CLIP-Fusion: A Spatio-Temporal Quality Metric for Frame Interpolation

Göksel Mert Çökmez[1], Yang Zhang[2], Christopher Schroers[2], Tunç Ozan Aydın[2]

[1]ETH Zürich, [2]Disney Research | Studios

mert.coekmez@alumni.ethz.ch,{yang.zhang,christopher.schroers,tunc.aydin}@disneyresearch.com

Abstract

Video frame interpolation (VFI) is an ill-posed problem, and a wide variety of methods have been proposed, ranging from more traditional computer vision strategies to the most recent developments with neural network models. Although there are many methods to interpolate video frames, quality assessment regarding the resulting artifacts from these methods remains dependent on off-the-shelf methods. Although these methods can make accurate quality predictions for many visual artifacts such as compression, blurring, and banding, their performance is mediocre for VFI artifacts due to the unique spatio-temporal qualities of such artifacts. To address this, we aim to leverage semantic feature extraction capabilities of the pre-trained visual backbone of CLIP. Specifically, we adapt its multi-scale approach to our feature extraction network and combine it with the spatio-temporal attention mechanism of the Video Swin Transformer. This allows our model to detect interpolation-related artifacts across frames and predict the relevant differential mean opinion score. Our model outperforms existing state-of-the-art quality metrics for assessing the quality of interpolated frames in both full-reference (FR) and no-reference (NR) settings.

1. Introduction

VFI methods aim to generate new frames from the existing frames of a video, which are then utilized to up- or down-sample the video. Although there are numerous algorithms to generate these new frames, the aesthetic quality of the resulting interpolated frames can vary substantially.

To address the problem of quality assessment of generated frames, there exist a multitude of general purpose video and image quality metrics. Examples of more conventional approaches include PSNR (Peak Signal to Noise Ratio) and SSIM (Structural Similarity Index Measure) [39], which compare pixel values or low-level visual patterns to perform quality assessment.

In addition to traditional methods, deep learning-based quality metrics perform well on general image [3, 9, 10, 13, 16, 18, 33, 34, 39] and video [42, 43] quality assessment. However, their performance on VFI-related distortions is limited [6] since they are not specifically designed for these distortions.

Our work aims to address this shortcoming, taking inspiration from LPVPS proposed by Hou *et al.* [14]. LPVPS extracts features from all levels of a convolutional neural network in a multi-scale manner with modified SwinIR (Swin Image Restoration) transformers [17]. The transformers then build spatio-temporal features using the outputs from different levels of the CNN. The temporally conscious approach of LPVPS inspired this work, as VFI generates unique artifacts not only in the spatial domain but also in the temporal domain.

Another critical component of this work is the zero-shot image classification capabilities of CLIP (Contrastive Language-Image Pre-training) [29]. Previous works have demonstrated its capabilities in general-purpose image [38, 50] and video [41, 43] quality assessment, where CLIP has been used to classify video frames or still images as *high quality* or *low quality*, and its predictions are in line with human judgments. In this work, the pre-trained visual backbone of CLIP is utilized, and the resulting semantic features are fed into modified Video Swin Transformers [19, 20], which are then fine-tuned on datasets [6, 14] where the dominant distortions are VFI related.

The main contributions of this paper are as follows:

- A novel quality metric combining CLIP's multi-scale feature extraction with Video Swin Transformer for improved video frame interpolation assessment.
- Captures both spatial and temporal features across multiple scales to effectively detect VFI artifacts in both FR and NR settings.

2. Related Work

This work utilizes the visual backbone of CLIP [29] as the main feature extractor. We cover CLIP's image classification and prior work on general quality metrics using

CLIP. Since our focus is on VFI artifacts, we discuss metrics tailored for such artifacts. We then describe VFI-related datasets, including BVI-VFI [6] and VFIPS [14]. Finally, we detail the state-of-the-art models on these datasets.

2.1. CLIP backbone

CLIP is an image classification framework that uses a visual backbone for extracting features from input images and a text backbone for extracting features from prompts. It computes cosine similarity between these features to determine the probability that each prompt matches the image content. CLIP is employed in this work, as its main purpose is zero-shot image classification. The zero-shot approach makes it easy to adapt to downstream vision tasks without fine-tuning, which makes it possible to deploy CLIP within a larger model with frozen weights to keep the number of learnable parameters on a manageable level.

2.2. General image and video quality assessment

Assessing the quality of VFI artifacts is a distinct problem due to its spatio-temporal nature. Therefore, examining the existing general-purpose image and video quality metrics (VQM) is necessary, as some sub-problems within this domain may already be investigated by prior works.

For FR-VQM, FAST [44] proposes a model based on the motion trajectories of pre-determined keypoints. A spatial quality score is computed using the optical flows of the selected points. RandkDVQA [8] proposes a method that uses a two-stage, ranking-based training strategy to enhance model generalization and performance, leveraging a large-scale training dataset without human-labeled ground truth. FUNQUE [37] proposed a method that uses a wavelet-domain transform and applies a contrast sensitivity function for FR video quality assessment (VQA).

LIQE [50] is an NR image quality metric using CLIP's zero-shot capabilities. It extracts scene type, dominant distortion, and perceived quality from the input image to make a blind quality assessment. SF-IQA [47] integrates image quality and image-text similarity using a Swin Transformer for feature extraction and a CLIP model for semantic similarity, merged through a score fusion module.

FAST-VQA [40], DOVER [42], and MaxVQA [43] are NR VQA models. These models were developed cumulatively, each model building upon the previous. FAST-VQA, along with its subsequent iterations, leverages spatial and temporal fragments from an input video to predict mean opinion scores. This is achieved by sampling fragments from an input frame instead of cropping or resizing the image to avoid introducing additional distortions. Spatial fragments are sampled from the same grid coordinates for every couple of frames, ensuring temporal fragmentation as well. DOVER augments this approach by adding a separate aesthetic input evaluation. The FAST-VQA backbone is preserved as the technical branch, which evaluates more technical distortions such as blur and compression artifacts. The aesthetic branch takes semantics and composition into account, focusing mainly on content for quality assessment. Then, MaxVQA builds on this by offloading aesthetic evaluation to CLIP and technical evaluation to DOVER.

2.3. Quality metrics for video frame interpolation

The unique spatio-temporal nature of VFI artifacts may reduce the general-purpose VQA model performance and may not necessarily indicate their performance on such datasets. This is evident when evaluating generic metrics on VFI datasets such as VFIPS [14] and BVI-VFI [6]. We review the best performing models on the BVI-VFI dataset, ST-GREED [25] and FRQM [49]; and the best performing model on the VFIPS dataset, LPVPS [14] to understand what enables accurate human judgment prediction for frame interpolation artifacts.

FRQM [49] is a conventional FR-VQM, designed to capture frame rate-related artifacts. The proposed method using temporal wavelet decomposition, subband comparison, and spatiotemporal pooling, FRQM estimates the relative quality of low frame rate videos compared to higher frame rate references.

ST-GREED [25] is a learning-based FR-VQM that evaluates frame rate effects using spatial and temporal generalized entropic differences, which are then mapped to quality scores via Support Vector Regressor (SVR) trained on the LIVE-YT-HFR [25] dataset. LPVPS [14] is a learning-based FR-VQA model that consists of a five-level pyramid network for feature extraction. The extracted features from each level are forwarded to the spatio-temporal module. In the spatio-temporal module, features from the reference video and distorted video are merged. This is accomplished by computing the difference per frame between all reference and distorted features and concatenating the resulting difference with reference and distorted features in the channel dimension. Finally, LPVPS employs a SwinIR transformer [17] at every level of the pyramid network to capture temporal features. The resulting features of all levels are then averaged to produce the final prediction.

3. Method

As illustrated in Fig. 1, our work is built around a modified pre-trained CLIP visual backbone [29] for spatial feature extraction. We modify the visual backbone so that it extracts features not only from the final layer of the backbone network, but also features from four other layers at different scales. This multi-scale approach is inspired by the feature extraction network of LPVPS [14], which helps it to yield good performance on VFI-related artifacts.

The extracted features are then fed into the multi-scale spatio-temporal module. There, we merge features from

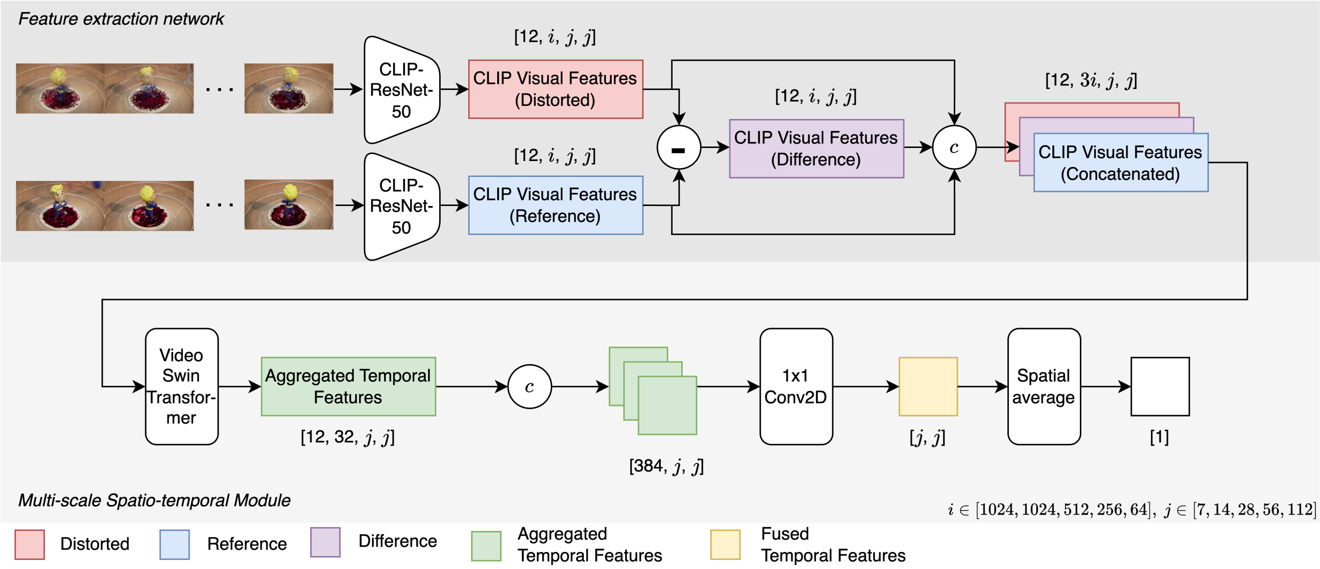

Feature extraction network
CLIP-ResNet-50
CLIP-ResNet-50
[12, i, j, j]
CLIP Visual Features (Distorted)
[12, i, j, j]
CLIP Visual Features (Reference)
[12, i, j, j]
CLIP Visual Features (Difference)
c
[12, 3i, j, j]
CLIP Visual Features (Concatenated)
Video Swin Transformer
Aggregated Temporal Features
[12, 32, j, j]
c
[384, j, j]
1x1 Conv2D
[j, j]
Spatial average
[1]
Multi-scale Spatio-temporal Module
$i \in [1024, 1024, 512, 256, 64]$, $j \in [7, 14, 28, 56, 112]$
Distorted
Reference
Difference
Aggregated Temporal Features
Fused Temporal Features

Figure 1. Full-reference (FR) architecture of our model.

the reference video and the distorted video, and use Video Swin Transformers [19, 20] to compute cross-frame attention. The resulting spatio-temporal features of each level are used to compute the final DMOS. We explain the processes for each module in detail below.

3.1. Full-reference (FR) metric setup

3.1.1 CLIP features

In the current release of CLIP, there are numerous pre-trained visual backbones available. One of these pre-trained backbones is the modified ResNet-50 [12]. Although it is not the highest performing visual backbone in the CLIP paper [29], it should be noted that these benchmarks are for image classification, which is the original purpose of CLIP. However, in [38], Wang *et al*. demonstrates that CLIP with ResNet-50 backbone produces the highest performance on quality assessment benchmarks. Thus, we also choose to employ ResNet-50 backbone for CLIP in our model. In its normal use case, CLIP is a language-supervised model, meaning that one of the strengths of the pre-trained release of CLIP we use in our feature extraction network is zero and few-shot performance. For this reason, unlike LPVPS, we keep the weights of our feature extraction network frozen during the training process.

To take advantage of the success of the multi-scale approach of LPVPS [14], we further modify CLIP-ResNet-50 to capture multiscale spatial features so that it extracts features from five layers from different depths of ResNet-50.

Specifically, we extract features from 12 consecutive frames by inputting each frame individually into our feature extraction network, as expressed in Eq. (1),

$$\mathbf{F}_{\{\text{ref}\,—\,\text{dist}\},l} = f_{\text{CLIP},l}\left(\mathbf{I}_{\{\text{ref}\,—\,\text{dist}\}}\right) \tag{1}$$

where $\mathbf{F}_{\text{ref},l}$ and $\mathbf{F}_{\text{dist},l}$ indicate the extracted reference and distorted video features from the corresponding level l of the CLIP feature extraction network. $\mathbf{I}_{\text{ref}}$ and $\mathbf{I}_{\text{dist}}$ denote the reference and distorted videos, respectively. $f_{\text{CLIP},l}(\cdot)$ refers to the operation at the corresponding level l of the CLIP extraction network, which is a mapping in the form

$$f_{\text{CLIP},l} : \mathbb{R}^{B\times 12\times 3\times 224\times 224} \mapsto \mathbb{R}^{B\times 12\times i\times j\times j},$$

where B denotes the input batch size, $i \in [1024, 1024, 512, 256, 64]$, and $j \in [7, 14, 28, 56, 112]$. We keep the weights of the extraction network frozen. Note that this process is repeated for both the distorted and reference frames.

After extracting the features from the distorted and reference frames, we fuse the distorted frame features $\mathbf{F}_{\text{dist},l}$ and the reference frames features $\mathbf{F}_{\text{dist},l}$ accordingly in each level. First, we normalize our extracted features $\mathbf{F}_{\text{dist},l}$ and $\mathbf{F}_{\text{ref},l}$ across frames to further highlight temporal features. Then we compute the element-wise absolute difference $\mathbf{F}_{\text{diff},l}$ between the reference and distorted frames features. This can be represented as the following in Eq. (2),

$$\mathbf{F}_{\text{diff},l} = abs(f_{\text{norm}}(\mathbf{F}_{\text{ref},l}) - f_{\text{norm}}(\mathbf{F}_{\text{dist},l})) \tag{2}$$

In Eq. (2), $\mathbf{F}_{\text{ref},l}$ and $\mathbf{F}_{\text{dist},l}$ represent the CLIP features extracted from the reference and distorted videos, respectively. $f_{\text{norm}}(\cdot)$ is the normalization operation across frames, $abs(\cdot)$ is the element-wise absolute difference operator, and

$\mathbf{F}_{\text{diff},l}$ is the element-wise absolute difference of the normalized reference and distorted features.

The resulting difference tensor is then concatenated with the reference features $\mathbf{F}_{\text{ref},l}$ and the distorted features $\mathbf{F}_{\text{dist},l}$ in the channel dimension. These operations can be represented as

$$\mathbf{F}_{\text{cat},l} = f_{\text{cat}}(f_{\text{norm}}(\mathbf{F}_{\text{ref},l}), \mathbf{F}_{\text{diff},l}, f_{\text{norm}}(\mathbf{F}_{\text{dist},l})), \quad (3)$$

where $f_{\text{cat}}(\cdot)$ represents the concatenation operation for the extracted features along the channel dimension, resulting in a shape $\mathbb{R}^{B\times 12\times 3i\times j\times j}$. Note that these operations are repeated for every level of the feature extraction network. The extracted features are input into the Video Swin Transformer [19, 20] to calculate spatio-temporal features.

3.1.2 Multi-scale spatio-temporal module

The extracted features from the CLIP-ResNet-50 of the reference and distorted frames are inputted into the Video Swin Transformers after each level of the feature extraction network individually. Video Swin Transformers compute attention inside sliding windows across three dimensions that are height, width, and video frames. This is in stark contrast to LPVPS, where temporal features from input frames are concatenated in the channel dimension, and sliding windows are only applied across the height and width, forgoing the sliding-window self-attention computation in the temporal dimension. Our method of computing attention ensures that temporal features are also represented in our latent features in addition to the spatial features. We denote this operation as

$$\mathbf{F}_{\text{VSwin},l} = f_{\text{VSwin},l}(\mathbf{F}_{\text{cat},l}), \quad (4)$$

where, $f_{\text{VSwin},l}(\cdot)$ represents the operation of the Video Swin Transformers on concatenated features $\mathbf{F}_{\text{cat},l}$ on every level l. The output features from individual levels of the Video Swin Transformer have the shape $\mathbb{R}^{B\times 12\times 32\times j\times j}$, and are then concatenated in the channel dimensions and passed through a 1×1 convolution layer, which fuses features from all channels to a singular channel, and all elements of the final single channel tensor are averaged. This is repeated at every level and can be represented as the following in Eq. (5),

$$\mathbf{F}_{\text{final},l} = f_{\text{emb}}(f_{\text{reshape}}(\mathbf{F}_{\text{VSwin},l})) \quad (5)$$

where, $f_{\text{emb},l}(\cdot)$ represents the 1×1 convolution operation to reduce the number of channels and $f_{\text{reshape}}(\cdot)$ represents the reshaping operation to shape $\mathbb{R}^{B\times 384\times j\times j}$, which merges the frame and channel dimensions of the tensor, effectively merging the spatial and temporal features. As the last step, the final DMOS is calculated by adding the average of the features of shape $\mathbb{R}^{B\times j\times j}$ from all levels as shown in Eq. (6),

$$dmos = \sum_{l=0}^{L} \frac{1}{N} \sum_{m,n=0}^{N} \mathbf{F}_{\text{final},l}^{(m,n)} \quad (6)$$

where, m and n denote the row and column indices for entries of the final features, $\mathbf{F}_{\text{final},l}$ in each level l. N represents the total number of entries at the level l of the tensor $\mathbf{F}_{\text{final},l}$.

3.2. No-reference (NR) metric setup

To evaluate our model against existing NR metrics, our model can be modified to function in an NR manner, as explained in the following sections.

3.2.1 Feature extraction network

Similarly to the FR approach, our feature extraction network consists of a modified CLIP visual ResNet-50 backbone. In the FR setting, the ground-truth and distorted videos are fed separately to the feature extraction network. As there is no ground-truth video available for the NR setting, multi-scale features from the distorted video are extracted uniquely and sent to the multi-scale spatio-temporal module for spatio-temporal feature computation.

3.2.2 Multi-scale satio-temporal module

The lack of ground-truth video means that it is not possible to compute the element-wise difference between ground-truth and distorted video features. As a result, instead of calculating the difference between ground truth and distorted videos and concatenating the resulting features in the channel dimension, this step is completely bypassed. Distorted video features coming from the video extraction network are fed directly into the Video Swin Transformer for spatio-temporal attention computation. The remaining steps for combining features across frames and MOS predictions with features from all levels remain identical to the FR setting.

4. Datasets

For training and testing, we selected datasets that predominantly feature VFI distortion. This study relies on two primary datasets: VFIPS [14] and BVI-VFI [6]. VFIPS includes 12 frame sequences of a ground-truth video alongside two distorted versions generated by VFI algorithms. The scores reflect perceptual judgments indicating which video has higher quality, akin to a binary classification task. BVI-VFI comprises videos of 36 subjects at two resolutions (12 videos at 960x540 and 24 videos at 1920x1080). Each ground truth video has versions with three frame rates (30 FPS, 60 FPS, 120 FPS) and five different interpolation algorithms, totaling 540 videos. DMOS and MOS values are processed using the P.910 subject screening procedure [36] to minimize participant bias.

5. Experiments and results

We utilize three different metrics to evaluate the performance of our model: Pearson linear correlation coefficient (PLCC), Spearman rank correlation coefficient (SRCC), and Kendall rank correlation coefficient (KRCC).

The evaluation procedure for our work consists of six total experiments in three sets. The first set consists of two experiments designed to compare our work with LPVPS [14], which was the main inspiration for our work. Thus, these first two experiments recreate the methodology used in [14], where the reported results are the average of local PLCC, SRCC and KRCC scores, namely ***Local scores***. We calculate these local scores by computing the PLCC, SRCC, and KRCC per reference video and reporting the mean PLCC, SRCC, and KRCC across reference videos.

The second set of experiments consists of two experiments designed to benchmark our model against other general-purpose VQA methods, as surveyed in the BVI-VFI paper [6]. For this reason, we employ the methodology detailed in [6], where the PLCC, SRCC, and KRCC scores are calculated on all data points (540 distorted videos), namely ***Global scores***. These experiments compare videos across different content types and therefore *Global scores* are more indicative of the general performance of the metrics tested.

The DMOS values have been provided in the BVI-VFI dataset since the FR setup. However, for the NR setup of the metrics evaluation, we conduct the third set of experiments that covers cross-validation on the BVI-VFI dataset using MOS. This experiment benchmarks our model against other NR models covered in [6].

5.1. Experiment setup

For experiments where we train the models on VFIPS, the LPVPS results were produced by the publicly available pre-trained LPVPS model. For experiments in which we perform training on BVI-VFI, we train LPVPS from scratch with the default training settings in its source paper [14]. It is trained over 20 epochs, with the AdamW [22] optimizer. We set the learning rate at $1 \cdot 10^{-4}$ and β_1 to 0.5. We set all other hyperparameters to their default values.

For all experiments, we train our model for only 5 epochs due to its fast convergence, using AdamW optimizer with a learning rate of $5 \cdot 10^{-5}$. Experiments are conducted on a single NVIDIA GeForce RTX 4090 GPU with a batch size of 8, the same for LPVPS and our model.

For experiments in which we train the model on the BVI-VFI dataset, we perform a cross-validation on the BVI-VFI dataset. We used 80% of the 36 source videos in the BVI-VFI dataset for training, while we use the remaining 20% for testing. We reshuffle the dataset splitting process 20 times. To maintain reproducibility, we save all 20 splits and reuse them for all experiments in which the model is trained on the BVI-VFI dataset. On average, every subject appears in the training set 15.56 times with a standard deviation of 1.95 compared to an average of 4.44 times in the test set with a standard deviation of 1.95. This is in line with the expected number of occurrences of 16 times in the training set and 4 times in the test set, meaning that there is no significant imbalance in the training and testing datasets.

Table 1. Average performance of the model benchmarked against LPVPS when training on the VFIPS and evaluating on the BVI-VFI dataset.

	VFIPS validation	BVI-VFI		
Model	2AFC	PLCC	SRCC	KRCC
LPVPS	0.81	0.70	0.63	0.55
Ours	**0.82**	**0.78**	**0.68**	**0.58**

5.2. Benchmarks on local scores

5.2.1 Trained on VFIPS and tested on BVI-VFI

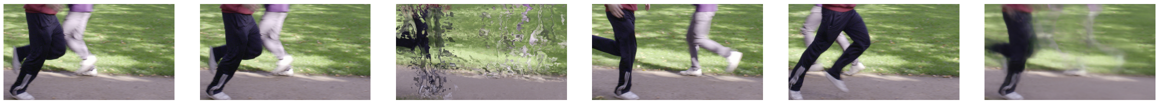
To evaluate the cross-dataset performance, we train our model on the VFIPS dataset and evaluate it on the BVI-VFI dataset. As mentioned in the previous section, the local PLCC, SRCC and KRCC scores are computed over the five different distorted versions of each ground truth, and the average local score is reported. As shown in Tab. 1, our model outperforms LPVPS in all available metrics. For completeness, we also report 2AFC scores on the VFIPS validation dataset. Tab. 2 present visual examples of the BVI-VFI dataset accompanied by human rankings for each distorted video. Our predictions demonstrate greater consistency with human evaluations compared to LPVPS.

5.2.2 BVI-VFI dataset cross-validation

As mentioned in Sec. 4, the VFIPS dataset is a binary classification dataset, with no MOS or DMOS provided. Since the BVI-VFI dataset, which is used for evaluation in this work, provides DMOS, this experiment will perform cross-validation on the BVI-VFI dataset to observe how the two models compare in the presence of actual quality scores. Therefore, 20-fold cross-validation splits of the BVI-VFI dataset are used as training and test sets for this experiment, as described in Sec. 5.1.

Similar to the previous experiment, the reported results are average local PLCC, SRCC, and KRCC scores. Since there are 20 splits, the final reported result is the mean and standard deviation of the average local PLCC, SRCC and KRCC of each split. The results of the experiment can be observed in Tab. 3. Compared to the results of Tab. 1, these results indicate an increase in performance for both models

Table 2. Quality ranking predictions by our model and LPVPS versus human for *BVI_joggers_3840x2160_30fps* and *LIVEHFR_3Runners-_960x540_30fps* videos. The leftmost image is the pristine ground truth image.



	Average	DVF [21]	QVI [46]	Repeat	ST-MFNet [5]
Human	3rd	5th	2nd	4th	1st
LPVPS	3rd	5th	1st	4th	2nd
Ours	3rd	5th	2nd	4th	1st

	Average	DVF	QVI	Repeat	ST-MFNet
Human	3rd	5th	1st	4th	2nd
LPVPS	4th	5th	2nd	3rd	1st
Ours	3rd	5th	1st	4th	2nd

when trained on DMOS values. Although our model outperforms LPVPS in all available metrics, its performance remains comparable to LPVPS, under the *Local score* evaluation method.

Table 3. Average performance of the model benchmarked against LPVPS when cross-validating on the BVI-VFI dataset.

Model	PLCC	SRCC	KRCC
LPVPS	0.78 (0.08)	0.72 (0.09)	**0.63 (0.09)**
Ours	**0.80 (0.06)**	**0.73 (0.06)**	**0.63 (0.05)**

5.3. Benchmarks on global scores

In previous experiments, our main focus was on comparing our model with LPVPS [14]. The first set of experiments covered in the previous section demonstrated that our model outperforms LPVPS. This set of experiments aims to demonstrate the performance of our model on *Global* PLCC and SRCC scores by comparing them with readily available results in [6], including conventional and deep learning-based metrics, these are PSNR, GMSD [48], FAST [44], SpEED [1], C3DVQA [45], FRQM [49], FovVideoVDP [26], ST-GREED [25] and FloLPIPS [4].

In this experiment, we follow the same evaluation method as utilized in [6]. We cross-validate our model on the BVI-VFI dataset, as specified in Sec. 5.1, and report the median global PLCC and SRCC scores along with their standard deviations. The results in Tab. 4 demonstrate that, barring its performance over non-DL based interpolation methods such as frame averaging and frame repeating, our model consistently ranks as the best or second best model in every category. Regarding overall performance, it is the best model with consistently high PLCC and SRCC scores and low standard deviation, indicating robust performance.

5.4. No-reference (NR) evaluation with MOS

As mentioned in Sec. 3.2, our model is also capable of running in an NR setting. To demonstrate the NR quality assessment capabilities, we perform cross-validation on the BVI-VFI dataset to train our model on MOS values, as described in Sec. 5.1. We compare the NR version of our model with other NR models tested in [6], including BRISQUE [32], ChipQA [7], VIDEVAL [35], deepIQA_NR [2], NIQE [28], VIIDEO [27], FastVQA [40], VBLIINDS [31], TLVQM [15], and FAVER [51]. The metrics reported are the median values of PLCC and SRCC across the test splits for each model, along with their standard deviations.

Except in the non-DL category, our model consistently performs the best among NR models, with a significant increase in performance, as shown in Tab. 5. It should be noted that it exhibits performance comparable to the FR models in Tab. 4, indicating that our architecture benefits from the zero-shot capabilities of CLIP.

5.5. Ablation studies

5.5.1 Feature extraction network

We investigate the impact of using the pre-trained CLIP model as our feature extraction network. Consequently, we replace our CLIP visual backbone with a five-level pyramid convolutional network, which serves as the feature extractor network in LPVPS. Since we train this network from scratch, we train our model for 20 epochs and keep the rest of the training parameters identical.

Table 4. Cross validation tests over all the DMOS values in BVI-VFI dataset. Best and second best models are marked accordingly.

	30fps		60fps		120fps		non-DL		DL		Overall	
Model	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
FAST	0.60 (0.13)	0.50 (0.12)	0.69 (0.17)	0.74 (0.08)	0.77 (0.19)	0.72 (0.14)	0.50 (0.12)	0.47 (0.12)	0.79 (0.07)	0.77 (0.06)	0.66 (0.08)	0.71 (0.07)
PSNR	0.60 (0.11)	0.51 (0.11)	0.66 (0.10)	0.69 (0.10)	0.63 (0.13)	0.63 (0.14)	0.54 (0.12)	0.46 (0.13)	0.72 (0.09)	0.71 (0.07)	0.62 (0.07)	0.65 (0.08)
FRQM	0.57 (0.12)	0.45 (0.13)	0.64 (0.10)	0.66 (0.11)	0.63 (0.14)	0.60 (0.15)	0.84 (0.06)	0.81 (0.07)	0.45 (0.07)	0.49 (0.06)	0.50 (0.06)	0.58 (0.06)
FovVideoVDP	0.55 (0.12)	0.46 (0.11)	0.62 (0.10)	0.66 (0.09)	0.59 (0.14)	0.60 (0.16)	0.63 (0.10)	0.57 (0.12)	0.64 (0.07)	0.66 (0.06)	0.59 (0.06)	0.64 (0.06)
FloLPIPS	0.57 (0.11)	0.47 (0.12)	0.62 (0.11)	0.59 (0.11)	0.69 (0.13)	0.60 (0.13)	0.53 (0.12)	0.47 (0.11)	0.67 (0.08)	0.66 (0.07)	0.61 (0.08)	0.61 (0.07)
GMSD	0.61 (0.11)	0.51 (0.12)	0.64 (0.10)	0.67 (0.11)	0.63 (0.12)	0.63 (0.14)	0.52 (0.11)	0.42 (0.14)	0.72 (0.09)	0.69 (0.07)	0.61 (0.08)	0.64 (0.08)
C3DVQA	0.43 (0.17)	0.28 (0.14)	0.51 (0.21)	0.57 (0.11)	0.50 (0.25)	0.64 (0.13)	0.46 (0.18)	0.40 (0.12)	0.56 (0.14)	0.59 (0.09)	0.48 (0.10)	0.55 (0.08)
SpEED	0.39 (0.19)	0.50 (0.12)	0.51 (0.21)	0.68 (0.09)	0.49 (0.19)	0.64 (0.14)	0.31 (0.15)	0.45 (0.11)	0.61 (0.17)	0.69 (0.06)	0.49 (0.25)	0.65 (0.07)
ST-GREED	0.60 (0.12)	0.53 (0.14)	0.72 (0.12)	0.71 (0.12)	0.71 (0.15)	0.61 (0.17)	0.46 (0.11)	0.41 (0.12)	0.77 (0.09)	0.70 (0.10)	0.66 (0.11)	0.64 (0.10)
LPVPS	0.45 (0.12)	0.61 (0.13)	0.63 (0.11)	0.65 (0.09)	0.65 (0.14)	0.58 (0.11)	0.32 (0.12)	0.36 (0.11)	0.61 (0.11)	0.66 (0.09)	0.53 (0.11)	0.61 (0.09)
Ours	0.67 (0.11)	0.63 (0.10)	0.76 (0.08)	0.76 (0.07)	0.75 (0.07)	0.67 (0.12)	0.31 (0.09)	0.34 (0.10)	0.77 (0.07)	0.76 (0.07)	0.73 (0.08)	0.72 (0.07)

Table 5. Cross validation tests over all the MOS values in BVI-VFI dataset. Best and second best models are marked accordingly.

	30fps		60fps		120fps		non-DL		DL		Overall	
Model	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
BRISQUE	0.29 (0.12)	0.05 (0.19)	0.30 (0.11)	0.05 (0.19)	0.25 (0.11)	0.11 (0.20)	0.28 (0.12)	0.16 (0.13)	0.26 (0.11)	0.08 (0.15)	0.21 (0.08)	0.02 (0.12)
ChipQA	0.54 (0.13)	0.49 (0.16)	0.57 (0.14)	0.57 (0.17)	0.56 (0.15)	0.43 (0.18)	0.35 (0.12)	0.27 (0.15)	0.60 (0.14)	0.56 (0.15)	0.50 (0.12)	0.47 (0.13)
VIDEVAL	0.58 (0.11)	0.55 (0.15)	0.55 (0.12)	0.51 (0.13)	0.61 (0.13)	0.53 (0.20)	0.51 (0.12)	0.47 (0.13)	0.56 (0.13)	0.51 (0.14)	0.49 (0.12)	0.45 (0.13)
deepIQA_NR	0.32 (0.12)	0.19 (0.17)	0.28 (0.10)	0.16 (0.18)	0.22 (0.10)	0.07 (0.19)	0.22 (0.10)	0.03 (0.13)	0.26 (0.10)	0.15 (0.12)	0.20 (0.08)	0.09 (0.11)
NIQE	0.30 (0.12)	0.02 (0.19)	0.30 (0.11)	0.05 (0.18)	0.28 (0.09)	0.03 (0.22)	0.32 (0.11)	0.19 (0.13)	0.27 (0.09)	0.02 (0.13)	0.22 (0.08)	0.04 (0.12)
VIIDEO	0.27 (0.14)	0.10 (0.13)	0.32 (0.15)	0.11 (0.12)	0.37 (0.14)	0.20 (0.14)	0.22 (0.08)	0.05 (0.05)	0.36 (0.11)	0.29 (0.14)	0.28 (0.10)	0.20 (0.08)
FastVQA	0.36 (0.16)	0.11 (0.17)	0.33 (0.12)	0.28 (0.15)	0.38 (0.14)	0.26 (0.18)	0.43 (0.10)	0.38 (0.10)	0.27 (0.10)	0.17 (0.12)	0.28 (0.09)	0.24 (0.11)
VBLIINDS	0.58 (0.11)	0.52 (0.13)	0.67 (0.12)	0.60 (0.13)	0.67 (0.14)	0.49 (0.17)	0.44 (0.11)	0.40 (0.12)	0.63 (0.11)	0.59 (0.11)	0.48 (0.11)	0.51 (0.11)
TLVQM	0.49 (0.14)	0.40 (0.18)	0.54 (0.13)	0.49 (0.15)	0.67 (0.15)	0.53 (0.19)	0.47 (0.13)	0.42 (0.14)	0.63 (0.11)	0.59 (0.12)	0.52 (0.12)	0.48 (0.12)
FAVER	0.50 (0.12)	0.44 (0.15)	0.50 (0.14)	0.41 (0.16)	0.54 (0.17)	0.39 (0.17)	0.66 (0.11)	0.62 (0.13)	0.58 (0.13)	0.52 (0.14)	0.52 (0.12)	0.49 (0.13)
Ours (no-ref)	0.72 (0.06)	0.74 (0.08)	0.77 (0.07)	0.76 (0.10)	0.75 (0.08)	0.55 (0.18)	0.33 (0.08)	0.34 (0.09)	0.75 (0.06)	0.72 (0.08)	0.68 (0.05)	0.64 (0.07)

5.5.2 Vision transformers

Another vital component of our model is the Video Swin Transformers at every level of our multi-scale spatio-temporal module. It replaces SwinIR [17] in LPVPS, as its 3D windows evaluate image patches across multiple frames, taking into account the features in the temporal domain. This contrasts with SwinIR in LPVPS, where multiple frames are concatenated before being inputted to SwinIR to model the temporal features.

Table 6. Results of our ablation studies.

	VFIPS validation	BVI-VFI		
Model	2AFC	PLCC	SRCC	KRCC
LPVPS	0.81	0.70	0.63	0.55
Ours w/ five-level pyramid	0.82	0.54	0.60	0.51
Ours w/ SwinIR	**0.83**	0.52	0.51	0.44
Ours w/ single layer CLIP	0.81	0.65	0.44	0.36
Ours w/ optical flow	0.82	0.75	0.61	0.52
Ours	0.82	**0.78**	**0.68**	**0.58**

5.5.3 Multi-scale CLIP features

Our model leverages high- and low-level semantic features from multiple levels of CLIP. To verify the performance contribution of our multi-scale architecture, we replace our multi-scale features with features solely from the last level of the CLIP visual backbone.

5.5.4 Optical flow

The experiments performed on the cross-validation of the BVI-VFI dataset demonstrate that FAST [44] is the second-best performing model after ours. FAST utilizes optical flow to assess the aesthetic quality of consecutive frames. Consequently, we also employ optical flow by implementing SPyNet [30] in our feature extraction network. We simply concatenate optical flow features to our CLIP features for every frame in our feature extraction network to evaluate the impact of optical flow on our model's overall performance.

We trained the ablation study models on the VFIPS dataset and evaluated them on the BVI-VFI dataset, the results in Tab. 6 demonstrate the performance of the variants. Both the SwinIR model and the five-level pyramid model are outperformed by our model. The multi-scale approach introduced in LPVPS yields a notable performance increase, as our model with multi-scale CLIP yields approximately 41% higher scores for PLCC, SRCC, and KRCC compared to our model with a single-layer CLIP feature. Finally, we observe that the inclusion of optical flow does not improve performance within our architecture, as it performs slightly worse than our model despite having a larger feature extraction network.

5.6. Discussion

Based on the conducted experiments, it can be concluded that our proposed model offers a significant performance improvement over existing state-of-the-art FR-VQMs for frame interpolation artifacts, over both same-subject and cross-subject comparisons.

Similarly to its FR performance, the NR setup of our model remains competitive, yielding a respectable performance among state-of-the-art FR metrics. In addition to surpassing all NR metrics, the NR setup of our model even manages to outperform all FR metrics barring FAST [44] in the overall category.

Notably, FRQM [49] performs the best in our weakest category, the *non-DL* category, as shown in Table 4. Although it is a conventional model, the optimal weighting values for the subband combination module are obtained through cross-validation on the BVI-HFR dataset [23, 24]. In this training process, the low frame rate input videos are upsampled to match the frame rate of the reference high frame rate videos using a nearest-neighbor interpolation [11] algorithm, which is similar to frame repeating. Our intuition is that this weight optimization process may be a key reason why FRQM performs the best in the *non-DL* category, while showing lower performance in the *DL* and other categories on the BVI-VFI dataset.

The inference time of a single frame from the BVI-VFI dataset is higher for our model (~ 0.243 seconds) compared to that of LPVPS (~ 0.157 seconds), as CLIP feature extraction is performed on the fly. Although the inference time is higher, the pre-trained CLIP model is not further fine-tuned and its zero-shot capabilities enable our model to achieve the reported performance in fewer epochs compared to LPVPS, necessitating only five epochs.

6. Conclusion and outlook

In this work, we introduce a novel model for the quality assessment of videos with frame interpolation artifacts. In both FR and NR settings, our proposed model outperforms state-of-the-art VQA metrics while requiring few training epochs. As another exemplar of CLIP's applicability in quality assessment tasks, our proposed network architecture and training pipeline demonstrate that CLIP features can be leveraged to extract semantic features across the temporal domain in addition to the spatial domain.

Our experiments show that the performance of our model remains dependent on the available training datasets. Our training pipeline benefits from the availability of MOS/DMOS values compared to the case where it only has access to the relative quality ranking of input videos. The reason for this performance differential and the effects of readily available MOS/DMOS values would pose an interesting topic for further research in the context of quality assessment.

The future work would encompass the utilization of text features from the CLIP model, as it also possesses a text encoder that was not employed in this work. Although its performance remains dependent on input prompts, future research could focus on identifying the optimal prompts for quality assessment of VFI. Another direction would be to forgo this *prompt engineering* approach for learnable prompts, where the model weights and optimal prompt could be learned in an alternating fashion.

References

[1] Christos George Bampis, Praful Gupta, Rajiv Soundararajan, and Alan C. Bovik. Speed-qa: Spatial efficient entropic differencing for image and video quality. *IEEE Signal Process. Lett.*, 24(9):1333–1337, 2017. 6

[2] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Trans. Image Process.*, 27(1):206–219, 2018. 6

[3] Alexandre G. Ciancio, André Luiz N. Targino da Costa, Eduardo A. B. da Silva, Amir Said, Ramin Samadani, and Pere Obrador. No-reference blur assessment of digital pictures based on multifeature classifiers. *IEEE Trans. Image Process.*, 20(1):64–75, 2011. 1

[4] Duolikun Danier, Fan Zhang, and David Bull. Flolpips: A bespoke video quality metric for frame interpolation. In *Picture Coding Symposium, PCS 2022, San Jose, CA, USA, December 7-9, 2022*, pages 283–287. IEEE, 2022. 6

[5] Duolikun Danier, Fan Zhang, and David R. Bull. St-mfnet: A spatio-temporal multi-flow network for frame interpolation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 3511–3521. IEEE, 2022. 6

[6] Duolikun Danier, Fan Zhang, and David R. Bull. BVI-VFI: A video quality database for video frame interpolation. *IEEE Trans. Image Process.*, 32:6004–6019, 2023. 1, 2, 4, 5, 6

[7] Joshua Peter Ebenezer, Zaixi Shang, Yongjun Wu, Hai Wei, Sriram Sethuraman, and Alan C. Bovik. Chipqa: No-reference video quality prediction via space-time chips. *IEEE Trans. Image Process.*, 30:8059–8074, 2021. 6

[8] Chen Feng, Duolikun Danier, Fan Zhang, and David Bull. Rankdvqa: Deep vqa based on ranking-inspired hybrid training. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1648–1658, 2024. 2

[9] D. Ghadiyaram and A. C. Bovik. Live in the wild image quality challenge database. Online: `http://live.ece.utexas.edu/research/ChallengeDB/index.html`, 2015. 1

[10] Deepti Ghadiyaram and Alan C. Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Trans. Image Process.*, 25(1):372–387, 2016. 1

[11] Rafael C. González and Richard E. Woods. *Digital image processing*. Addison-Wesley, 1992. 8

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 3

[13] Vlad Hosu, Hanhe Lin, Tamás Szirányi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Trans. Image Process.*, 29:4041–4056, 2020. 1

[14] Qiqi Hou, Abhijay Ghildyal, and Feng Liu. A perceptual quality metric for video frame interpolation. In *European Conference on Computer Vision*, pages 234–253. Springer, 2022. 1, 2, 3, 4, 5, 6

[15] Jari Korhonen. Two-level approach for no-reference consumer video quality assessment. *IEEE Trans. Image Process.*, 28(12):5923–5938, 2019. 6

[16] Eric C. Larson and Damon M. Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *J. Electronic Imaging*, 19(1):011006, 2010. 1

[17] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021, Montreal, BC, Canada, October 11-17, 2021*, pages 1833–1844. IEEE, 2021. 1, 2, 7

[18] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted IQA database. In *11th International Conference on Quality of Multimedia Experience QoMEX 2019, Berlin, Germany, June 5-7, 2019*, pages 1–3. IEEE, 2019. 1

[19] 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 *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9992–10002. IEEE, 2021. 1, 3, 4

[20] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 3192–3201. IEEE, 2022. 1, 3, 4

[21] Ziwei Liu, Raymond A. Yeh, Xiaoou Tang, Yiming Liu, and Aseem Agarwala. Video frame synthesis using deep voxel flow. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 4473–4481. IEEE Computer Society, 2017. 6

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 5

[23] Alex Mackin, Fan Zhang, and David R Bull. A study of subjective video quality at various frame rates. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 3407–3411. IEEE, 2015. 8

[24] Alex Mackin, Fan Zhang, and David R Bull. A study of high frame rate video formats. *IEEE Transactions on Multimedia*, 21(6):1499–1512, 2018. 8

[25] Pavan C. Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C. Bovik. ST-GREED: space-time generalized entropic differences for frame rate dependent video quality prediction. *IEEE Trans. Image Process.*, 30:7446–7457, 2021. 2, 6

[26] Rafal K. Mantiuk, Gyorgy Denes, Alexandre Chapiro, Anton Kaplanyan, Gizem Rufo, Romain Bachy, Trisha Lian, and Anjul Patney. Fovvideovdp: a visible difference predictor for wide field-of-view video. *ACM Trans. Graph.*, 40(4):49:1–49:19, 2021. 6

[27] Anish Mittal, Michele A. Saad, and Alan C. Bovik. A completely blind video integrity oracle. *IEEE Trans. Image Process.*, 25(1):289–300, 2016. 6

[28] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a ”completely blind” image quality analyzer. *IEEE Signal Process. Lett.*, 20(3):209–212, 2013. 6

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3

[30] Anurag Ranjan and Michael J. Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 8

[31] Michele A. Saad, Alan C. Bovik, and Christophe Charrier. Blind prediction of natural video quality. *IEEE Trans. Image Process.*, 23(3):1352–1365, 2014. 6

[32] Wenting Shao and Xuanqin Mou. No-reference image quality assessment based on edge pattern feature in the spatial domain. *IEEE Access*, 9:133170–133184, 2021. 6

[33] Hamid R. Sheikh, Muhammad F. Sabir, and Alan C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Trans. Image Process.*, 15(11):3440–3451, 2006. 1

[34] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik. Live image quality assessment database release 2. *http://live.ece.utexas.edu/research/quality*, 2012. 1

[35] 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, 2021. 6

[36] International Telecommunication Union. Recommendation itu-t p. 910. subjective video quality assessment methods for multimedia applications. Technical report, 2022. 5

[37] Abhinau K Venkataramanan, Cosmin Stejerean, Ioannis Katsavounidis, and Alan C Bovik. One transform to compute them all: Efficient fusion-based full-reference video quality assessment. *IEEE Transactions on Image Processing*, 2023. 2

[38] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2555–2563, 2023. 1, 3

[39] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004. 1

[40] Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Fastvqa: Efficient end-to-end video quality assessment with fragment sampling. In *Proceedings of European Conference of Computer Vision (ECCV)*, 2022. 2, 6

[41] Haoning Wu, Liang Liao, Jingwen Hou, Chaofeng Chen, Erli Zhang, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring opinion-unaware video quality assessment with semantic affinity criterion. In *IEEE International Conference on Multimedia and Expo, ICME 2023, Brisbane, Australia, July 10-14, 2023*, pages 366–371. IEEE, 2023. 1

[42] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *ICCV*, 2023. 1, 2

[43] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Towards explainable in-the-wild video quality assessment: a database and a language-prompted approach. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1045–1054, 2023. 1, 2

[44] Jinjian Wu, Yongxu Liu, Weisheng Dong, Guangming Shi, and Weisi Lin. Quality assessment for video with degradation along salient trajectories. *IEEE Trans. Multim.*, 21(11):2738–2749, 2019. 2, 6, 8

[45] Munan Xu, Junming Chen, Haiqiang Wang, Shan Liu, Ge Li, and Zhiqiang Bai. C3DVQA: full-reference video quality assessment with 3d convolutional neural network. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pages 4447–4451. IEEE, 2020. 6

[46] Xiangyu Xu, Li Siyao, Wenxiu Sun, Qian Yin, and Ming-Hsuan Yang. Quadratic video interpolation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 1645–1654, 2019. 6

[47] Zihao Yu, Fengbin Guan, Yiting Lu, Xin Li, and Zhibo Chen. Sf-iqa: Quality and similarity integration for ai generated image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6692–6701, 2024. 2

[48] Bo Zhang, Pedro V. Sander, and Amine Bermak. Gradient magnitude similarity deviation on multiple scales for color image quality assessment. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 1253–1257. IEEE, 2017. 6

[49] Fan Zhang, Alex Mackin, and David R Bull. A frame rate dependent video quality metric based on temporal wavelet decomposition and spatiotemporal pooling. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 300–304. IEEE, 2017. 2, 6, 8

[50] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 14071–14081. IEEE, 2023. 1, 2

[51] Qi Zheng, Zhengzhong Tu, Pavan C Madhusudana, Xiaoyang Zeng, Alan C Bovik, and Yibo Fan. FAVER: Blind quality prediction of variable frame rate videos. *Signal Processing: Image Communication*, 122:117101, 2024. 6