In-the-wild Pretrained Models Are Good Feature Extractors for Video Quality Assessment

Anonymous authors
Paper under double-blind review

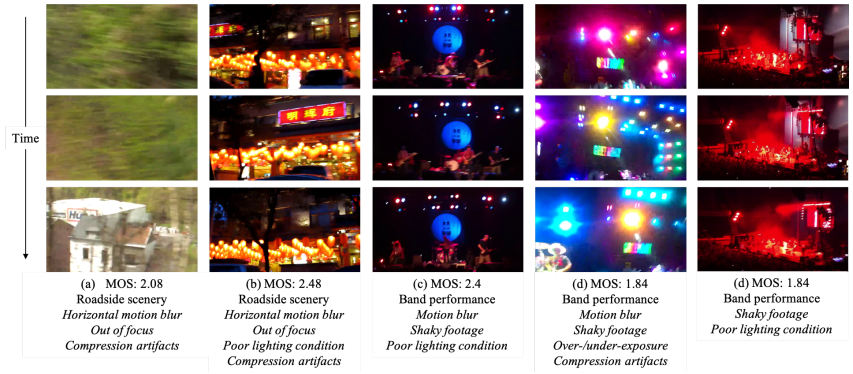
Abstract

Video quality assessment (VQA) is a challenging problem since the perceptual quality of a video can be affected by many factors, *e.g.*, content attractiveness, distortion type and level, motion pattern, and level. Further, the huge expense of annotating limits the scale of VQA datasets, which becomes the main obstacle for deep learning-based VQA methods. In this paper, we propose a VQA method leveraging PreTrained Models, named PTM-VQA, to transfer knowledge from models pretrained on various pre-tasks to benefit VQA from different aspects. Specifically, features of input videos are extracted by different pretrained models with frozen weights, transformed to the same dimension, and integrated to generate the final representation. Since these models possess various fields of knowledge and are often trained with labels irrelevant to quality, we propose an Intra-Consistency and Inter-Divisibility (ICID) loss, which imposes constraints on features extracted by multiple pretrained models from different samples. The intra-consistency constrain is model-wise and requires features extracted by different pretrained models to be in the same unified quality-aware latent space, while the sample-wise inter-divisibility introduces pseudo clusters based on the annotation of samples and tries to separate features of samples from different clusters. Further, confronted with a constantly growing number of pretrained models, it is crucial to determine which ones to use and how to use them. To tackle the problem, we propose an efficient scheme to choose suitable candidates: models that possess better clustering performance on a VQA dataset are chosen to be our candidate backbones. Extensive experiments demonstrate the effectiveness of the proposed method.

1 Introduction

Social network platforms focused on videos have gone viral in recent years. According to the Visual Networking Index (VNI) by Cisco, by the year 2022, the global IP video traffic will account for 82% of all IP traffic (both business and consumer) (Barnett et al., 2018). The substantial growth in the consumption of video content brings tremendous challenges for video providers to deliver better services. Since the perceptual quality of videos significantly affect Quality of Experience (QoE), how to identify quality of videos becomes one of the most important problems (Klink & Uhl, 2020; Chikkerur et al., 2011; Shahid et al., 2014; Fan et al., 2019; Chen et al., 2015). Imitating subjective feedback of human when viewing a video, video quality assessment (VQA) aims to assess the perceptual quality of input videos automatically, and has been studied extensively in the context of assessing compression artifacts, transmission error, and overall quality (Saad et al., 2014; Liu et al., 2018; Mittal et al., 2016; Korhonen, 2019; Li et al., 2019). Compared with conventional methods based on hand-crafted features, data-driven deep learning based methods possess better performance and has been drawing more and more attention (Chen et al., 2022c; 2020; Xu et al., 2021; Kossi et al., 2022; Chen et al., 2022b; You, 2021; Qian et al., 2021; Li et al., 2019; 2021a; Wang et al., 2021).

Compared with other high-level computer vision tasks, datasets for VQA are much smaller. One of the most popular datasets for human action classification Kinetics (Carreira et al., 2019) has 650,000 clips, while the popular VQA dataset KoNViD-1k (Hosu et al., 2017) has only 1,200 videos. One of the reasons is because VQA is a highly subjective task (Winkler, 1999; Wang & Li, 2007). To obtain an unbiased label, it is recommended by annotation guidelines (Rec, 2006) that the subjective quality of a single video should be measured in a laboratory test by calculating the arithmetic mean



(a) MOS: 2.08
Roadside scenery
Horizontal motion blur
Out of focus
Compression artifacts

(b) MOS: 2.48
Roadside scenery
Horizontal motion blur
Out of focus
Poor lighting condition
Compression artifacts

(c) MOS: 2.4
Band performance
Motion blur
Shaky footage
Poor lighting condition

(d) MOS: 1.84
Band performance
Motion blur
Shaky footage
Over-/under-exposure
Compression artifacts

(d) MOS: 1.84
Band performance
Shaky footage
Poor lighting condition

Figure 1: Video frames sampled from KoNViD-1k dataset, which illustrate a certain correlation between content/motion patterns and video quality. We specify some of the reasons that may lead to poor perceptual video quality in italic, following the labeled MOS.

value of multiple subjective judgments, *i.e.*, Mean Opinion Score (MOS). Take KoNViD-1k as an example, it has 114 votes for each video on average. This significantly raises the cost of labeling and limits the size of the VQA dataset. Such a small amount of data limits the power of data-driven VQA methods. To deal with the problem, most existing methods (You, 2021; Chen et al., 2022b; Kossi et al., 2022; Xu et al., 2021) choose to finetune their model on VQA datasets using weights pretrained on common larger datasets (*e.g.*, ImageNet (Deng et al., 2009)). However, existing works (Li et al., 2022; 2019; Wang et al., 2021) show that the perceptual quality of a video is related to many factors, *e.g.*, content attractiveness, aesthetic quality, distortion type and level, motion pattern and level, *etc*. Only considering content-based pretrained models may not be sufficient for VQA. Thus, in this work, we focus on how to better utilize a large amount of available pretrained models to benefit VQA.

To begin with, we notice that there is a certain correlation between VQA tasks and other computer vision tasks. Figure 1 shows several examples from KoNViD-1k dataset. It is natural to expect models pretrained on datasets for various pre-tasks to capture characteristics of different aspects with respect to video quality. In this paper, we conduct a simple clustering experiment using LMNN (Weinberger & Saul, 2009) to observe the correlation between typical pretrained models and the VQA task. Based on the observation, we propose a practical **VQA method leveraging PreTrained Models**, named **PTM-VQA**, which takes pretrained models as pure feature extractors and predicts the quality of input videos based on integrated features. Since the parameters of pretrained models are frozen, we can introduce more pretrained models with limited computational resources. Moreover, we notice that labels in common datasets for pretraining (*e.g.*, object/scene/action) are quite quality-irrelevant. For instance, a clear photo of a puppy with high quality and a blurred photo of a puppy may have the same object-wise label, whereas their quality-wise label may be significantly different. This will confuse the learning process for the VQA task. To tackle the problem, we propose a **I**ntra-**C**onsistency and **I**nter-**D**ivisibility loss, which imposes constraints on features extracted by multiple pretrained models from different samples. Model-wise intra-consistency requires features extracted by different pretrained models to be in the same unified quality-aware latent space, while sample-wise inter-divisibility introduces pseudo clusters based on the MOS label of samples and tries to separate features of samples from different clusters. Further, since the number of pretrained models is constantly growing in the past decade (*e.g.*, PyTorch image models library (Timm) (Wightman, 2019) itself currently supports more than 700 pretrained models), finding models suitable for VQA task through trail-and-error is unpractical. We propose to use Davies–Bouldin Index (DBI) (Davies & Bouldin, 1979) to measure the clustering results and adopt it as the basis of model selection and the weighting for feature integration. To summarize, the main contributions are specified below:

- We verify the correlation between models pretrained on different pre-text tasks and the VQA task and propose a practical NR-VQA method, named PTM-VQA, to exploit cutting-edge pretrained models with diversity to benefit VQA effectively.

- To constrain features with diversity into a unified quality-aware space and eliminate the mismatch between objective (common vision tasks) and perceptual (VQA tasks) annotations, we propose an Intra-Consistency and Inter-Divisibility loss. To avoid looking for a needle in a haystack, we propose an effective way to select candidate models based on DBI, which also determines the contributions of different pretrained models during feature integration.

- PTM-VQA achieves state-of-the-art performance with a rather small amount of learnable weights on three VQA datasets, improving the results to **0.8718 (+0.0303)** and **0.8570 (+0.0273)** in PLCC for KoNViD-1k and YouTube-UGC datasets, respectively. Extensive experiments and ablation studies prove the effectiveness of our method.

2 Related Work

VQA. Based on whether the pristine reference video is required, VQA methods can be classified as Full Reference (FR), Reduced Reference (RR), and No Reference (NR). Our work will be focused on the NR-VQA method, which directly quantifies the perceptual quality of input video, without any other information. Traditional NR-VQA methods either measure video quality by rule-based metric (Yang et al., 2005), or predict MOS by an estimator (*e.g.*, Multi-Layer Perceptron, Support Vector Machine) based on hand-crafted features (Culibrk et al., 2009). In recent years, deep learning-based VQA methods have been studied and surpassed traditional methods. STDAM (Xu et al., 2021) introduced a graph convolution to extract features and a bidirectional long short-term memory network to handle motion information. StarVQA (Xing et al., 2021) proposed encode space-time position information of each patch on video frames and feed them into a Transformer architecture. RAPIQUE (Tu et al., 2021b) proposed to combine conventional features and deep convolutional features. These works, however, neglected the correlation between VQA and other tasks and did not utilize datasets of other tasks. BVQA (Li et al., 2022) took one step further and proposed to transfer knowledge from image quality assessment (IQA) and action recognition datasets to VQA. Our work further investigates the possibility of using more kinds of tasks.

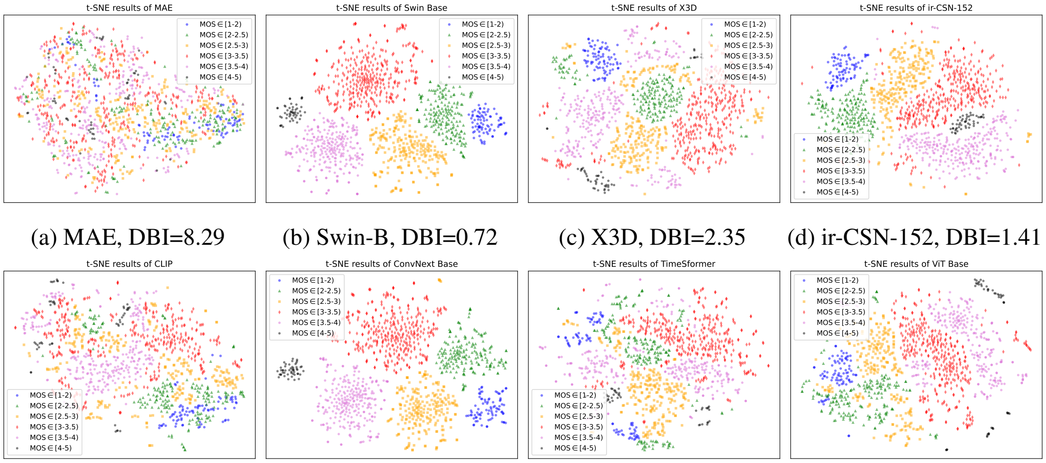
Pretrained models. Pretrained models reveal the great potential in deep learning. In Natural Language Processing (NLP), BERT (Devlin et al., 2019) and GPT-3 (Brown et al., 2020) demonstrated substantial gains on many NLP tasks and benchmarks by pretraining on a large corpus of text followed by finetuning on a specific task. The advent of ViT (Dosovitskiy et al., 2021) had migrated this capability into the visual realm. Some subsequent literature (Radford et al., 2021; He et al., 2021; Li et al., 2021b) had shown that the same benefits can be achieved. For example, CLIP (Radford et al., 2021) trained on the WebImageText matched the performance of the original ResNet-50 on ImageNet zero-shot, without using any of the original labeled data. In the field of quality assessment (QA), there are also efforts (Li et al., 2019; Chen et al., 2020; Mittal et al., 2012; Li et al., 2022) to introduce pretrained models to improve performance. Among them, VSFA (Li et al., 2019) extracted features from a pretrained image classification neural network for its inherent content-aware property. And BVQA (Li et al., 2022) proposed to transfer knowledge from IQA datasets and action recognition datasets with motion patterns. But combining more types of tasks has not been studied.

Metric learning. Metric learning can learn distance metrics from data to measure the difference between samples. It has been used in many research, including QA. RankIQA (Liu et al., 2017) trained a siamese network to rank synthesized images with different levels of distortions constrained by pairwise ranking hinge loss and then finetune the model on the target IQA dataset. UNIQUE (Zhang et al., 2021) sampled ranked image pairs from individual IQA datasets and used a fidelity loss (Tsai et al., 2007) and a hinge constraint to supervise the training process. FPR (Chen et al., 2022a) extracted distortion/reference feature from the input/reference, hallucinated pseudo reference feature from the input alone, and used a triplet loss (Schroff et al., 2015) to pull the pristine and hallucinated reference features closer while pushing the distortion feature away. In our work, we group samples into clusters and propose a centroid triplet loss, trying to pull features of samples within one cluster closer while pushing those from different clusters farther.

3 Method

3.1 Observations

Recently, many researches (Devlin et al., 2019; Brown et al., 2020; Radford et al., 2021; He et al., 2021) are focused on pretraining and demonstrate the effectiveness of applying pretrained models to downstream tasks. This meets the main obstacle of VQA tasks, where huge expense of annotating



(a) MAE, DBI=8.29 (b) Swin-B, DBI=0.72 (c) X3D, DBI=2.35 (d) ir-CSN-152, DBI=1.41

(e) CLIP, DBI=2.49 (f) ConvNeXt, DBI=0.62 (g) TimeSformer, DBI=4.47 (h) ViT-B, DBI=3.15

Figure 2: Visualization of clustering results of features extracted by different pretrained models using t-SNE (van der Maaten, 2009). Videos in KoNViD-1k (Hosu et al., 2017) are used. The number of cluster centers is set to be 6 according to the range of MOS values. And DBI scores, which will be introduced in detail in Sec. 3.4, measure the divergence of clustering results (the smaller, the better).

limits the scale of datasets. In the field of VQA, there are also efforts (Li et al., 2019; Chen et al., 2020; Li et al., 2022) to introduce pretrained models to capture their inherent content-aware properties or motion-related patterns, benefiting the representation of the perceptual quality of videos. However, multitudinous factors in pretrained models may affect the transferring performance (*e.g.*, architecture of neural networks, pre-text tasks, and pretrained databases). Yet to the best of our knowledge, these factors as well as newly-appeared cutting-edge pretrained models are rarely explored and exploited in the VQA field. So we intend to find a way to make full use of these models.

To verify the correlation between pretrained models and VQA tasks, we construct a simple clustering experiment. First, a selected pretrained model, whose weights are frozen, is performed as a feature extractor to obtain the corresponding features of videos. Then these features are clustered into multiple centers using LMNN (Weinberger & Saul, 2009) according to their range of MOS values. Based on aforementioned factors, we take eight models for example, including MAE (He et al., 2021) trained on ImageNet-1k (Deng et al., 2009), Swin-Base (Liu et al., 2021) trained on ImageNet-22k (Deng et al., 2009), X3D (Feichtenhofer, 2020) trained on Kinetics-400 (Kay et al., 2017), ir-CSN-152 (Tran et al., 2019) trained on Sports-1M (Karpathy et al., 2014), CLIP (Radford et al., 2021) trained on WebImageText (Radford et al., 2021), ConvNeXt (Liu et al., 2022) trained on ImageNet-22k, TimeSformer (Bertasius et al., 2021) trained on Kinetics-400 and ViT-Base (Dosovitskiy et al., 2021) trained on ImageNet-22k. As shown in Fig. 2, some models show surprising discriminant results even though they had not been exposed to quality-related labels during the training of pre-text tasks. We speculate that during the training of pre-text tasks, some quality-aware representations have been learned simultaneously. Take CLIP which learns visual concepts from natural language supervision as an example, some texts may contain emotional descriptions related to the quality of images. And other models trained on action recognition tasks (*e.g.*, ir-CSN-152, X3D) could be sensitive to motion-related distortions (Li et al., 2022) (*e.g.*, camera shaking or motion blurriness). These broader pretrained models have the potential to help VQA tasks achieve better performance.

3.2 Pipeline of the proposed PTM-VQA

Suppose there exist multiple available pretrained models, the most intuitive way to apply them to VQA tasks is finetuning on target datasets and integrating extracted features for quality prediction. Nevertheless, this is highly computationally resource-consuming and becomes less practical as the number of pretrained models increases and the models get larger. For example, the training of ViT (Dosovitskiy et al., 2021) requires a TPUv3 with 8 cores in approximately 30 days. And MAE (He et al., 2021) consumes 128 TPUv3 cores for its 800-epoch training. This would be unaffordable in a VQA task. Fortunately, the above results in Fig. 2 suggest that pretrained models can be applied

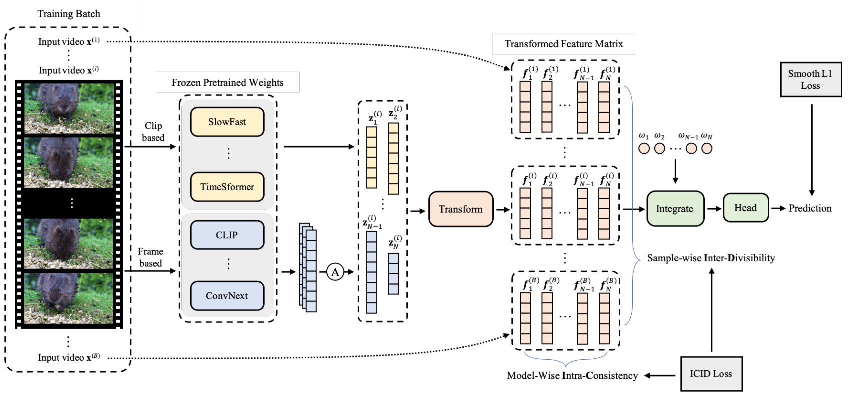

Training Batch
Input video $\mathbf{x}^{(1)}$
Input video $\mathbf{x}^{(i)}$
Input video $\mathbf{x}^{(B)}$
Clip based
Frame based
Frozen Pretrained Weights
SlowFast
TimeSformer
CLIP
ConvNext
A
$\mathbf{z}_1^{(i)}$ $\mathbf{z}_2^{(i)}$
$\mathbf{z}_{N-1}^{(i)}$ $\mathbf{z}_N^{(i)}$
Transform
Transformed Feature Matrix
$\boldsymbol{f}_1^{(1)}$ $\boldsymbol{f}_2^{(1)}$ $\boldsymbol{f}_{N-1}^{(1)}$ $\boldsymbol{f}_N^{(1)}$
$\boldsymbol{f}_1^{(i)}$ $\boldsymbol{f}_2^{(i)}$ $\boldsymbol{f}_{N-1}^{(i)}$ $\boldsymbol{f}_N^{(i)}$
$\boldsymbol{f}_1^{(B)}$ $\boldsymbol{f}_2^{(B)}$ $\boldsymbol{f}_{N-1}^{(B)}$ $\boldsymbol{f}_N^{(B)}$
ω_1 ω_2 ω_{N-1} ω_N
Integrate
Head
Prediction
Smooth L1 Loss
Sample-wise Inter-Divisibility
Model-Wise Intra-Consistency
ICID Loss

Figure 3: The pipeline of the proposed PTM-VQA. Features of input videos are extracted by pretrained models with frozen weights, transformed to the same dimension, and integrated to generate the final representation. Expect for the ordinary smooth $\mathcal{L}_1$ loss for regression, we add an ICID loss to ensure model-wise consistency and sample-wise divisibility.

directly with their weights frozen. In this paper, we propose a simple framework, named PTM-VQA, to utilize the knowledge from pretrained models of diversity efficiently.

As shown in Fig. 3, given an input video $\boldsymbol{x}^{(i)}$, N pretrained models, whose weights are frozen, are utilized to extract features, resulting in representations from different perspectives. Specifically, for video clip-based models, we uniformly sample frames in the temporal dimension to form the input clip. Corresponding representations are then generated by these models. For frame-based models, they are fed with sampled frames and the output features are averaged to perform the spatiotemporal representation. Features extracted by models can be noted as $\mathbf{z}_n^{(i)}$, where $n \in \{1, \dots, N\}$. To further distill quality-aware features and perform dimension alignment, we apply a learnable transformation module following each feature extractor. Structurally, the transformation module consists of two fully connected layers, each followed by a normalization layer and an activation layer of GELU. The transformed features are defined as $\boldsymbol{f}_n^{(i)} \in \mathbb{R}^D$, where D represents the aligned dimension. Then features are integrated to obtain a unified representation through:

$$\boldsymbol{h}^{(i)} = \frac{\sum_{n=1}^{N} \omega_n \boldsymbol{f}_n^{(i)}}{\sum_{n=1}^{N} \omega_n}, \tag{1}$$

where ω_n is the coefficient for each model. When ω_n is $\frac{1}{N}$, it means calculating an average, with each model contributing equally to the final representation. Last, $\boldsymbol{h}^{(i)}$ is used to get the quality prediction through a regression head, which is a single fully-connected layer.

Based on this design, the training procedure becomes very efficient and avoids the computational burden met by the aforementioned finetuning paradigm. Referring to the performance in subsequent Tab. 1, the whole training process can be completed in about two hours, on a single GPU. This retains the information of the pretrained models well, but it also increases the difficulty of obtaining preferable performance due to the reduction of learnable parameters. Some concerns are as follows:

- Due to various pre-texts of pretrained models, features generated by different models are of large diversity, which may distribute over inconsistent feature spaces (Wortsman et al., 2022). How to constrain these abundant features into a unified quality-aware space is important.
- Different from the objective category in common classification tasks, the perceptual quality of a video is more implicit and related to various factors (*e.g.*, content attractiveness, distortion type and level, motion pattern and level), whereas videos of the same quality often render completely different content and vice versa. Therefore, it is difficult for the models trained based on objective annotations to distinguish these samples of the same category but with a large perceptual quality difference. A more comprehensive contrast approach beyond sample-wise comparison needs to be proposed to deal with these outliers.
- There exists hundreds of pretrained models available in public libraries. How to select the desired models efficiently and how determining the contribution of these models to represent the perceptual quality effectively is an urgent problem to be solved.

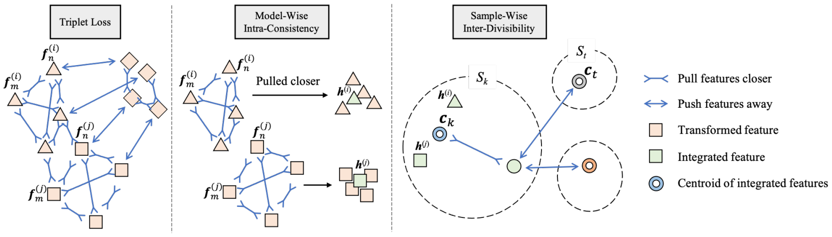


Figure 4: Illustration of triplet loss and the proposed ICID loss. The figure shows examples of several triplets of triplet loss; two sets of model-wise intra-consistency between features extracted by four pretrained models; and one sample (with two triplets) for sample-wise inter-divisibility.

3.3 Intra-Consistency and Inter-Divisibility Loss

To solve the above concerns and better satisfy VQA tasks, we intend to constrain the features between different pretrained models and different samples using metric learning. Triplet loss, which is one of the most widely adopted metric learning measures, can be formed as follows:

$$\mathcal{L}_{\text{triplet}}(\boldsymbol{f}_{\hat{a}}, \boldsymbol{f}_{\hat{p}}, \boldsymbol{f}_{\hat{n}}) = \max(\|\boldsymbol{f}_{\hat{a}} - \boldsymbol{f}_{\hat{p}}\|^2 - \|\boldsymbol{f}_{\hat{a}} - \boldsymbol{f}_{\hat{n}}\|^2 + \alpha, 0), \tag{2}$$

where $\boldsymbol{f}_{\hat{a}}$,$\boldsymbol{f}_{\hat{p}}$,$\boldsymbol{f}_{\hat{n}}$ are features of an anchor sample $\hat{a}$, a positive sample $\hat{p}$ of the same class as $\hat{a}$, and a negative sample $\hat{n}$ which has a different class of $\hat{a}$. And α is a margin between anchor-positive and anchor-negative pairs. Some previous studies (Chen et al., 2022a; Golestaneh et al., 2022) in QA also applied triplet loss to measure the distance between the distorted feature and the reference feature of the same sample. Since the MOS values are continuous, the original triplet loss cannot be directly used to constrain the distance between arbitrary samples. We make some modifications to constrain features generated by different pretrained models and samples, as given in Fig. 4.

Intra-consistency constraint. To constrain features generated by different pretrained models into a unified quality-aware latent space, we propose a model-wise intra-consistency constraint. Formally, it is defined to minimize the distance between arbitrary two of the transformed features through computing a cosine similarity, which is widely used in deep metric learning (Wang et al., 2019a):

$$\mathcal{L}_{\text{intra}} = \frac{2}{N \cdot (N-1)} \sum_{n=1}^{N} \sum_{m, m \neq n}^{N} \big(1 - \frac{\boldsymbol{f}_n^{(i)} \cdot \boldsymbol{f}_m^{(i)}}{\|\boldsymbol{f}_n^{(i)}\|_2 \|\boldsymbol{f}_m^{(i)}\|_2}\big). \tag{3}$$

Inter-divisibility constraint. To constrain features generated by different samples, we split videos into distinct pseudo clusters under different numerical intervals, according to the annotated MOS values (on a scale of 1.0 to 5.0). For example, videos with MOS in the range of 1.0 to 2.0 are generally considered to be of poor quality, and whose content cannot be normally recognized due to the existence of various distortions. And videos with MOS in the range of 4.0 to 5.0 are of high quality, whose content is unambiguous, without noise, shaking, and blurring. We identify the videos within the same range as the same category, thus dividing them into K clusters. Each cluster can be noted as $\mathcal{S}_k = \{\boldsymbol{x}^{(i)} | y^{(i)} \in (p_k, q_k], q_k > p_k \in [1.0, 5.0]\}$, where $y^{(i)}$ is the labeled MOS for the i-th input video, p_k and q_k are the endpoints of the interval. Through this pseudo cluster, triplet loss can be utilized for samples belonging to the same cluster to be closer and samples of the different clusters to be farther away. Then Eq. (2) can be rewritten as:

$$\mathcal{L}_{\text{triplet}}(\boldsymbol{h}^{(i)}, \boldsymbol{h}^{(j)}, \boldsymbol{h}^{(l)}), \text{ where } \boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)} \in \mathcal{S}_k, \boldsymbol{x}^{(l)} \notin \mathcal{S}_k. \tag{4}$$

Besides, the original feature $\boldsymbol{f}$ extracted by individual models are replaced by the integrated feature $\boldsymbol{h}$.

As shown in Fig. 5, the original triplet loss performs a sample-to-sample form, which is highly affected by the sampling of triples. When facing outliers that are of the same quality but render different contents, or vice versa, it may lead to bad local minima and prevent the model from achieving top performance. To solve the second concern aforementioned, we propose using the centroid of the cluster to represent the positive and negative points as:

$$\mathcal{L}_{\text{inter}} = \mathcal{L}_{\text{triplet}}(\boldsymbol{h}^{(i)}, \boldsymbol{c}_k, \boldsymbol{c}_t), \text{ and } \boldsymbol{c}_k = \frac{1}{|\mathcal{S}_k|} \sum_{\{i | \boldsymbol{x}^{(i)} \in \mathcal{S}_k\}} \boldsymbol{h}^{(i)}, \boldsymbol{c}_t = \frac{1}{|\mathcal{S}_t|} \sum_{\{j | \boldsymbol{x}^{(j)} \in \mathcal{S}_t\}} \boldsymbol{h}^{(j)}. \tag{5}$$

Given a batch consisting B inputs, during training, the optimization objective can be summarized as:

$$\min \mathcal{L}_1 + \beta\big(\sum\nolimits_{i=1}^{B} \mathcal{L}_{\text{intra}} + \mathcal{L}_{\text{inter}}\big), \tag{6}$$

where β is the coefficient balancing smooth $\mathcal{L}_1$ regression loss and the proposed ICID loss.

3.4 Selection scheme through DBI

We observe an obvious difference in the clustering results of different pretrained models in Fig. 2. Since the weights of models are frozen both in the clustering test and subsequent training process, the divergence of clustering results can reflect the relevance of VQA tasks. We propose using the Davies–Bouldin Index (DBI) (Davies & Bouldin, 1979), which is used as a metric for evaluating clustering algorithms, for model selection. In our setting, it can be noted as:

$$\psi = \frac{1}{K}\sum_{k=1}^{K}\max_{t\neq k}\frac{d_k+d_t}{\|\boldsymbol{c}_k-\boldsymbol{c}_t\|_2}, \text{ and } \boldsymbol{c}_k = \frac{1}{|\mathcal{S}_k|}\sum_{\mathcal{S}_k}\mathbf{z}^{(i)}, d_k = \frac{1}{|\mathcal{S}_k|}\sum_{\mathcal{S}_k}\|\mathbf{z}^{(i)}-\boldsymbol{c}_k\|_2, \tag{7}$$

where $\boldsymbol{c}_k$ is the centroid of cluster $\mathcal{S}_k$ for the set of extracted feature $\mathbf{z}^{(i)}$, d_k represents the average distance between each sample and its corresponding centroid. For the n-th model, its DBI score can be noted as ψ_n. A lower DBI indicates better clustering performance, which means that the pretrained model (*e.g.*, ConvNeXt, Swin-Base, ir-CSN-152, CLIP in Fig. 2) is more relevant to the VQA task. During training, the DBI scores computed offline can be used in the aggregation procedure as given in Eq. (1), where ω_n can be replaced by $\frac{1}{\psi_n}$. It means the models that are more relevant to the VQA task contribute more to the feature representation.

4 Experiments

4.1 Datasets and evaluation criteria

Datasets. Our method is evaluated on 3 public NR-VQA datasets, including KoNViD-1k (Hosu et al., 2017), LIVE-VQC (Sinno & Bovik, 2019) and YouTube-UGC (Wang et al., 2019b). In detail, KoNViD-1k contains 1,200 videos that are fairly filtered from a large public video dataset YFCC100M. The videos are 8 seconds long with 24/25/30 FPS and a resolution of 960×540. The MOS ranges from 1.22 to 4.64. Each video owns 114 annotations to get a reliable MOS. LIVE-VQC consists of 585 videos with complex authentic distortions captured by 80 different users using 101 different devices, with 240 annotations for each video. YouTube-UGC has 1,380 UGC videos sampled from YouTube with a duration of 20 seconds and resolutions from 360P to 4K, with 123 annotations for each video. All the datasets contain no pristine videos, thus only NR methods can be evaluated on them. Following (Xu et al., 2021; Su et al., 2020), we split the dataset into a 80% training set and a 20% testing set randomly for all three datasets. We perform 10 repeat runs in each dataset using different splittings to get the mean values of PLCC and SRCC to eliminate the bias.

Evaluation criteria. Pearson's Linear Correlation Coefficient (PLCC) and Spearman's Rank-Order Correlation Coefficient (SRCC) are selected as criteria to measure the accuracy and monotonicity. They are in the range of $[0, 1]$. A larger PLCC means a more accurate numerical fit with MOS scores. A larger SRCC shows a more accurate ranking between samples. Besides, the mean average of PLCC and SRCC is also reported as a comprehensive criterion.

4.2 Implementation details

Our experiments are performed using PyTorch (Paszke et al., 2019) and MMAction2 (Contributors, 2020), and are all conducted on **one** Nvidia V100 GPU by training for 60 epochs. For KoNViD-1k, we select ConvNeXt, ir-CSN-152, and CLIP as feature extractors. For LIVE-VQC, we use CLIP and TimeSformer. For YouTube-UGC, an extra Video Swin-Base is used together with those selected on KoNViD-1k. For KoNViD-1k, we sample 16 frames with a frame interval of 2. As videos in LIVE-VQC and YouTube-UGC has a longer time duration, we use larger intervals for these two datasets. Since most augmentations will introduce extra interference to the quality of videos (Ke et al., 2021), we only choose the center crop to produce an input with a size of 224×224. During training, we use AdamW optimizer with a weight decay of 0.02. Cosine annealing with a warmup of 2 epochs

Table 1: Training details for different datasets of PTM-VQA.

Dataset	Pretrained Models	Frames	Interval	Initial LR	Time(h)	Param(M)	Mem(G)
KoNViD-1k	ConvNeXt, ir-CSN-152, CLIP	16	2	1e-3	2.00	0.66	4.94
LIVE-VQC	CLIP, TimeSformer	16	4	5e-3	1.97	0.30	4.32
YouTube-UGC	CLIP, ir-CSN-152, CLIP, Video Swin-B	32	8	1e-3	2.34	0.86	5.32

Table 2: Comparisons with existing methods. The "-" is an unreported result. The "*" means using extra labeled QA data for training. The best and second best results are **bolded** and underlined.

Method	KoNViD-1k			LIVE-VQC			YouTube-UGC		
	PLCC	SRCC	Mean	PLCC	SRCC	Mean	PLCC	SRCC	Mean
VIIDEO (Mittal et al., 2016)	0.3030	0.2980	0.3005	0.2164	0.0332	0.1248	0.1534	0.0580	0.1057
NIQE (Mittal et al., 2013)	0.5530	0.5417	0.5473	0.6286	0.5957	0.6121	0.2776	0.2379	0.2577
BRISQUE (Mittal et al., 2012)	0.626	0.654	0.640	0.638	0.592	0.615	0.395	0.382	0.388
VSFA (Li et al., 2019)	0.744	0.755	0.749	-	-	-	-	-	-
TLVQM (Korhonen, 2019)	0.7688	0.7729	0.7708	0.8025	0.7988	0.8006	0.6590	0.6693	0.6641
RIRNet (Chen et al., 2020)	0.7812	0.7755	0.7783	0.7982	0.7713	0.7847	-	-	-
UGC-VQA (Tu et al., 2021a)	0.7803	0.7832	0.7817	0.7514	0.7522	0.7518	0.7733	0.7787	0.7760
CSPT (Chen et al., 2022c)	0.8062	0.8008	0.8035	0.8194	0.7989	0.8091	-	-	-
RAPIQUE (Tu et al., 2021b)	0.8175	0.8031	0.8103	0.7863	0.7548	0.7705	0.7684	0.7591	0.7637
StarVQA (Xing et al., 2021)	0.796	0.812	0.804	0.808	0.732	0.770	-	-	-
BVQA* (Li et al., 2022)	0.8335	0.8362	0.8348	**0.8415**	**0.8412**	**0.8413**	0.8194	0.8312	0.8253
STDAM* (Xu et al., 2021)	0.8415	0.8448	0.8431	0.8204	0.7931	0.8067	0.8297	0.8341	0.8319
PTM-VQA	**0.8718**	**0.8568**	**0.8643**	0.8198	0.8110	0.8154	**0.8570**	**0.8578**	**0.8574**

is adopted to control the learning rate. The dimension D of transformed features is set to 128. The margin α is set to be 0.05. β is set to be 0.2. By default, we select the checkpoint generated by the last iteration for evaluation. During inference, we follow a similar procedure as given in (Arnab et al., 2021) by using 4×5 views. To be specific, 4 clips are uniformly sampled from a video in the temporal domain. For each clip, we take 5 crops in the four corners and the center. The final score is computed as the average score of all the views. More training details are given in Tab. 1.

4.3 Comparison with SOTA methods

We select existing VQA methods for comparison in three datasets. As shown in Tab. 2, our method obtains competitive results on all three datasets. Compared with traditional methods that rely on statistical regularities (*e.g.*, VIIDEO, NIQE, and BRISQUE), PTM-VQA models outperform by large margins. Compared with some deep learning-based methods that apply well-designed networks (*e.g.*, TLVQM, StarVQA), PTM-VQA still obtains higher performances. Especially, VSFA and RIRNet also adopt pretrained models that contain content-dependency or motion information to finetune in VQA tasks. PTM-VQA demonstrates that features extracted directly from pretrained models can also achieve better results. As the best two SOTA methods BVQA and STDAM who utilize extra IQA datasets, PTM-VQA proves that transferring knowledge from pretrained models can achieve competitive results compared with a model trained with additional data. PTM-VQA improved SOTA's PLCC by 3.02%, SRCC by 1.20%, and mean score by 2.12% on KoNViD-1k. And PTM-VQA improved SOTA's PLCC by 2.73%, SRCC by 2.37%, and mean score by 2.55% on YouTube-UGC.

4.4 Experimental analysis

In this section, we conduct a performance analysis to evaluate the effectiveness of each proposed component. By default, experiments are performed following the best configurations in KoNViD-1k.

Ablation on different constraints. As given in Tab. 3, direct usage of triplet loss cannot obtain satisfy results. When either or both constraints are absent, performance degrades significantly. These prove the effectiveness of intra-consistency in transferring knowledge from different pretrained models and inter-divisibility in generating stable predictions.

Ablation on the clustering settings. Tab. 4 gives the results with different number of clusters. When K is 2, videos are simply classified as low-quality and high-quality ones. When K is 4, videos are evenly divided into four parts on a scale of 1.0 to 5.0. Due to the relatively small amount of data at both endpoints, a 6-split setting can be obtained by using fine-grained division in the middle fraction segment. Since the need to ensure the number of samples per cluster within the batch, a larger number of clusters are not attempted. The best result can be acquired when K is 6.

Table 3: Ablation on constraints.

$\mathcal{L}_1$	$\mathcal{L}_{\text{intra}}$	$\mathcal{L}_{\text{inter}}$	$\mathcal{L}_{\text{tri}}$	PLCC	SRCC
✓	✓	✓		**0.8718**	**0.8568**
✓			✓	0.7968	0.7850
✓				0.7867	0.7655
✓	✓			0.8545	0.8299
✓		✓		0.8172	0.7707

Table 4: Ablation on cluster settings.

K	intervals	PLCC	SRCC
2	$\mathcal{S}_1$=[1,3), $\mathcal{S}_2$=[3, 5]	0.8277	0.8066
4	$\mathcal{S}_1$=[1,2), $\mathcal{S}_2$=[2,3), $\mathcal{S}_3$=[3,4), $\mathcal{S}_4$=[4,5]	0.8431	0.8012
6	$\mathcal{S}_1$=[1,2), $\mathcal{S}_2$=[2, 2.5), $\mathcal{S}_3$=[2.5, 3), $\mathcal{S}_4$=[3, 3.5), $\mathcal{S}_5$=[3.5, 4), $\mathcal{S}_6$=[4, 5]	**0.8718**	**0.8568**

Ablation on the effectiveness of DBI. The effectiveness of DBI can be evaluated in 2 aspects: (1) correlation between offline metrics and final results as given in Tab. 5; (2) integration of different contributions during training as shown in Tab. 6 and combinations using random strategy in Tab. 7. Tab. 5 indicates that models with higher DBI scores (*e.g.*, MAE and ViT-B) are more likely to perform poorly on downstream tasks, and vice versa (*e.g.*, CLIP, ConvNeXt). Tab. 6 shows the effectiveness of DBI in guiding the integration of different models. Tab. 7 shows that the selection scheme based on DBI is better than random selection.

Table 5: Correlation of DBIs and performances.

Model	DBI	KoN1k	LIVE	YT
MAE	8.29	0.7169	0.7807	0.7631
Swin-B	0.72	0.7892	0.7701	0.7816
X3D	2.35	0.6165	0.5995	0.6432
ir-CSN152	1.41	0.7647	0.6304	0.7349
CLIP	2.49	0.8398	0.7832	0.8089
ConvNeXt	0.62	0.7794	0.7554	0.7988
TimeSformer	4.47	0.8044	0.7427	0.7541
ViT-B	3.15	0.7879	0.7353	0.7218

Table 6: Ablation on the types of ω_n.

Datasets	ω_n	PLCC	SRCC
KoNViD-1k	$1/N$	0.8631	0.8521
	$1/\psi$	0.8718	0.8568
LIVE-VQC	$1/N$	0.8205	0.8197
	$1/\psi$	0.8198	0.8110
YouTube-UGC	$1/N$	0.8427	0.8446
	$1/\psi$	0.8570	0.8578

Table 7: Ablation on the selection of pretrained models on different datasets.

Dataset	N	selected pretrained models	PLCC	SRCC
KoNViD-1k	1	CLIP	0.8398	0.8083
	2	CLIP, ir-CSN-152	0.8520	0.8180
	3	**CLIP, ir-CSN-152, ConvNeXt**	**0.8718**	**0.8568**
	4	CLIP, ir-CSN-152, ConvNeXt, Swin-Base	0.8634	0.8423
	3	ViT Base, MAE, Swin-Base (Random)	0.8317	0.7912
	3	SlowFast, TimeSformer, MAE (Random)	0.8135	0.7863
LIVE-VQC	1	CLIP	0.7832	0.7779
	2	**CLIP, TimeSformer**	**0.8198**	**0.8110**
	3	CLIP, TimeSformer, Video Swin-Base	0.8192	0.8075
	4	CLIP, TimeSformer, Video Swin-Base, MAE	0.8107	0.8038
	4	X3D, TimeSformer, SlowFast, Swin Base (Random)	0.7851	0.7712
	4	ConvNeXt,SlowFast, MAE, Swin Base (Random)	0.7768	0.7626
YouTube-UGC	1	CLIP	0.8089	0.8236
	2	CLIP, ir-CSN-152	0.8067	0.8243
	3	CLIP, ir-CSN-152, ConvNeXt	**0.8589**	0.8552
	4	**CLIP, ir-CSN-152, ConvNeXt, Swin-Base**	0.8570	**0.8578**
	4	SlowFast, TimeSformer, MAE, Swin-Base (Random)	0.8343	0.8208
	4	X3D, SlowFast, ViT Base, ConvNeXt (Random)	0.8167	0.8126

5 Conclusion and future work

In this paper, we propose a new PTM-VQA framework that utilizes in-the-wild pretrained models as feature extractors for VQA tasks. The DBI scores are utilized to select candidates from a large amount of available pretrained models. To constrain features with large diversity into a unified quality-aware latent space and tackle outliers (*e.g.*, render different content but of the same perceptual quality), we propose a new Intra-Consistency and Inter-Divisibility loss. Under small computational cost, PTM-VQA models obtain SOTA results in widely-used NR-VQA benchmarks. Furthermore, how to further use the pretrained models is still an open question with great practical significance. There are some problems worthy of research, which we would like to explore in future work: (1) is there a more effective way to select models? (2) We find that different datasets require different model combinations (by trial-and-error) for optimal performance. Is there an automatic selection manner? (3) What exactly did the pretrained models migrate? We hope to inspire subsequent related research.

REFERENCES

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. Vivit: A video vision transformer. *CoRR*, abs/2103.15691, 2021.

Thomas Barnett, Shruti Jain, Usha Andra, and Taru Khurana. Cisco visual networking index (vni) complete forecast update, 2017–2022. *Americas/EMEAR Cisco Knowledge Network (CKN) Presentation*, 2018.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 813–824. PMLR, 2021.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020.

João Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *CoRR*, abs/1907.06987, 2019.

Baoliang Chen, Lingyu Zhu, Chenqi Kong, Hanwei Zhu, Shiqi Wang, and Zhu Li. No-reference image quality assessment by hallucinating pristine features. *IEEE Trans. Image Process.*, 2022a.

Baoliang Chen, Lingyu Zhu, Guo Li, Fangbo Lu, Hongfei Fan, and Shiqi Wang. Learning generalized spatial-temporal deep feature representation for no-reference video quality assessment. *IEEE Trans. Circuits Syst. Video Technol.*, 32(4):1903–1916, 2022b.

Pengfei Chen, Leida Li, Lei Ma, Jinjian Wu, and Guangming Shi. Rirnet: Recurrent-in-recurrent network for video quality assessment. In *ACM Multimedia*, pp. 834–842. ACM, 2020.

Pengfei Chen, Leida Li, Jinjian Wu, Weisheng Dong, and Guangming Shi. Contrastive self-supervised pre-training for video quality assessment. *IEEE Trans. Image Process.*, 31:458–471, 2022c.

Yanjiao Chen, Kaishun Wu, and Qian Zhang. From qos to qoe: A tutorial on video quality assessment. *IEEE Commun. Surv. Tutorials*, 17(2):1126–1165, 2015.

Shyamprasad Chikkerur, Vijay Sundaram, Martin Reisslein, and Lina J. Karam. Objective video quality assessment methods: A classification, review, and performance comparison. *IEEE Trans. Broadcast.*, 57(2):165–182, 2011.

MMAction2 Contributors. Openmmlab's next generation video understanding toolbox and benchmark. `https://github.com/open-mmlab/mmaction2`, 2020.

Dubravko Culibrk, Dragan Kukolj, Petar Vasiljevic, Maja Pokric, and Vladimir Zlokolica. Feature selection for neural-network based no-reference video quality assessment. In *ICANN (2)*, volume 5769 of *Lecture Notes in Computer Science*, pp. 633–642. Springer, 2009.

David L. Davies and Donald W. Bouldin. A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1(2):224–227, 1979.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255. IEEE Computer Society, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pp. 4171–4186. Association for Computational Linguistics, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*. OpenReview.net, 2021.

Qiang Fan, Wang Luo, Yuan Xia, Guozhi Li, and Daojing He. metrics and methods of video quality assessment: a brief review. *Multim. Tools Appl.*, 78(22):31019–31033, 2019.

Christoph Feichtenhofer. X3D: expanding architectures for efficient video recognition. In *CVPR*, pp. 200–210. Computer Vision Foundation / IEEE, 2020.

S. Alireza Golestaneh, Saba Dadsetan, and Kris M. Kitani. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *WACV*, pp. 1220–1230, January 2022.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *CoRR*, abs/2111.06377, 2021.

Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe. The konstanz natural video database (konvid-1k). In *QoMEX*, pp. 1–6. IEEE, 2017.

Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: multi-scale image quality transformer. In *ICCV*, pp. 5128–5137. IEEE, 2021.

Janusz Klink and Tadeus Uhl. Video quality assessment: Some remarks on selected objective metrics. In *SoftCOM*, pp. 1–6. IEEE, 2020.

Jari Korhonen. Two-level approach for no-reference consumer video quality assessment. *IEEE Trans. Image Process.*, 28(12):5923–5938, 2019.

Koffi Kossi, Stéphane Coulombe, Christian Desrosiers, and Ghyslain Gagnon. No-reference video quality assessment using distortion learning and temporal attention. *IEEE Access*, 10:41010–41022, 2022.

Bowen Li, Weixia Zhang, Meng Tian, Guangtao Zhai, and Xianpei Wang. Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception. *IEEE Trans. Circuits Syst. Video Technol.*, 32(9):5944–5958, 2022.

Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *ACM Multimedia*, pp. 2351–2359. ACM, 2019.

Dingquan Li, Tingting Jiang, and Ming Jiang. Unified quality assessment of in-the-wild videos with mixed datasets training. *Int. J. Comput. Vis.*, 129(4):1238–1257, 2021a.

Yanghao Li, Saining Xie, Xinlei Chen, Piotr Dollár, Kaiming He, and Ross B. Girshick. Benchmarking detection transfer learning with vision transformers. *CoRR*, abs/2111.11429, 2021b.

Wentao Liu, Zhengfang Duanmu, and Zhou Wang. End-to-end blind quality assessment of compressed videos using deep neural networks. In *ACM Multimedia*, pp. 546–554. ACM, 2018.

Xialei Liu, Joost van de Weijer, and Andrew D. Bagdanov. Rankiqa: Learning from rankings for no-reference image quality assessment. In *ICCV*, pp. 1040–1049. IEEE Computer Society, 2017.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pp. 9992–10002. IEEE, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *CoRR*, abs/2201.03545, 2022.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.*, 21(12):4695–4708, 2012.

Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a ”completely blind” image quality analyzer. *IEEE Signal Process. Lett.*, 20(3):209–212, 2013.

Anish Mittal, Michele A. Saad, and Alan C. Bovik. A completely blind video integrity oracle. *IEEE Trans. Image Process.*, 25(1):289–300, 2016.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pp. 8024–8035, 2019.

Lihui Qian, Tianxiang Pan, Yunfei Zheng, Jiajie Zhang, Mading Li, Bing Yu, and Bin Wang. No-reference nonuniform distorted video quality assessment based on deep multiple instance learning. *IEEE Multim.*, 28(1):28–37, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

ITUT Rec. P. 800.1, mean opinion score (mos) terminology. *International Telecommunication Union, Geneva*, 2006.

Michele A. Saad, Alan C. Bovik, and Christophe Charrier. Blind prediction of natural video quality. *IEEE Trans. Image Process.*, 23(3):1352–1365, 2014.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pp. 815–823. IEEE Computer Society, 2015.

Muhammad Shahid, Andreas Rossholm, Benny Lövström, and Hans-Jürgen Zepernick. No-reference image and video quality assessment: a classification and review of recent approaches. *EURASIP J. Image Video Process.*, 2014:40, 2014.

Zeina Sinno and Alan Conrad Bovik. Large-scale study of perceptual video quality. *IEEE Trans. Image Process.*, 28(2):612–627, 2019.

Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *CVPR*, pp. 3664–3673. Computer Vision Foundation / IEEE, 2020.

Du Tran, Heng Wang, Matt Feiszli, and Lorenzo Torresani. Video classification with channel-separated convolutional networks. In *ICCV*, pp. 5551–5560. IEEE, 2019.

Ming-Feng Tsai, Tie-Yan Liu, Tao Qin, Hsin-Hsi Chen, and Wei-Ying Ma. Frank: a ranking method with fidelity loss. In *SIGIR*, pp. 383–390. ACM, 2007.

Zhengzhong Tu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C. Bovik. UGC-VQA: benchmarking blind video quality assessment for user generated content. *IEEE Trans. Image Process.*, 30:4449–4464, 2021a.

Zhengzhong Tu, Xiangxu Yu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C. Bovik. RAPIQUE: rapid and accurate video quality prediction of user generated content. *CoRR*, abs/2101.10955, 2021b.

Laurens van der Maaten. Learning a parametric embedding by preserving local structure. In *AISTATS*, volume 5 of *JMLR Proceedings*, pp. 384–391. JMLR.org, 2009.

Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R. Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *CVPR*, pp. 5022–5030. Computer Vision Foundation / IEEE, 2019a.

Yilin Wang, Sasi Inguva, and Balu Adsumilli. Youtube UGC dataset for video compression research. In *MMSP*, pp. 1–5. IEEE, 2019b.

Yilin Wang, Junjie Ke, Hossein Talebi, Joong Gon Yim, Neil Birkbeck, Balu Adsumilli, Peyman Milanfar, and Feng Yang. Rich features for perceptual quality assessment of UGC videos. In *CVPR*, pp. 13435–13444. Computer Vision Foundation / IEEE, 2021.

Zhou Wang and Qiang Li. Video quality assessment using a statistical model of human visual speed perception. *JOSA A*, 24(12):B61–B69, 2007.

Kilian Q. Weinberger and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.*, 10:207–244, 2009.

Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

Stefan Winkler. Issues in vision modeling for perceptual video quality assessment. *Signal Process.*, 78(2):231–252, 1999.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23965–23998. PMLR, 2022.

Fengchuang Xing, Yuan-Gen Wang, Hanpin Wang, Leida Li, and Guopu Zhu. Starvqa: Space-time attention for video quality assessment. *CoRR*, abs/2108.09635, 2021.

Jiahua Xu, Jing Li, Xingguang Zhou, Wei Zhou, Baichao Wang, and Zhibo Chen. Perceptual quality assessment of internet videos. In *ACM Multimedia*, pp. 1248–1257. ACM, 2021.

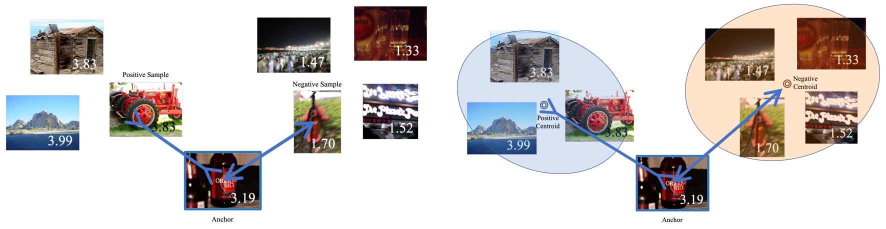
Fuzheng Yang, Shuai Wan, Yilin Chang, and Hong Ren Wu. A novel objective no-reference metric for digital video quality assessment. *IEEE Signal Process. Lett.*, 12(10):685–688, 2005.

Junyong You. Long short-term convolutional transformer for no-reference video quality assessment. In *ACM Multimedia*, pp. 2112–2120. ACM, 2021.

Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Trans. Image Process.*, 30:3474–3486, 2021.

A Appendix

A.1 Comparison with the original triplet loss



(a) Sample-based triplet.

(b) Centroid-based triplet.

Figure 5: Comparison of the sample-based and centroid-based triplet. Videos of the same quality often render completely different content and vice versa, increasing the difficulty for models trained on objective tasks to generate discriminate features. Sample-based triplets mainly focus on the content, while centroid-based triplet concerns the similarity lies in the cluster, which contributes for stable predictions when facing these outliers.

A.2 Pre-selection of pretrained models

In practice, we also used some rules to conduct preliminary selection before computing DBI scores. First, candidate pretrained models should achieve top performance in their original pre-text fields (*e.g.*, high Top-1 accuracy in ImageNet classification). Using this, homogeneous models can be filtered out (*e.g.*, ResNet-18 (ψ=4.36), MobileNet (ψ=4.67)). And the DBI scores also prove this attempt. Second, there should be a considerable divergence between candidate models. It ensures that each pretrained model has its own strengths (*e.g.*, building spatiotemporal relation [TimeSformer], content-aware [Swin], human emotional tendency [CLIP]). By using these two rules, we can filter the optimal models in each field that own the lowest DBI scores. This effectively reduces the cost of subsequent attempts for model combinations.

A.3 Selection of hyper-parameters

We test using different values of β in Table 9. The best results can be achieved when β is 0.2. Excessive constraints may cause the features to be too close to affecting the representation. So we set the default value of β to be 0.2 in all other experiments.

Table 8: Performances using different values of β.

β	PLCC	SRCC
0.1	0.8514	0.8230
0.2	**0.8734**	**0.8472**
0.5	0.8485	0.8190
1.0	0.8457	0.8102
2.0	0.8457	0.8085

A.4 Statistical significant over SOTA methods

We perform 10 repeat runs in each dataset using different splittings to get the mean values of PLCC and SRCC (0.8718±0.0041 of PLCC, 0.8568 ± 0.0103 of SRCC in KoNViD-1k, 0.8198± 0.0022 of PLCC, 0.8110 ± 0.0051 of SRCC in LIVE-VQC, 0.8570± 0.0043 of PLCC, 0.8578 ± 0.0029 of SRCC in YouTube-UGC). Then unequal variance t-tests are calculated to show whether these results are statistically significant, compared with the current SOTA method of STDAM. For KoNViD-1k, the t-value and degrees of freedom (DF) are 6.54 and 13.04 (reduced to 13), respectively. Using the degree of freedom value as 13 and 5% level of significance, a look at the t-value distribution table gives a value of 1.77. For YouTube-UGC, the t-value and DF is 4.81 and 11.71 (reduced to 11), respectively. Using the degree of freedom value as 11 and 5% level of significance, a look at the t-value distribution table gives a value of 1.79. Therefore, it is safe to reject the null hypothesis that there is no difference between means. And PTM-VQA is better than STDAM in statistics in KoNViD-1k and YouTube-UGC.

A.5 Statistical significant over random selection

Table 9: Performances using randomly selected pretrained models.

Models	PLCC	SRCC
ResNet152, ViT, RegNet	0.8022	0.7524
ResNet152, X3D, ir-CSN-IG	0.6221	0.6197
EfficientNet-b7, ViT, Swin Base	0.8117	0.7675
ConvNext, iBot, SlowFast	0.8230	0.7915
TimeSformer, EfficientNet-B7, MAE	0.8049	0.7777
SlowFast, MAE	0.7993	0.7743
SlowFast, TimeSformer, MAE, Swin Base	0.8214	0.7868
ViT	0.7879	0.7353
ViT, MAE, Swin Base	0.8317	0.7912
SlowFast, TimeSformer, MAE	0.8135	0.7863
avg	0.7917±0.0578	0.7583 ± 0.0492

We perform 10 runs (an extra 8) using random model selection in KoNViD-1k. Then a t-test evaluation is used to show the statistically significant DBI. Compared with random results, the t-value and DF are 10.18 and 10.27 (reduced to 10), respectively. Using the degree of freedom value as 10 and 5% level of significance, a look at the t-value distribution table gives a value of 1.79. Using the degree of freedom value as 11 and 5% level of significance, a look at the t-value distribution table gives a value of 1.81. Therefore, it is safe to reject the null hypothesis that there is no difference between means. And DBI gives better results than random selection.