on with labelled videos varying between 100-500 keeping the number of unlabelled videos fixed. We see a steady increase in performance as the number of labelled videos increases. Similarly, in Figure 3c, and 3d, we train SSL-VQA using 500 labelled and 500-2500 unlabelled videos. We see that when the number of unlabelled videos is considerably high (more than 1500), the model's prediction nearly saturates.

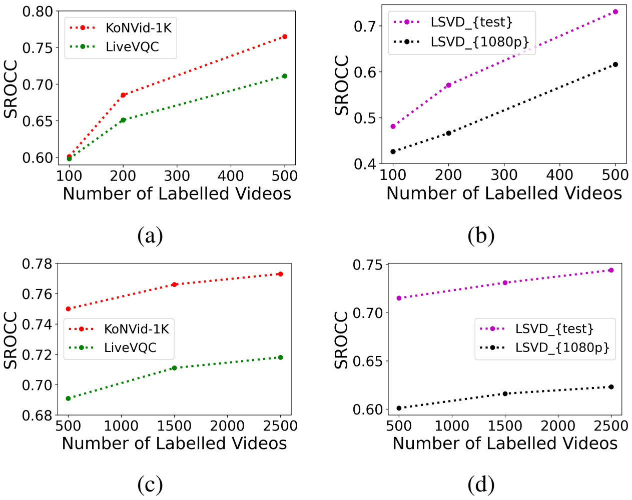

Figure 3: (a), and (b) correspond to SSL-VQA performance on KoNVid-1K, LIVE-VQC, $\text{LSVQ}_{\text{test}}$, and $\text{LSVQ}_{1080\text{p}}$ datasets when the number of labelled videos for training ranges from 100-500. In (c), and (d), we show the performance for 500 labelled videos with different amounts of unlabelled videos.

Conclusion

We presented a novel SSL approach with a robust quality aware feature encoder ST-VQRL. Through extensive experiments, we showed that SSL-VQA achieves higher performance than existing state-of-the art VQA methods even when learned with very few human annotated videos. We also benchmarked SSL methods for VQA and showed the superiority of our framework on the use of unlabelled

	VBLIIND	TLVQM	VIDEVAL	VSFA	FAST-VQA	TCSVT-BVQA	VISION	CONVIQT	SSL-VQA^{-}
VBLIIND	-	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
TLVQM	1 1 1 1	-	0 0 1 1	0 0 0 0	0 0 0 0	0 1 0 0	1 1 0 0	0 0 0 0	0 0 0 0
VIDEVAL	1 1 1 1	1 1 0 0	-	0 0 0 0	0 0 0 0	0 1 0 0	1 1 0 0	0 1 0 0	0 0 0 0
VSFA	1 1 1 1	1 1 1 1	1 1 1 1	-	0 0 0 1	0 1 0 0	1 1 1 1	1 1 0 1	0 0 0 0
FAST-VQA	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 0	-	0 1 0 0	1 1 1 1	1 1 1 1	0 0 0 0
TCSVT-BVQA	1 1 1 1	1 0 1 1	1 0 1 1	1 0 1 1	1 0 1 1	-	1 1 1 1	1 0 1 1	0 0 0 0
VISION	1 1 1 1	0 0 1 1	0 0 1 1	0 0 0 0	0 0 0 0	0 0 0 0	-	0 0 0 1	0 0 0 0
CONVIQT	1 1 1 1	1 1 1 1	1 0 1 1	0 0 1 0	0 0 0 0	0 1 0 0	1 1 1 0	-	0 0 0 0
SSL-VQA^{-}	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	-

Table 6: Results of one-sided Wilcoxon Rank Sum Test performed between the SROCC values of the other VQA algorithms and SSL-VQA^{-}. Each entry in the table consists of a codeword with 4 symbols corresponding to the testing on LSVQ$_{\text{test}}$, LSVQ$_{\text{1080p}}$, KoNVid-1K, and LIVE VQC databases in that order. A code value of "1" indicates that the VQA model in the row is statistically superior to the VQA model in the column. While a value of "0" indicates row model is inferior to the column model and " – " indicates a statistically similar performance.

Method	LSVQ Test	LSVQ 1080p	KoNVid	LIVE VQC
VSFA	0.801	0.675	0.784	0.734
PatchVQ	0.827	0.711	0.791	0.770
CSVT-BVQA	0.852	0.771	0.834	0.816
FAST-VQA	0.876	0.779	0.859	0.823
SSL-VQA^{-}	0.891	0.799	0.877	0.839

Table 7: SROCC performance of various VQA methods

videos. As our model works on video fragments similar to FAST-VQA (Wu et al. 2022), it is also computationally efficient. We believe that SSL-VQA can make deep learning based VQA perform robustly with limited labels.

A Evaluation on Full LSVQ Train Data

We provide a quantitative analysis between VQA methods and SSL-VQA^{-} with ST-VQRL feature backbone trained on the full annotated train set of LSVQ in Table 7. We infer that our ST-VQRL representation achieves superior performance compared to various benchmark methods even at full scale supervision.

B Statistical Significance Test

In Tables 1 and 2, we reported the median performance of SSL-VQA^{-} and SSL-VQA against various VQA and SSL methods in limited labelled data or in semi-supervised settings respectively over 3 random training splits. We conduct a statistical significance test to validate the superiority of our method. In particular, the non-parametric Wilcoxon Rank Sum Test is used to compare the rank of two sets of correlation coefficients for a pair of methods across 3 splits. Similar to (Yu et al. 2019), we consider the null hypothesis as that the the median of one algorithm is equal to that of the other at 95 % significance level. The alternate hypothesis is that the medians differ. From Tables 6 and 8, we observe that our SSL-VQA^{-} and SSL-VQA outperform other methods with regard to f-test in both the experimental settings.

	Mean Teacher	Meta Pseudo Label	FixMatch	SSL-VQA
Mean Teacher	-	0 1 0 1	0 1 0 0	0 0 0 0
Meta Pseudo-Label	1 0 1 0	-	0 1 0 0	0 0 0 0
FixMatch	1 0 1 1	1 0 1 1	-	0 0 0 0
SSL-VQA	1 1 1 1	1 1 1 1	1 1 1 1	-

Table 8: Results of one-sided Wilcoxon Rank Sum Test performed between the SROCC values of the other semisupervised algorithms and SSL-VQA. The code word has similar representation as in Table 6.

C Training Details

SSL-VQA and all other benchmarking methods were trained in Python 3.8 using Pytorch 2.0 on a 3 × 24 GB NVIDIA RTX 3090 GPU. We consider the optimizer hyperparameters to train SSL-VQA is similar to that of described in FAST-VQA as both uses a Video Swin-T backbone.

Acknowledgments

This work was supported in part by Department of Science and Technology, Government of India, under grant CRG/2020/003516.

References

Berthelot, D.; Carlini, N.; Goodfellow, I.; Oliver, A.; Papernot, N.; and Raffel, C. 2019. MixMatch: A Holistic Approach to Semi-Supervised Learning. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc.

Chen, P.; Li, L.; Ma, L.; Wu, J.; and Shi, G. 2020a. RIRNet: Recurrent-In-Recurrent Network for Video Quality Assessment. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, 834–842. New York, NY, USA: Association for Computing Machinery. ISBN 9781450379885.

Chen, P.; Li, L.; Wu, J.; Dong, W.; and Shi, G. 2022. Contrastive Self-Supervised Pre-Training for Video Quality Assessment. *IEEE Transactions on Image Processing*, 31: 458–471.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020b. A Simple Framework for Contrastive Learning of Visual Representations. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 1597–1607. PMLR.

Conde, M. V.; Burchi, M.; and Timofte, R. 2022. Conformer and blind noisy students for improved image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 940–950.

De Simone, F.; Tagliasacchi, M.; Naccari, M.; Tubaro, S.; and Ebrahimi, T. 2010. A H.264/AVC video database for the evaluation of quality metrics. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2430–2433.

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

Ghadiyaram, D.; Pan, J.; Bovik, A. C.; Moorthy, A. K.; Panda, P.; and Yang, K.-C. 2018. In-Capture Mobile Video Distortions: A Study of Subjective Behavior and Objective Algorithms. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(9): 2061–2077.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Hosu, V.; Hahn, F.; Jenadeleh, M.; Lin, H.; Men, H.; Szirányi, T.; Li, S.; and Saupe, D. 2017. The Konstanz natural video database (KoNViD-1k). In *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, 1–6. IEEE.

Kancharla, P.; and Channappayya, S. S. 2022. Completely Blind Quality Assessment of User Generated Video Content. *IEEE Transactions on Image Processing*, 31: 263–274.

Korhonen, J. 2019. Two-Level Approach for No-Reference Consumer Video Quality Assessment. *IEEE Transactions on Image Processing*, 28(12): 5923–5938.

Kumar, A.; and Rawat, Y. S. 2022. End-to-End Semi-Supervised Learning for Video Action Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14700–14710.

Li, B.; Zhang, W.; Tian, M.; Zhai, G.; and Wang, X. 2022. Blindly Assess Quality of In-the-Wild Videos via Quality-aware Pre-training and Motion Perception. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9): 5944–5958.

Li, D.; Jiang, T.; and Jiang, M. 2019. Quality Assessment of In-the-Wild Videos. MM '19, 2351–2359. New York, NY, USA: Association for Computing Machinery. ISBN 9781450368896.

Li, D.; Jiang, T.; and Jiang, M. 2021. Unified Quality Assessment of in-the-Wild Videos with Mixed Datasets Training. *International Journal of Computer Vision*, 129(4): 1238–1257.

Liao, L.; Xu, K.; Wu, H.; Chen, C.; Sun, W.; Yan, Q.; and Lin, W. 2022. Exploring the Effectiveness of Video Perceptual Representation in Blind Video Quality Assessment. MM '22, 837–846. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392037.

Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; and Hu, H. 2022. Video Swin Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3202–3211.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Madhusudana, P. C.; Birkbeck, N.; Wang, Y.; Adsumilli, B.; and Bovik, A. C. 2022. CONVIQT: Contrastive Video Quality Estimator.

Mitra, S.; and Soundararajan, R. 2022. Multiview Contrastive Learning for Completely Blind Video Quality Assessment of User Generated Content. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, 1914–1924. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392037.

Mittal, A.; Soundararajan, R.; and Bovik, A. C. 2013. Making a "Completely Blind" Image Quality Analyzer. *IEEE Signal Processing Letters*, 20(3): 209–212.

Moorthy, A. K.; Choi, L. K.; Bovik, A. C.; and de Veciana, G. 2012. Video Quality Assessment on Mobile Devices: Subjective, Behavioral and Objective Studies. *IEEE Journal of Selected Topics in Signal Processing*, 6(6): 652–671.

Pham, H.; Dai, Z.; Xie, Q.; and Le, Q. V. 2021. Meta Pseudo Labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11557–11568.

Rimac-Drıje, S.; Vranje, M.; and Žagar, D. 2010. Foveated Mean Squared Error–a Novel Video Quality Metric. *Multimedia Tools Appl.*, 49(3): 425–445.

Saad, M. A.; Bovik, A. C.; and Charrier, C. 2014. Blind Prediction of Natural Video Quality. *IEEE Transactions on Image Processing*, 23(3): 1352–1365.

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

Shen, W.; Zhou, M.; Liao, X.; Jia, W.; Xiang, T.; Fang, B.; and Shang, Z. 2022. An End-to-End No-Reference Video Quality Assessment Method With Hierarchical Spatiotemporal Feature Representation. *IEEE Transactions on Broadcasting*, 68(3): 651–660.

Singh, A.; Chakraborty, O.; Varshney, A.; Panda, R.; Feris, R.; Saenko, K.; and Das, A. 2021. Semi-Supervised Action Recognition With Temporal Contrastive Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10389–10399.

Sinno, Z.; and Bovik, A. C. 2019. Large-Scale Study of Perceptual Video Quality. *IEEE Transactions on Image Processing*, 28(2): 612–627.

Sohn, K.; Berthelot, D.; Li, C.-L.; Zhang, Z.; Carlini, N.; Cubuk, E. D.; Kurakin, A.; Zhang, H.; and Raffel, C. 2020. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, 596–608. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713829546.

Tao, L.; Wang, X.; and Yamasaki, T. 2020. *Self-Supervised Video Representation Learning Using Inter-Intra Contrastive Framework*, 2193–2201. New York, NY, USA: Association for Computing Machinery. ISBN 9781450379885.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 1195–1204.

Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive Multiview Coding. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J.-M., eds., *Computer Vision – ECCV 2020*, 776–794. Cham: Springer International Publishing. ISBN 978-3-030-58621-8.

Tu, Z.; Wang, Y.; Birkbeck, N.; Adsumilli, B.; and Bovik, A. C. 2021a. UGC-VQA: Benchmarking Blind Video Quality Assessment for User Generated Content. *IEEE Transactions on Image Processing*, 30: 4449–4464.

Tu, Z.; Yu, X.; Wang, Y.; Birkbeck, N.; Adsumilli, B.; and Bovik, A. C. 2021b. RAPIQUE: Rapid and Accurate Video Quality Prediction of User Generated Content. *CoRR*, abs/2101.10955.

Vu, P. V.; and Chandler, D. M. 2014. ViS3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices. *Journal of Electronic Imaging*, 23(1): 1 – 25.

Wang, Y.; Inguva, S.; and Adsumilli, B. 2019. YouTube UGC Dataset for Video Compression Research. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, 1–5.

Wang, Z.; Li, D.; and Ma, K. 2021. Semi-supervised deep ensembles for blind image quality assessment.

Wu, H.; Chen, C.; Hou, J.; Liao, L.; Wang, A.; Sun, W.; Yan, Q.; and Lin, W. 2022. FAST-VQA: Efficient End-to-End Video Quality Assessment with Fragment Sampling. In Avidan, S.; Brostow, G.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision – ECCV 2022*, 538–554. Cham: Springer Nature Switzerland. ISBN 978-3-031-20068-7.

Wu, H.; Zhang, E.; Liao, L.; Chen, C.; Hou, J.; Wang, A.; Sun, W.; Yan, Q.; and Lin, W. 2023. Towards Explainable In-the-Wild Video Quality Assessment: a Database and a Language-Prompted Approach. *arXiv preprint arXiv:2305.12726*.

Xu, J.; Ye, P.; Liu, Y.; and Doermann, D. 2014. No-reference video quality assessment via feature learning. In *2014 IEEE International Conference on Image Processing (ICIP)*, 491–495.

Xu, Y.; Wei, F.; Sun, X.; Yang, C.; Shen, Y.; Dai, B.; Zhou, B.; and Lin, S. 2022. Cross-Model Pseudo-Labeling for Semi-Supervised Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2959–2968.

Ying, Z.; Mandal, M.; Ghadiyaram, D.; and Bovik, A. 2021. Patch-VQ: 'Patching Up' the Video Quality Problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14019–14029.

Yu, X.; Bampis, C. G.; Gupta, P.; and Bovik, A. C. 2019. Predicting the Quality of Images Compressed After Distortion in Two Steps. *IEEE Transactions on Image Processing*, 28(12): 5757–5770.

Yue, G.; Cheng, D.; Li, L.; Zhou, T.; Liu, H.; and Wang, T. 2022. Semi-Supervised Authentically Distorted Image Quality Assessment with Consistency-Preserving Dual-Branch Convolutional Neural Network. *IEEE Transactions on Multimedia*, 1–13.