Full-Reference and No-Reference Quality Assessment for Video Frame Interpolation

Jinliang Han, Xiongkuo Min*, *Member, IEEE*, Jun Jia, Yixuan Gao, Xiaohong Liu, *Member, IEEE*, and Guangtao Zhai, *Fellow, IEEE*

***Abstract*—Video frame interpolation (VFI) synthesizes new frames from original video frames to produce high frame-rate videos and enhance their visual appeal. The quality of these interpolated frames significantly affects the perceptual experience of the synthesized video. Recent research in VFI has increasingly focused on perceptual quality of the interpolated frames and the overall video. However, most existing quality metrics do not align well with human perceptual experiences and often suffer from unnatural artifacts in the interpolated frames. Consequently, there is an urgent need for VFI video quality assessment (VFIVQA) methods to assess the quality of the synthesized videos. In this paper, we propose both a full-reference (FR) method and a no-reference (NR) method for VFIVQA. The FR method employs two feature extraction blocks to measure continuous frame changes, extracting flow features with short temporal spans and motion features with long temporal spans. By calculating multilevel similarities in the temporal dimension of 3D convolutional neural networks and fusing these similarity features, the quality score of the VFI video is obtained from the quality regression network. Since the flow feature extraction block does not utilize the reference VFI video, the proposed NR method consists solely of this feature block. Extensive validation on several VFIVQA datasets demonstrates that the proposed methods outperform state-of-the-art FR and NR methods.**

***Index Terms*—Video frame interpolation, video quality assessment, perceptual quality, video incoherence.**

I. Introduction

VIDEO has become the primary medium for information communication in modern times, and a large number of videos are available on the Internet. However, due to bandwidth constraints, the frame rate of videos has not seen significant improvements. Fortunately, video frame interpolation (VFI) techniques [1]–[3], which synthesize new frames between the original video frames, can enhance the frame rate or smooth motion [4], [5] for low frame rate (LFR) videos, thereby alleviating bandwidth demands.

In recent years, novel VFI techniques have been proposed to generate videos with higher perceptual quality [6]–[8]. This development raises a new challenge: existing image and video quality assessment (IQA/VQA) methods may not be suitable for VFI videos [9], [10]. On the one hand, due to the persistence of vision, observers are particularly sensitive to frame continuity when watching videos. In other words, it is not only the quality of individual frames but also the coherence of multiple frames that affects the human viewing experience. Current IQA methods severely overlook the coherence of video frames and the temporal changes of objects, making them unsuitable for the task of VFI video quality assessment (VFIVQA) [10]–[12]. On the other hand, most VQA methods are specifically designed for videos affected by typical distortions, such as compression and blur. However, during the VFI synthesis process, the inherent uncertainty often introduces distortions throughout the video, combining high-level content distortions with low-level texture distortions [9]. Despite their impact on the perceptual quality of the video, VQA studies focusing on these perceptual distortions have received limited attention.

Research on perceptual VQA is typically divided into two categories: subjective and objective studies [13]–[17]. Several subjective VQA datasets have been constructed. To investigate specific perceptual effects, distortions within the datasets are synthesized on a set of original videos. In subjective VFIVQA studies [18], videos synthesized by different VFI methods are evaluated by subjects who are instructed to assess the overall quality of the VFI videos. However, there are still very few subjective VQA studies that specifically address the issue of frame incoherence in the VFI videos.

While subjective studies in VQA are generally accurate, they are also time-consuming and laborious. Consequently, objective studies have become the primary research focus. With the development of VQA towards perceptual consistency, objective studies rely on quality assessment scores provided by subjective evaluation experiments. Among objective VQA methods, full-reference (FR) and no-reference (NR) methods are categorized based on whether or not the original video is used as a reference [19]–[22]. Specifically for VFIVQA, the methods proposed by Danier *et al.* [11] and Hou *et al.* [10] draw on the concept of perceptual distance from the distortion to the reference [23] to predict quality scores. These are FR methods, that require the original HFR video as a reference. Although FR VQA methods usually have better perceptual consistency due to the additional information, obtaining the original HFR video is still a challenge in VFIVQA. This limits the application scenarios of FR methods.

Accordingly, in this paper, we propose not only a FR method (VFIVQA-FR) but also a NR method to predict the quality of VFI videos (VFIVQA-NR). Both methods are designed within the same VQA framework, utilizing an end-to-end neural network architecture that involves video feature extraction and

Jinliang Han, Xiongkuo Min, Jun Jia, Yixuan Gao, Xiaohong Liu, and Guangtao Zhai are with the Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: hanjinliang@sjtu.edu.cn; minxiongkuo@sjtu.edu.cn; jiajun0302@sjtu.edu.cn; gaoyixuan@sjtu.edu.cn; xiaohongliu@sjtu.edu.cn; zhaiguangtao@sjtu.edu.cn).

* Corresponding author.

quality regression. For VFIVQA-FR, two types of feature extraction blocks, namely the flow feature block and the motion feature block, are designed to learn the quality from frames of VFI videos. These two blocks extract features from frames with short and long temporal spans, respectively. For short temporal spans, three consecutive frames in the VFI video are considered as a triplet, and flow features are extracted from the triplet using 3D convolutional neural networks (CNNs). For long spans, motion features are extracted from key frames in the VFI video and the corresponding reference frames using pre-trained video networks. The final perceptual quality scores of the VFI video are obtained through the fusion of these two types of features and the learning of feature-quality mappings. For VFIVQA-NR, as motion features require reference frames, only the flow feature extraction block is included along with the quality regression network.

The performances of VFIVQA-FR and VFIVQA-NR are extensively verified on the BVI-VFI [24], VFIPS [10], and VFIIQA datasets [9], and compared with state-of-the-art (SOTA) FR and NR IQA/VQA methods. Experimental results demonstrate that VFIVQA-FR and VFIVQA-NR are superior to SOTA methods. Additionally, validation and ablation studies are conducted to evaluate the effectiveness of each component of the proposed methods. The main contributions of this work are as follows:

- Two novel methods, VFIVQA-FR and VFIVQA-NR, are proposed for FR and NR VFIVQA, respectively. Especially, the NR method is the first NR VQA specifically designed for VFI videos.
- VFIVQA-FR and VFIVQA-NR consider two important factors in human video perception: the continuity of the video over short periods and the memory effect over long periods. Meanwhile, the two proposed methods also introduce two types of features to obtain quality scores.
- A novel flow feature extraction block is proposed based on the context of frame triplets, applicable to both FR and NR VFIVQA.
- Quantitative comparison of IQA/VQA methods and cross-validation experiments are performed on VFIVQA datasets, and the proposed methods outperform SOTA methods.

The rest of the paper is organized as follows. Section II reviews related works on VQA methods. Section III introduces the proposed VFIVQA models. Section IV provides the experimental results and discusses the performance. Finally, concluding remarks are given in Section V.

II. Related Works

This section reviews previous works on FR and NR VQA, and summarizes subjective and objective VQA methods related to VFI.

A. Video Quality Assessment

1) FR VQA: FR VQA methods assess video quality by comparing the distorted video with the reference video, typically achieving high accuracy [25]. The most commonly used methods are PSNR and SSIM [26]. These two methods compute the Euclidean distance and structural similarity, respectively, between the distorted and reference videos at each frame. They then average the frame scores to obtain an overall quality score. Combining temporal information with IQA metrics, VMAF [27] extracts multiple IQA and motion features, and employs support vector regression (SVR) to map these features to video quality. With the increase in computing resources, VQA models based on deep neural networks have emerged, demonstrating exceptional performance owing to their powerful learning capabilities. The CVQA-FR [28] method calculates structural similarity and texture similarity on multi-level feature maps generated from multiple frames. C3DVQA [29] introduces 3D CNNs to learn spatio-temporal features by combining 2D CNNs for VQA.

2) NR VQA: NR VQA methods take only the distorted video as input, making them more widely applicable in practical scenarios [19]–[21], [30]. Earlier NR VQA methods compute the quality of each frame using NR IQA techniques [31], and then consider the average of all frame scores as the final score. More recent methods draw upon the spatial information acquisition strategies from advanced IQA methods [32]–[38]. However, IQA overlooks important temporal features. TLVQM [39] improves regression accuracy by extracting rich spatio-temporal features at two levels for training SVR models. VIDEVAL [40] enhances classic NR VQA methods by regressing features to video quality. However, these handcrafted features are limited in their ability to perceive video content. Currently, many methods utilize deep features and feedforward neural networks to learn the relationship between videos and quality scores [28], [41]–[44]. VSFA [41] extracts features using a pre-trained ResNet backbone and employs a gated recurrent unit (GRU) network as a temporal quality regressor. CVQA-NR [28] enhances the feature extraction capability with a unique ladder structure and regresses quality scores through fully connected layers. SimpleVQA [42] incorporates SlowFast features to integrate temporal information, improving quality prediction accuracy. FAST-VQA [43] samples videos using an efficient fragment sampling block with attention networks and achieves the best VQA accuracy for general videos.

B. Subjective VFIVQA

The existing subjective VFIVQA datasets include BVI-VFI, VFIPS, and VFIIQA. The detailed specifications of the three datasets are listed in Table I. In addition, some sample frames in the datasets with different types of visual distortions are shown in Fig. 1.

1) BVI-VFI: The first subjective VFIVQA study is the BVI-VFI [24]. In the subjective experiments, the scores of VFI videos are annotated using the double stimulus continuous quality scale methodology [45], and the Differential Mean Opinion Score (DMOS) values are computed for each video. The BVI-VFI dataset contains 540 VFI videos generated from 36 source videos, with various spatial resolutions and frame rates. Based on the types of VFI algorithms, distortions can be categorized into two main types: Traditional (Trad)

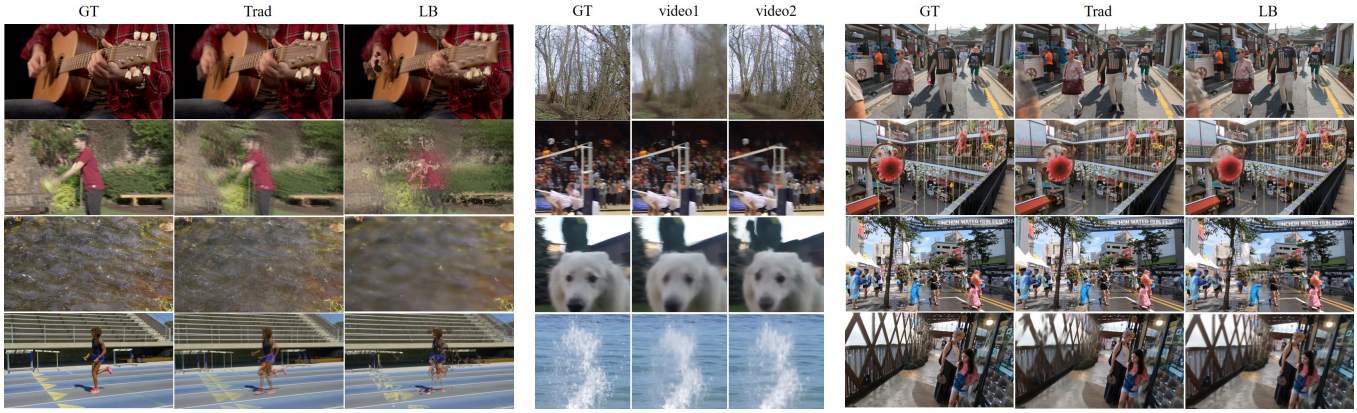


Fig. 1. Sample frames from the video contents contained in the VFI quality assessment datasets: (a) BVI-VFI (b) VFIPS (c) VFIIQA. GT stands for Ground Truth. Trad and LB refer to the two kinds of distortions (traditional and learning-based) in BVI-VFI and VFIIQA. VFIPS contains LB distortions, represented as pairs of videos.

TABLE I
VIDEO QUALITY ASSESSMENT DATASETS FOR VIDEO FRAME INTERPOLATION.

Item	Dataset		
	BVI-VFI	VFIPS	VFIIQA
Count	36	500	56
Total	540	5948	488
Resolution	540p, 1080p, 2160p	256×256	1280×720
Framerate	30, 60, 120fps	2fps	-
Length	5s	12 frames	3 frames
Format	MP4	Patch	PNG
Distortion	Trad, LB	LB	Trad, LB
Ground-truth	DMOS	2AFC score	MOS

and Learning-based (LB). The Trad distortions include those generated by averaging and repeating algorithms, while the LB distortions [46]–[49] stem from DVF, QVI, and STMFNet [50]–[52]. Fig. 1(a) illustrates the two types of visual distortions. The Trad distortions are primarily involve single-frame ghosting and reduced sharpness, while the LB distortions manifest as unstable object structures and the appearance of artifacts. All of these factors significantly impact the perceptual quality of VFI.

2) VFIPS: The VFIPS [10] dataset is a large-scale VFIVQA dataset. It focuses on perceptual similarity and consists of thousands of cropped video patches extracted from 12-frame VFI video clips. The subjective scores are annotated using the Two-Alternative Forced Choice (2AFC) [23] experiment. Accordingly, in the subjective experiments, two VFI videos are presented with the reference video, and then the subjects are required to choose the better video. Although automatic annotation is also conducted to expand the dataset, the data annotated by subjects includes 5978 sequences generated from 500 source videos. Only LB distortions are contained in the dataset.

3) VFIIQA: Given that not only the overall quality of videos but also the quality of individual frames significantly impacts the perception of VFI videos, the VFIIQA [9] dataset was constructed in our earlier work to study VFIVQA. The subjective experiments in VFIIQA were specifically designed to evaluate the synthesized frames in VFI videos. Two consecutive frames adjacent to each frame under evaluation were extracted to form a triplet of frames. The triplets were observed, and the scores for the synthesized frames were annotated by subjects. The data annotation methodology used is the single stimulus continuous quality rating and the Mean Opinion Score (MOS) values were calculated. Eight academic VFI algorithms with LB distortions and one industrial algorithm with Trad distortions were employed to extensively study the perceptual impact of single frames.

C. Objective VFIVQA

The basic metrics for assessing visual quality in VFI algorithm research [2] are PSNR and SSIM. However, these metrics do not consistently align with human subjective perception of VFI videos [12]. Although some algorithms have introduced LPIPS [23] as the evaluation criterion to pursue high-quality VFI, these metrics overlook motion information in the temporal dimension, resulting in limited perceptual accuracy. Taking into account temporal information, more recent objective VFIVQA methods are learning-based VQA methods. FloLPIPS [11] integrates the optical flow [53], [54] features between frames into the perceptual metric. It is trained and validated on the BVI-VFI dataset, confirming the effectiveness of temporal information for the perception of VFI videos. VFIPS [10] incorporates attention mechanisms into the LPIPS through a transformer architecture and uses a continuous sequence of frames as input to capture inter-frame motion information. Both of these methods are FR methods, and their effectiveness for single-frame quality perception in VFI videos has not been validated.

III. PROPOSED METHOD

This section provides a detailed introduction to the proposed VFIVQA methods. The frameworks of the proposed VFIVQA-FR and VFIVQA-NR methods are illustrated in Fig. 2 and

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2025.3540104

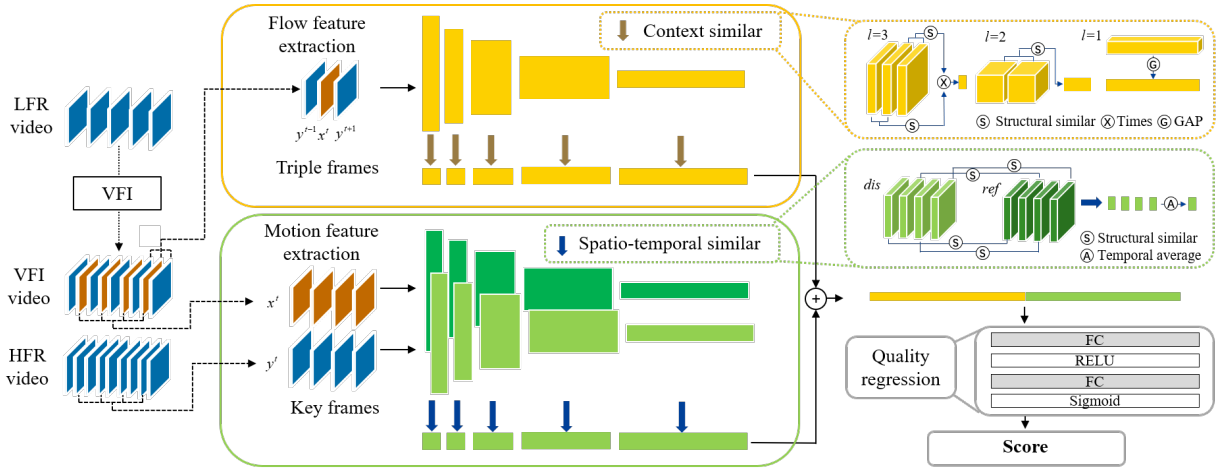


Fig. 2. Overall architecture of the proposed VFIVQA-FR method. The distorted inputs are the VFI videos generated from the LFR videos and the references are the corresponding HFR videos. Solid arrows represent the data flow, and dotted arrows indicate the frame extraction process. The solid line boxes encompass the feature extraction block and the quality regression block, and the dotted line boxes illustrate the similarity calculation.

Fig. 3, respectively. Both methods employ an end-to-end VQA architecture that contains feature extraction blocks and quality regression blocks.

A. Feature Extraction

In the VFIVQA-FR method, video features of both VFI videos and reference videos are extracted, as shown in Fig. 2. The frames synthesized by VFI methods are denoted by x^t, while the original frames in both the LFR video and the HFR video are represented by y^t. Features representing short-term continuity are extracted from the flow feature extraction block, while features representing long-term semantics are extracted from the motion feature extraction block.

1) Flow Feature Extraction: Given that VFI involves interpolating two frames in the time dimension to synthesize a new frame, continuous playback often results in unnatural motion or deformation of objects in the synthesized frame compared to the original frames [10]. Moreover, the human visual system assesses video quality based on the smoothness between consecutive frames [9]. Although optical flows are computed from consecutive input frames and can represent the smoothness of objects and scenes [54]–[56], the computation is expensive, and the feature scale is singular [11]. Therefore, a novel extraction block is designed to extract multi-scale flow features and perceive inter-frame smoothness within short temporal spans. Similar to the VFI process, a triplet of frames, comprising one synthesized frame and the two adjacent original frames from VFI videos, is input into the flow feature block.

In the process of feature extraction, a multi-layer neural network architecture is introduced as the backbone to extract multi-scale features. Considering that the ResNet3D architecture [57], [58] has been demonstrated to be effective in learning spatio-temporal information, the flow features in our method are learned from continuous frames by the feature layers of ResNet3D. Specifically, the backbone is composed of multiple layers of cascaded 3D CNNs. To facilitate the learning of desired features with fewer parameters, skip connections are incorporated. For the feature map f_k of the k-th residual block, the skip connection can be represented by:

$$f_k = f_{k-1} + F(f_{k-1}; \theta_k), \tag{1}$$

where $F(f_{k-1}; \theta_k)$ represents the function of convolutions parameterized by weights θ_k, and the activation function of rectified linear unit (ReLU).

TABLE II
PARAMETERS OF THE FEATURE MAPS IN RESNET3D.

Layer name	Chanel (C)	Size ($l \times h \times w$)
conv1	64	$L \times H \times W$
conv2	64	$L \times H \times W$
conv3	128	$[L/2] \times H/2 \times W/2$
conv4	256	$[L/4] \times H/4 \times W/4$
conv5	512	$[L/8] \times H/8 \times W/8$

The entire feature learning process in the networks is divided into five stages based on the size of the feature maps, with the corresponding networks labeled conv1 to conv5 in Fig. 2. As the feature extraction stage progresses, the temporal dimension of the feature maps decreases due to the action of 3D CNNs. As a result, information across the temporal scale of video frames is gradually integrated into the feature maps. Specifically, the output sizes $C \times l \times h \times w$ of the feature maps for each stage of the network are provided in Table II, where C is the number of channels and l is the temporal size of the convolution. L, H and W refer to the size of output feature maps at conv1.

Learning the temporal continuity for each triplet of frames, the network can utilize the flow features as the primary basis for quality assessment. With the extraction backbone handling temporal dimension feature, the flow feature vector is calculated with the proposed ConTextual Similarity (CTS) to represent frame coherence within a short temporal span. The

CTS calculation is integrated into the feature maps, which will be described in detail in Section III-B.

2) Motion Feature Extraction: Given that objects within the video may not always exhibit consistent motion, the perception of the video is a complex process [42]. The effects of distortion in VFI videos depend not only on individual frames but also on multiple frames with long time spans. Consequently, a motion feature extraction block is proposed to assist VFIVQA-FR in conducting comprehensive evaluations. For motion features of video content, models pre-trained on a large number of videos have sufficient representation abilities. The pre-trained ResNet3D is commonly used for video feature extraction, especially in applications such as video recognition and action recognition. In VQA methods, the effectiveness of using pre-trained models for video feature extraction has been proven [41], [43]. Following the principles, motion features are extracted using the pre-trained ResNet3D, with the network parameters frozen. Inheriting the stability and comprehensiveness of the human visual system, the pre-trained model can perceive videos over long time spans.

The motion feature block simultaneously takes the generated VFI video and the corresponding HFR video as inputs for FR VFIVQA. For each video clip under evaluation, the key frames, which are usually the frames synthesized from the VFI video, are selected as the distorted input. The corresponding frames from the HFR video serve as the reference input. Motion features are extracted from both the distorted and reference frames to obtain comparable feature maps with multi-scale representations. Since the feature maps output by ResNet3D gradually integrate information from multiple key frames in the temporal dimension, a spatio-temporal similarity calculation is proposed in Section III-B to better quantify the feature maps as vectors in the block.

B. Similarity Computation

1) Context Similarity: In the proposed flow feature extraction block, feature maps of the triplet frames are obtained, covering multiple scales in both the temporal and spatial dimensions. To enable the 3D CNNs to learn smoothness features between consecutive frames, a CTS computation of the feature map is designed in this section. Considering the primary form of perceptual loss between VFI frames and original frames, the CTS introduces structural similarity by focusing on the global correlations of feature maps at each stage:

$$S_k(f_k^t, f_k^{t+1}) = \frac{\sigma_k^{t(t+1)} + c}{\left(\sigma_k^t\right)^2 + \left(\sigma_k^{t+1}\right)^2 + c}, \tag{2}$$

where f_k^t and f_k^{t+1} represent two adjacent feature components in the time dimension for the k-th stage feature map, with $t = 1, \ldots, l-1$. The $\left(\sigma_k^t\right)^2$ and $\left(\sigma_k^{t+1}\right)^2$ refer to the global variances of f_k^t and f_k^{t+1}, while $\sigma_k^{t(t+1)}$ represents the global covariance between f_k^t and f_k^{t+1}, and c is a constant avoiding numerical singularity. The CTS aims to calculate similarity in temporal dimension from the overall feature representation. Accordingly, the CTS in the k-th stage CTS_k is:

$$CTS_k = \prod_{t=1}^{l-1} S_k(f_k^t, f_k^{t+1}). \tag{3}$$

Since the input designed in the extraction block is a triplet, $l \leq 3$. Specifically, the size of the feature map is reduced stage by stage in the time dimension as shown in Table II. To provide a clearer explanation, the calculation of the CTS for different values of l is shown in the dotted box in Fig. 2. When l is 1, the temporal information across feature maps is fully integrated by the networks. Consequently, the global average pooling (GAP) is introduced to generate high-level features. Finally, by incorporating multi-scale features, all vectors are connected into the flow feature v_F:

$$v_F = \text{concat}(\{CTS_k\}_{k=1}^5). \tag{4}$$

2) Spatio-temporal Similarity: Although motion features are typically treated as high-level features in VQA methods, they may result in the loss of motion information at the pixel level [28]. Given that distortions in VFI videos can affect both pixel-level and semantic-level features, spatio-temporal similarity (STS) is designed at each stage in the motion feature extraction block. STS relies on a reference-distortion comparison strategy, computing the structural similarity across multiple key frames and the corresponding reference. The STS_k at the k-th stage is defined as:

$$STS_k = \frac{1}{l} \sum_{t=1}^{l} S_k(f_k^t(x), f_k^t(y)), \tag{5}$$

where x and y represent the distorted and reference video clips, and l is the temporal size of the feature maps. Integrating over temporal scales, the motion feature vector is:

$$v_M = \text{concat}(\{STS_k\}_{k=1}^5). \tag{6}$$

C. Feature Fusion and Quality Regression

Feature fusion and regression to quality scores play a pivotal role in the subjective consistency and generalizability of VQA methods. It has been demonstrated that feature fusion can significantly impact the performance of features [44]. The most commonly used feature fusion method in VQA is the concatenation of feature vectors, although there are also methods that directly sum similarity measures [28]. In this paper, as indicated in Eq. (4) and Eq. (6), the concatenation operation is employed, which allows for maximum information retention within the features. Depending on the exceptional performance of the quality regression network in high-dimensional spaces, the fused features can be accurately mapped to quality scores. The feature vectors obtained from the two parts of the feature extraction blocks are concatenated as shown in Fig. 2, denoted as $v = v_F \oplus v_M$, where $\oplus$ stands for the concatenation operation.

The subsequent quality regression network follows a consistent design principle in VQA, employing a mapping network composed of two cascaded fully connected layers. This design ensures a sufficiently high-dimensional representation of the

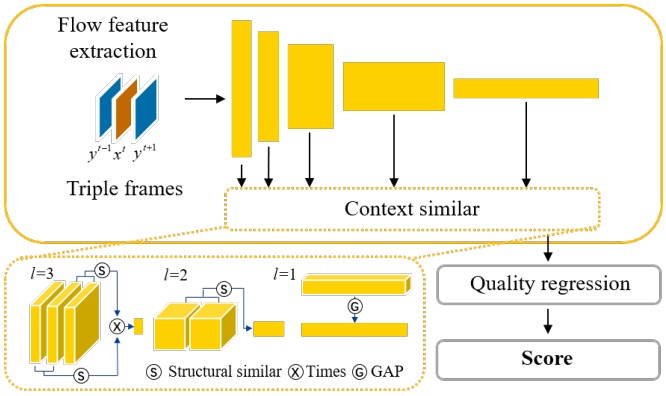


Fig. 3. The architecture of the proposed VFIVQA-NR method. The inputs are the distorted VFI videos and only the flow feature extraction block is included to obtain the contextual similarity.

extracted feature vectors and adapts to the trends in subjective scores. The activation functions are ReLU and Sigmoid, as shown in Fig. 2. ReLU mitigates the interdependence between neurons, thereby alleviating overfitting, and Sigmoid accomplishes the final score mapping.

D. No-reference Framework

As mentioned in Section III-A, the designed flow feature extraction block takes a triplet of frames from the VFI video as input. Due to the absence of the reference video, the extraction block can be independently designed for NR VFIVQA. The proposed VFIVQA-NR framework, depicted in Fig. 3, extracts features at multiple scales from the input triplet frames and computes the CTS. After concatenating all flow feature vectors into a vector, denoted as v_F, it serves as the input to the quality regression network, to obtain the final score. The experimental validation demonstrates the simplicity and effectiveness of the proposed NR framework, making it suitable for addressing NR VFIVQA problems.

IV. Experiments

In this section, the experimental settings are first introduced, which include datasets, experimental details, SOTA IQA/VQA methods used for comparison, and the evaluation metrics. Then, experiments are presented to demonstrate the effectiveness and superiority of the proposed method. In addition, statistical significance analysis and ablation studies validate the importance of individual modules. The generalization abilities of the methods are also verified through cross-database evaluation. Finally, we evaluate the computational efficiency of our methods.

A. Experimental Settings

1) Datasets: Following the tradition [40], all three relevant datasets introduced in Section II-B are used separately for training and validation. Given the differences in data format and distribution, separate validation ensures the fairness of the experiments and enables a comprehensive evaluation of the methods across different perceptions. Each dataset is divided into non-overlapping training and test sets, with an 80%-20% split. To prevent information leakage between training and testing, the random split is based on the reference video content. Videos with different distortion types but from the same reference video are assigned to the same set.

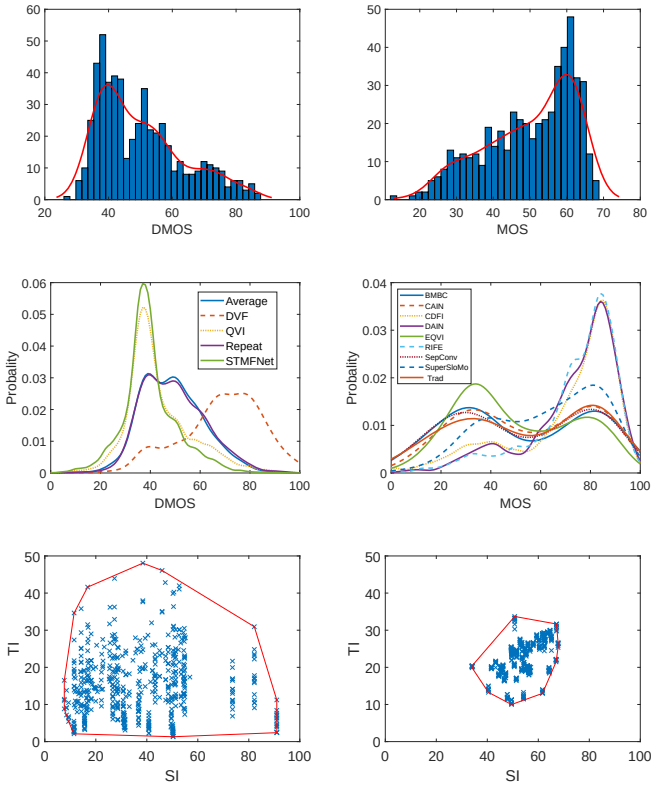


Fig. 4. Comparison between the BVI-VFI and VFIIQA datasets. The first row shows the histograms and the fitted kernel distributions of the two datasets. The second row shows the distribution comparisons of all opinion scores between VFI algorithms. The third row is the distribution of SI and TI in the paired feature space with the corresponding convex hulls.

Furthermore, since the BVI-VFI dataset [24] comprises subsets with varying spatial resolutions and frame rates, the partition ratios across different subsets are kept consistent. For the VFIPS dataset [10], in pursuit of higher perceptual quality accuracy, only the manually annotated portions are utilized. Compared to the two datasets, the VFIIQA dataset [9] focuses more on the quality of single frames. A comparison between the VFIIQA and BVI-VFI datasets is illustrated in Fig. 4, where both datasets include quality scores as ground truths. Histograms of data with fitted kernel distribution curves and probability distributions of raw scores for various VFI methods are also given in this figure. From the histograms, it can be observed that the VFIIQA exhibits a symmetric distribution with BVI-VFI, which is attributed to differences in the annotation methodology. This is because the VFIIQA employs MOSs as annotations, whereas the BVI-VFI dataset is annotated with DMOSs. Regarding the distributions of distortions from various VFI algorithms in both datasets, VFIIQA shows a greater variety of distortions in the low-quality range, serving as a valuable supplement to the types of

VFI distortions. The Spatial perceptual Information (SI) and the Temporal perceptual Information (TI) [59] for each video in both datasets are also depicted in Fig. 4. In the SI and TI feature spaces, the BVI-VFI contains more spatio-temporal information, while VFIIQA provides less information, posing a greater challenge for VQA methods.

2) Implementation Details: Both VFIVQA-FR and VFIVQA-NR are trained end-to-end on the datasets and the motion feature extraction backbone in VFIVQA-FR is pre-trained on the Kinetics-400 dataset [60]. The model parameters are optimized by the ADAM optimizer with an initial learning rate of 10^{-4}, which is reduced by a factor of 2 after every 50 iterations. For the dataset where the score is MOS, the training loss is the Mean Square Error (MSE) loss, which is a common choice for the quality regression tasks:

$$\mathrm{MSE} = \frac{1}{N}\sum_{n=1}^{N}(Q_n - \hat{Q_n})^2, \tag{7}$$

where N is the number of scores, Q_n and $\hat{Q_n}$ are the ground-truth quality score and the estimated score, respectively.

Limited by the size of the datasets, a 5-fold cross-validation process is employed for evaluating performance on the BVI-VFI and VFIIQA datasets in accordance with the data partitioning principles described in Section IV-A1. The process is iterated 10 times, during which performance metrics are computed, and the average of all results is reported for the final assessment. During training, the spatial resolution of each video is downsampled to 256×256. For the BVI-VFI dataset, the key-frame selection process follows a fixed time interval method, where one frame is selected for each second of the video. In contrast, for the other two datasets (VFIIQA and VFIPS), all frames are included without subsampling. For the VFIPS dataset, which involves subjective 2AFC scores, the output scores of each pair of videos are used as estimates for comparison, and the loss function used is Binary Cross Entropy (BCE). These data processing and training procedures are consistent across all comparison methods.

3) Compared Methods: The VQA methods specifically designed for VFI are FloLPIPS [11] and VFIPS [10]. Following their practices, the SOTA IQA/VQA methods are compared to validate the superiority of the proposed methods. For FR methods, commonly adopted IQA methods include PSNR, SSIM, LPIPS [23], and DISTS [61]. FR VQA methods compared include VMAF [27], CVQA-FR [28], and C3DVQA [29]. Furthermore, specific evaluation metrics for VFI like Interpolation Error (IE) and Normalized Interpolation Error (NIE) [53] as Eq.(8) and Eq.(9) are also compared:

$$\mathrm{IE} = [\frac{1}{M}\sum_{(x,y)}(I(x,y) - I_{GT}(x,y))^2]^{\frac{1}{2}}, \tag{8}$$

$$\mathrm{NIE} = [\frac{1}{M}\sum_{(x,y)}\frac{(I(x,y) - I_{GT}(x,y))^2}{\|\nabla I_{GT}(x,y)\|^2 + 1}]^{\frac{1}{2}}, \tag{9}$$

where I and I_{GT} are the distorted image and the ground-truth image respectively, (x, y) is the coordinate of the pixel and M is the number of pixels. Although IE and NIE are not IQA or VQA methods, these two metrics are included as FR methods for using ground-truth image information. the evaluation of NR methods, the IQA methods compared are BRISQUE [31], DBCNN [33], MANIQA [62], and VFIPQA [9]. The VQA methods performing well in user-generated videos include TLVQM [39], VIDEVAL [40], VSFA [41], CVQA-NR [28], SimpleVQA [42], FAST-VQA [43], and Q-Align [63].

Comparison methods and their associated parameters are sourced from official code repositories and are maintained at their default settings. For trainable methods, fine-tuning is performed on the respective datasets to achieve the best performance. It should be noted that not all methods are applicable to all test datasets, and reasonable adaptations are made to ensure fairness in evaluation. The loss function of the VFIPS method is adjusted to MSE due to the absence of 2AFC scores in the BVI-VFI and VFIIQA datasets. Moreover, as the VFIIQA dataset is presented in a format of triplets, the configurations of the input and the VFIPS network are adjusted from 12 frames to 3 frames. To accommodate VQA, IQA methods tested in the BVI-VFI take one frame per second and compute the average score, while for the other two datasets, scores for all frames are averaged for evaluation. These adjustments and adaptations are made to ensure that all the methods are comparable with the VFIVQA methods in each dataset, thus facilitating a fair evaluation.

4) Evaluation Criteria: Four commonly used performance criteria: Spearman Rank Order Correlation Coefficient (SROCC), Kendall Rank-Order Correlation Coefficient (KROCC), Pearson Linear Correlation Coefficient (PLCC), and Root Mean Square Error (RMSE) are employed as evaluation metrics [64]. The correlation coefficients are used to assess the consistency between the predicted scores with subjective perceptions, and the RMSE is used to evaluate the fitting error. When calculating PLCC and RMSE, the five-parameter logistic function [65] is used to non-linearly map the objective score to the subjective score:

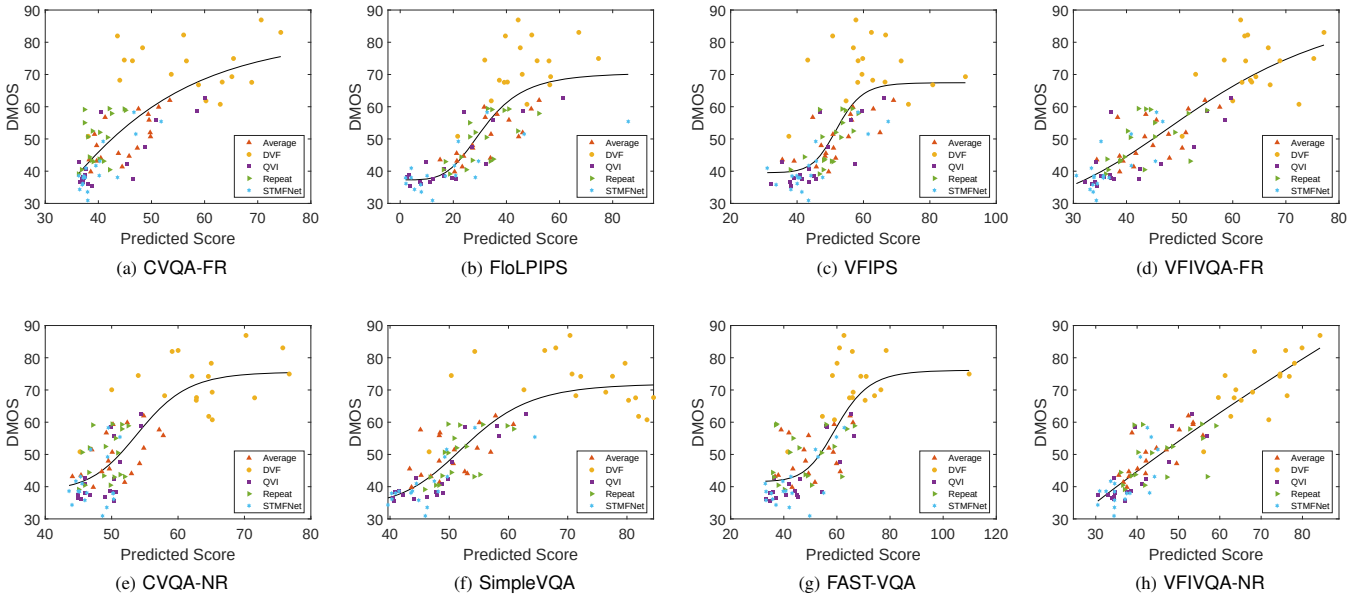
$$Q' = \beta_1(\frac{1}{2} - \frac{1}{1+\exp^{\beta_2(Q-\beta_3)}}) + \beta_4 Q + \beta_5, \tag{10}$$

where Q and Q' are the predicted and fitted quality scores, and $\{\beta_i, i = 1, \ldots, 5\}$ are the parameters determined by the curve fitting.

As a supplementary metric evaluation, the receiver operating characteristic (ROC) analysis has also been introduced [66]. The ROC analysis is based on the principle of assessing whether there are qualitative differences between two videos and is used to evaluate from two aspects. First, all possible pairs of videos are compared and categorized into those with significant quality differences and those without such differences. The ROC analysis is then employed to determine whether various objective metrics can distinguish between videos with significant differences and those without, referred to as "Different/Similar ROC analysis". Subsequently, videos with significant differences are divided into positive difference and negative difference video pairs, and ROC analysis is used to assess whether objective metrics can differentiate between these positive and negative difference videos, denoted as "Better/Worse ROC analysis". Area under the ROC curve

TABLE III
PERFORMANCE COMPARISON BETWEEN FR VQA METHODS AND THE PROPOSED METHOD. PERFORMANCE EVALUATION METRICS COMMONLY USED IN VFI RESEARCH ARE REPRESENTED IN GRAY. THE BEST MODEL IN EACH COLUMN IS IN BOLD, AND THE SECOND-BEST MODEL IS UNDERLINED.

Methods	BVI-VFI				VFIPS	VFIIQA			
	SROCC ↑	KROCC ↑	PLCC ↑	RMSE ↓	2AFC ↑	SROCC ↑	KROCC ↑	PLCC ↑	RMSE ↓
PSNR	0.587 (.097)	0.409 (.074)	0.578 (.106)	11.019 (0.934)	0.795	0.131 (.155)	0.085 (.107)	0.310 (.087)	16.738 (1.100)
SSIM [26]	0.582 (.123)	0.412 (.094)	0.580 (.100)	11.019 (0.947)	0.813	0.259 (.174)	0.182 (.121)	0.350 (.131)	16.367 (1.226)
LPIPS [23]	0.700 (.079)	0.511 (.071)	0.695 (.096)	9.651 (1.339)	0.858	0.714 (.064)	0.522 (.055)	0.725 (.063)	12.542 (2.095)
DISTS [61]	0.603 (.093)	0.430 (.071)	0.573 (.110)	11.046 (1.034)	0.820	0.755 (.050)	0.562 (.045)	0.781 (.052)	10.895 (0.744)
VMAF [27]	0.527 (.094)	0.368 (.074)	0.524 (.104)	11.514 (0.896)	0.836	0.281 (.146)	0.198 (.105)	0.369 (.119)	16.241 (1.241)
CVQA-FR [28]	0.708 (.101)	0.513 (.088)	0.722 (.096)	9.356 (1.476)	0.822	0.790 (.042)	0.595 (.041)	0.786 (.058)	10.818 (1.772)
C3DVQA [29]	0.677 (.114)	0.489 (.092)	0.610 (.130)	12.016 (1.450)	0.788	0.474 (.126)	0.330 (.095)	0.455 (.110)	17.301 (2.492)
IE [53]	0.561 (.094)	0.389 (.069)	0.515 (.085)	11.657 (0.964)	0.795	0.130 (.155)	0.085 (.107)	0.281 (.107)	16.816 (0.996)
NIE [53]	0.597 (.104)	0.425 (.075)	0.554 (.093)	11.286 (0.808)	0.807	0.375 (.147)	0.258 (.103)	0.430 (.127)	15.698 (0.784)
FloLPIPS [11]	0.733 (.083)	0.536 (.081)	0.725 (.094)	9.199 (1.622)	0.869	0.724 (.073)	0.530 (.065)	0.738 (.083)	11.686 (1.392)
VFIPS [10]	0.743 (.091)	0.553 (.084)	0.691 (.124)	11.020 (1.684)	0.866	0.667 (.080)	0.480 (.068)	0.650 (.063)	14.574 (2.156)
VFIVQA-FR	**0.828** (.056)	**0.634** (.040)	**0.839** (.057)	**7.295** (1.123)	**0.871**	**0.823** (.041)	**0.638** (.041)	**0.833** (.050)	**9.582** (1.381)



(a) CVQA-FR (b) FloLPIPS (c) VFIPS (d) VFIVQA-FR
(e) CVQA-NR (f) SimpleVQA (g) FAST-VQA (h) VFIVQA-NR

Fig. 5. Scatter plots of VQA metrics on the BVI-VFI dataset. The lines are the nonlinear fitted curves with a five-parameter logistic function. Different distortion types are represented by different scatter points.

(AUC) values are primarily reported for these two types of analysis, with higher values indicating better performance.

B. Overall Results

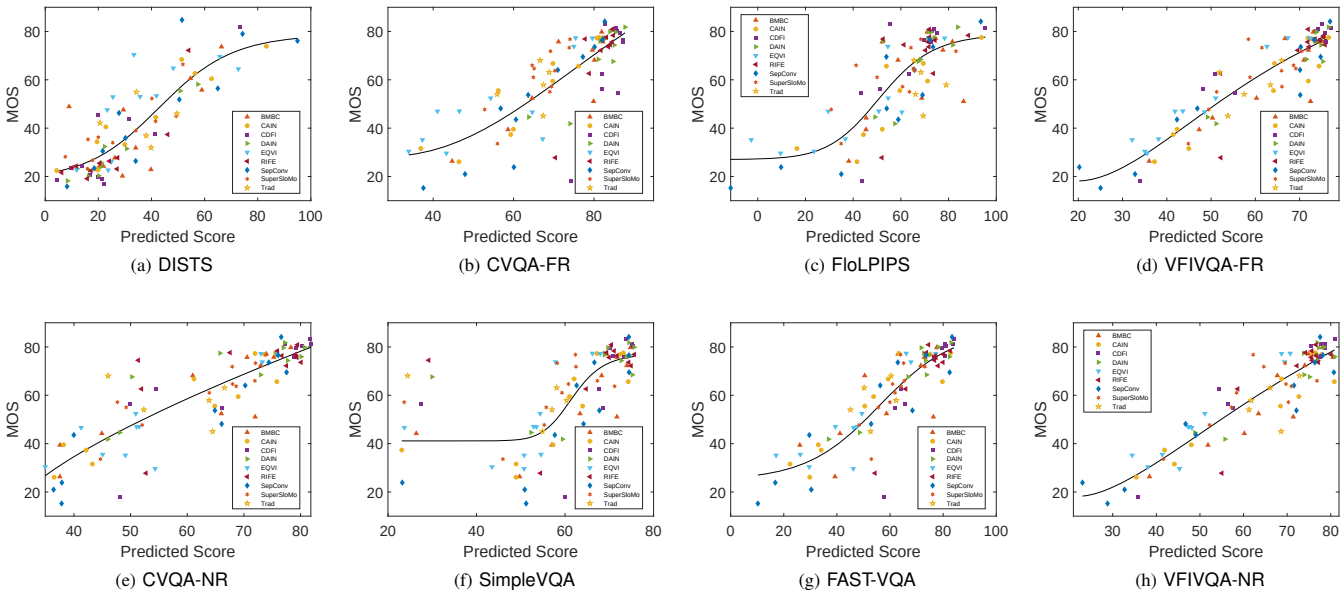
1) Quantitative Comparison: The quantitative experimental results on three datasets are presented in Table III and Table IV. The numbers in parentheses are the standard deviations of 10 times of cross-validation. The upper half of Table III includes the general IQA/VQA methods, and the lower half includes the methods specifically designed for VFI quality assessment. As shown in both tables, the proposed methods outperform the SOTA methods in both FR and NR VFIVQA.

Among the FR methods, the proposed VFIVQA-FR method outperforms SOTA methods across all four evaluation metrics. The performance of VFIPS is relatively close to FloLPIPS on the BVI-VFI dataset, and consistently lower than VFIVQA-FR across various metrics. VFIVQA-FR exhibits an improvement of more than 10% compared to these two methods. Additionally, the standard deviation of correlation coefficients indicates that the proposed method is the most stable during cross-validation. On the VFIPS and VFIIQA datasets, VFIVQA-FR also achieves optimal performance on their respective metrics. Notably, on the VFIIQA dataset, some general-purpose full-reference methods perform better than the specialized methods, which may be due to differences in perception tasks. The VFIVQA-FR overcomes this issue, providing improved performance and robustness.

The proposed VFIVQA-NR method also stands out as compared to all SOTA NR methods. Although there is a decrease in

TABLE IV
PERFORMANCE COMPARISON BETWEEN NR VQA METHODS AND THE PROPOSED METHOD. THE BEST MODEL IN EACH COLUMN IS IN BOLD, AND THE SECOND-BEST MODEL IS UNDERLINED.

Methods	BVI-VFI				VFIPS	VFIIQA			
	SROCC ↑	KROCC ↑	PLCC ↑	RMSE ↓	2AFC ↑	SROCC ↑	KROCC ↑	PLCC ↑	RMSE ↓
BRISQUE [31]	0.185 (.071)	0.131 (.048)	0.280 (.078)	13.455 (0.772)	0.794	0.481 (.211)	0.349 (.149)	0.492 (.171)	15.310 (1.681)
DBCNN [33]	0.686 (.054)	0.503 (.046)	0.738 (.068)	9.041 (1.124)	0.799	0.792 (.045)	0.607 (.043)	0.806 (.051)	10.156 (1.183)
MANIQA [62]	0.692 (.041)	0.512 (.040)	0.753 (.054)	8.902 (1.064)	0.787	0.707 (.075)	0.517 (.067)	0.694 (.088)	12.525 (1.403)
VFIPQA [9]	0.531 (.082)	0.371 (.058)	0.580 (.076)	11.077 (0.736)	0.842	**0.825** (.047)	**0.642** (.043)	0.820 (.052)	9.928 (1.214)
TLVQM [39]	0.446 (.080)	0.316 (.064)	0.474 (.104)	11.939 (1.058)	0.807	0.536 (.079)	0.376 (.057)	0.578 (.097)	14.233 (1.121)
VIDEVAL [40]	0.120 (.093)	0.086 (.064)	0.236 (.078)	13.308 (0.689)	0.814	0.504 (.091)	0.350 (.067)	0.528 (.074)	14.583 (1.249)
Q-Align [63]	0.534 (.109)	0.348 (.075)	0.548 (.075)	12.500 (0.780)	0.714	0.593 (.084)	0.428 (.062)	0.624 (.099)	13.597 (1.283)
VSFA [41]	0.552 (.093)	0.386 (.072)	0.588 (.096)	10.939 (1.182)	0.811	0.633 (.100)	0.453 (.083)	0.637 (.125)	13.272 (1.582)
CVQA-NR [28]	0.608 (.089)	0.442 (.071)	0.681 (.084)	14.516 (4.480)	0.825	0.776 (.114)	0.588 (.092)	0.721 (.230)	16.458 (5.631)
SimpleVQA [42]	0.750 (.057)	0.555 (.056)	0.778 (.049)	8.507 (0.766)	0.834	0.640 (.123)	0.457 (.096)	0.657 (.087)	13.155 (1.707)
FAST-VQA [43]	0.744 (.076)	0.552 (.073)	0.783 (.080)	8.329 (1.437)	0.821	0.734 (.129)	0.549 (.115)	0.722 (.146)	12.734 (3.054)
VFIVQA-NR	**0.810** (.047)	**0.626** (.042)	**0.838** (.049)	**7.506** (0.962)	**0.858**	0.821 (.041)	0.635 (.042)	**0.835** (.042)	**9.555** (1.164)



(a) DISTS (b) CVQA-FR (c) FloLPIPS (d) VFIVQA-FR
(e) CVQA-NR (f) SimpleVQA (g) FAST-VQA (h) VFIVQA-NR

Fig. 6. Scatter plots of VQA metrics on the VFIIQA dataset. The lines are the nonlinear fitted curves with a five-parameter logistic function. Different distortion types are represented by different scatter points.

consistency compared to FR methods, it still outperforms other NR VQA methods. Compared to the second-best method, VFIVQA-NR shows an improvement of approximately 7% on the BVI-VFI dataset. Furthermore, the proposed method is the first NR VQA method designed specifically for VFI videos, achieving high consistency with human perception across all three datasets, which is advantageous for the direct deployment of VFIVQA applications.

2) Qualitative Comparison: In order to provide a more intuitive representation of the consistency between objective predictions and subjective perceptual qualities, scatter plots and logistic function fitting results of the test outcomes have been presented. Scatter plots for the BVI-VFI and VFIIQA datasets are shown in Fig. 5 and Fig. 6, respectively. Each figure consists of two rows, presenting the results of representative FR methods and NR methods, respectively. In addition to the proposed methods, the three methods with the highest SRCC are selected for comparison. The scatter plots indicate that the proposed FR and NR methods exhibit better linearity on the two datasets compared to other methods, which implies a better consistency with MOS. Furthermore, through annotations of different distortion types, the proposed method also demonstrates good discrimination ability for videos with the same distortion type.

C. Performance on Different Subsets

The BVI-VFI dataset comprises videos with varying spatial resolutions and frame rates, thus this dataset can be divided into different subsets for further validation and analysis. Analyzing the differences in video quality perception from the

TABLE V
PERFORMANCE COMPARISON ON DIFFERENT SUBSETS OF BVI-VFI WITH DIFFERENT RESOLUTIONS.

Methods	540p		1080p		2160p		30fps		60fps		120fps	
	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
PSNR	0.681	0.683	0.605	0.683	0.595	0.622	0.543	0.611	0.563	0.581	0.389	0.572
SSIM [26]	0.657	0.638	0.658	0.651	0.668	0.699	0.533	0.668	0.585	0.588	0.417	0.502
IE [53]	0.697	0.697	0.593	0.642	0.480	0.500	0.519	0.574	0.557	0.533	0.334	0.465
NIE [53]	0.703	0.714	0.646	0.678	0.589	0.580	0.427	0.522	0.602	0.579	0.419	0.579
LPIPS [23]	0.751	0.788	0.718	0.766	0.724	0.737	0.577	0.668	0.713	0.717	0.585	0.716
DISTS [61]	0.697	0.702	0.673	0.723	0.639	0.657	0.490	0.604	0.621	0.607	0.435	0.636
VMAF [27]	0.650	0.627	0.569	0.639	0.408	0.579	0.451	0.511	0.549	0.518	0.402	0.553
CVQA-FR [28]	0.714	0.787	0.676	0.788	0.600	0.728	0.616	0.666	0.718	0.763	0.554	0.804
FloLPIPS [11]	0.793	0.816	0.736	0.762	0.718	0.745	0.718	0.745	0.762	0.758	0.573	0.714
VFIPS [10]	0.758	0.859	0.788	0.886	0.728	0.795	0.626	0.751	0.680	0.744	0.587	0.774
VFIVQA-FR	**0.847**	**0.876**	**0.830**	**0.894**	**0.806**	**0.848**	**0.796**	**0.804**	**0.802**	**0.864**	**0.693**	**0.913**
BRISQUE [31]	0.117	0.284	0.237	0.399	0.265	0.474	0.229	0.343	0.184	0.323	0.026	0.236
DBCNN [33]	0.717	0.754	0.677	0.854	0.699	0.755	0.674	0.684	0.747	0.765	0.566	0.830
MANIQA [62]	0.706	0.723	0.687	0.827	0.767	0.816	0.656	0.697	0.753	0.808	0.598	0.855
VFIPQA [9]	0.475	0.616	0.645	0.781	0.585	0.639	0.442	0.544	0.490	0.613	0.309	0.724
TLVQM [39]	0.576	0.654	0.497	0.638	0.460	0.557	0.383	0.472	0.428	0.510	0.337	0.544
VIDEVAL [40]	0.087	0.353	0.447	0.577	0.180	0.319	0.216	0.379	0.079	0.228	0.046	0.216
CVQA-NR [28]	0.640	0.713	0.642	0.843	0.635	0.733	0.626	0.669	0.653	0.751	0.427	0.776
SimpleVQA [42]	0.743	0.766	**0.820**	**0.907**	0.712	0.754	0.685	0.715	0.762	0.810	0.649	0.871
FAST-VQA [43]	0.720	0.827	0.770	0.827	0.765	0.818	0.734	0.758	0.726	0.792	0.523	0.798
VFIVQA-NR	**0.855**	**0.893**	0.790	0.870	**0.853**	**0.888**	**0.796**	**0.814**	**0.823**	**0.892**	**0.653**	**0.929**

perspectives of spatial resolution and frame rate allows for a more comprehensive assessment of VFI algorithms and further validates the generalizability of VFIVQA methods.

1) Resolution Subsets: Based on the spatial resolution of videos in the BVI-VFI dataset, it is divided into three sub-datasets with resolutions of 540p, 1080p, and 2160p, respectively. Each of these three subsets comprises 180 videos. During the training process, a cross-validation method is consistently employed. The dataset is split into training and test sets for each of the three subsets, and this process is iterated 10 times to obtain average results. The experimental results are presented in the left part of Table V. From the table, it is evident that among FR methods, VFIVQA-FR achieves the best performance in all three subsets of different resolutions. This indicates that the proposed method has a strong generalization ability across various spatial resolutions. In the case of NR methods, VFIVQA-NR performs optimally on the 540p and 2160p subsets. On the 1080p subset, the performance of the proposed method is second only to SimpleVQA, which might be attributed to the specific capabilities of SimpleVQA for this particular resolution.

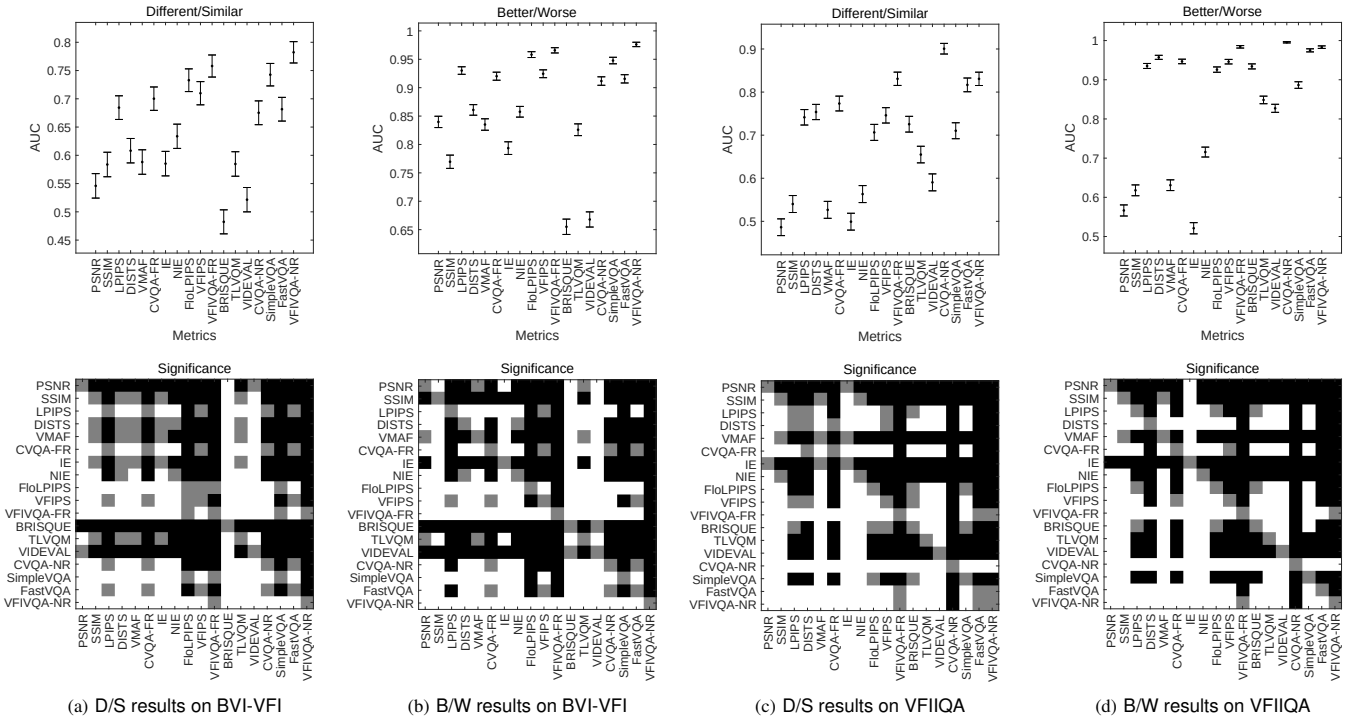
Additionally, it can be observed that for most methods, especially those designed for VFIVQA, there is a more noticeable decrease in correlation as the resolution increases to 2160p. The reason for this could be that many methods downsample the spatial resolution during the training process, which may hinder the effective extraction of features from high-resolution videos. However, it is worth noting that the proposed VFIVQA-NR method performs well on 2160p videos, indicating that the flow features are sensitive to the loss of high-resolution details in the VFI process and are effective in extracting features relevant to perceptual quality even for high-resolution videos.

2) Frame-rate Subsets: The BVI-VFI dataset is divided into three subsets based on video frame rates: 30fps, 60fps, and 120fps, with each subset containing 180 videos. The test results shown in the right part of Table V reveal that the proposed methods exhibit the best correlation for each frame rate subset, indicating good generalization across frame rate variations. It is worth noting that most methods perform better on the 60fps subset compared to other subsets. This suggests that the dataset contains more distinguishable features in this subset. Therefore, it may be necessary for VFI methods to focus on the quality of videos at the frame rate. In the 30fps subset, IQA-based methods typically show a larger discrepancy from VQA methods. However, as the frame rate increases, this gap gradually diminishes. Additionally, when the frame rate increases to 120fps, all methods show a significant decrease in SROCC, while PLCC remains relatively high. This decrease in SROCC may be due to the presence of subtle distortion variations in high frame rate videos, leading to fluctuations in objective predictions. This highlights the need for more robust quality assessment methods capable of handling such challenges.

D. Statistical Significance

The results of the ROC evaluation metrics mentioned in Section IV-A4 are presented and analyzed in this section. Due to the limitations of dataset labels, only datasets annotated with MOSs and DMOSs are used for analysis. Following the procedures given in [66], raw absolute ratings are preprocessed. Fig. 7 illustrates the Different/Similar and Better/Worse ROC analyses and their corresponding statistical results. In each analysis category, both FR and NR methods are compared to provide a more intuitive understanding of the results. The AUC results indicate that using the proposed framework on the BVI-VFI dataset, whether in FR or NR scenarios, outperforms other methods in accurately distinguishing between Different/Similar and Better/Worse video pairs. Notably, the proposed NR method exhibits particularly outstanding performance. From the statistical significance results, it is evident that both the VFIVQA-FR and VFIVQA-NR surpass other methods. Similar conclusions can be drawn from the AUC

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2025.3540104



(a) D/S results on BVI-VFI (b) B/W results on BVI-VFI (c) D/S results on VFIIQA (d) B/W results on VFIIQA

Fig. 7. Results of the Different/Similar (D/S) and Better/Worse (B/W) ROC analysis, and the corresponding statistical significance test results on the BVI-VFI and VFIIQA datasets. In the ROC analysis (top line), 95% confidence intervals of the AUC values are shown. In the statistical significance results (bottom line), a white/black block indicates that the row model is statistically better/worse than the column model. A gray block indicates that the row and column models are statistically indistinguishable.

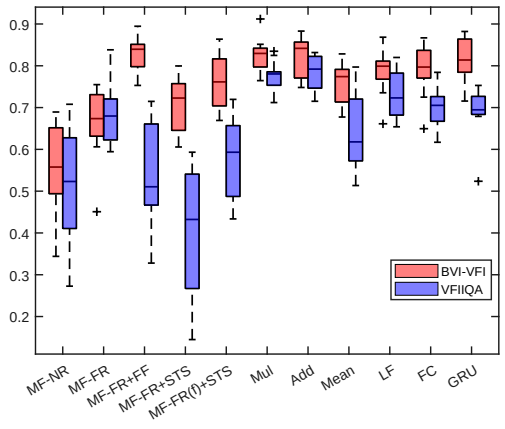


Fig. 8. SROCC results of different feature extraction blocks, feature fusion methods, and quality regressors. The box plots refer to ten results on the BVI-VFI and VFIIQA datasets.

TABLE VI
ABLATION EXPERIMENT RESULTS ON BVI-VFI AND VFIIQA DATASETS.

Methods	BVI-VFI			VFIIQA		
	SROCC	KROCC	PLCC	SROCC	KROCC	PLCC
MF-NR	0.560	0.396	0.634	0.506	0.359	0.547
MF-FR	0.662	0.481	0.730	0.696	0.506	0.710
MF-FR+FF	0.828	0.638	0.856	0.527	0.368	0.562
MF-FR+STS	0.709	0.515	0.686	0.412	0.289	0.499
MF-FR(f)+STS	0.761	0.565	0.754	0.584	0.421	0.658
Mul	0.824	0.639	**0.855**	0.774	0.583	0.784
Add	0.822	**0.640**	0.847	0.783	0.592	0.791
Mean	0.760	0.571	0.780	0.645	0.467	0.678
LF	0.786	0.600	0.818	0.733	0.548	0.750
FC	0.789	0.607	0.828	0.699	0.509	0.733
GRU	0.814	0.632	0.851	0.690	0.503	0.716
VFIVQA-NR	0.810	0.626	0.838	0.821	0.635	0.835
VFIVQA-FR	**0.828**	0.634	0.839	**0.823**	**0.638**	**0.833**

and statistical results on the VFIIQA dataset, where both FR and NR methods achieve a relatively high level of accuracy.

E. Ablation Study

In this section, ablation experiments are conducted to verify the effect of individual blocks at different stages in the FR and NR methods. Aligned with the phases of the proposed framework, three ablation experiments are designed, including the feature extraction stage, the feature fusion stage, and the quality regression stage.

In the feature extraction stage, the introduced flow feature (FF) block and motion feature (MF) block are the two pivotal components. First, as in [42], the motion feature extracted from the last layer of the backbone without reference videos is denoted as MF-NR. As proposed in VFIVQA-FR, the method of extracting MFs from both reference videos and VFI videos and concatenating them is referred to MF-FR. Based on MF-FR, incorporating FF block as a supplementary feature extraction method is represented as MF-FR+FF. In this method, the extraction of FF includes the computation

of CTS. On the other hand, combining the proposed STS calculation with the MF-FR is represented as MF-FR+STS. Furthermore, the MF block in VFIVQA-FR is frozen and denoted as MF-FR(f)+STS in the ablation experiments. For VFIVQA-NR, the feature extraction consists only of the FF block, while for VFIVQA-FR, the feature extraction strategy is MF-FR(f)+STS+FF.

For the feature fusion stage, the comparative strategies include element-wise multiplication (Mul), addition (Add), averaging (Mean), and the late fusion (LF) [44] of features. In this paper, the feature fusion method employed is concatenation within an early phase. Regarding the quality regression stage, the common practice involves regression using a fully connected (FC) layer. Additionally, the GRU [67] is utilized for regression modeling across the temporal dimension as a supplementary strategy for continuity modeling in VQA.

The results of these experiments are presented in Table VI, with more visually comparative results depicted in box plots in Fig. 8 for the SROCC performances. All ablation experiments are conducted on the same test set. Firstly, concerning feature extraction, MF-NR exhibits poor performance in both datasets, suggesting that independent motion features are insufficient to perceive the quality of VFI videos. MF-FR achieves a significant improvement in accuracy compared to MF-NR by introducing reference information, but it is not optimal. MF-FR+FF relies on flow features to effectively enhance performance on the BVI-VFI dataset, demonstrating the effectiveness of flow features in perceptual continuity perception. However, for VFIIQA, which focuses on single-frame VFI distortions, the accuracy is relatively poor. Comparison with VFIVQA-NR indicates that the inappropriate use of MF leads to a deterioration in perceptual accuracy for frames. On the other hand, the results of MF+STS and MF-FR(f)+STS prove that the calculation of STS with the MFs is beneficial for the VFIVQA. The freezing of model parameters ensures more stable feature learning. More importantly, incorporating FF into VFIVQA-FR achieves optimal performance for both entire videos and single frames.

The comparison of feature fusion strategies reveals that simple fusion methods as well as the LF methods can impact the representation of all features. The early fusion strategy in VFIVQA-FR integrates multiple features more effectively and performs better across different VFIVQA tasks. Finally, regarding quality regression, both FC and GRU strategies do not show significant improvement with the proposed features. Instead, the strategies lead to a notable decrease in accuracy for single-frame perceptual evaluation, indicating that the proposed quality regression network is more effective.

F. Cross-database Evaluation

The generalization is an essential ability of the VQA models. Several cross-database evaluation experiments are conducted by training both IQA/VQA models on the BVI-VFI and VFIIQA datasets and testing them on the remaining datasets. The experimental results are shown in Table VII. The VFIPS network trained on the BVI-VFI cannot be tested on the VFIIQA dataset due to frame limitation. When training on the VFIIQA dataset, the network is adjusted as described in Section IV-A1. From Table VII, it can be seen that the VFIVQA-FR model trained on BVI-VFI achieves the best performance in both FR and NR VQA models, demonstrating the generalization capability of the proposed framework. The performance of VFIVQA-NR is relatively low but still outperforms NR VQA methods. Although the results of VFIVQA-FR and VFIVQA-NR are not the best when trained on the VFIIQA dataset, they remain competitive. This may be due to the lack of temporal information in this dataset, which prevents the VQA models from temporal feature learning.

TABLE VII
CROSS-DATABASE VALIDATION. THE BEAT PERFORMANCES ARE IN BOLD.

Train	**BVI-VFI**			**VFIIQA**		
Test	VFIIQA		VFIPS	BVI-VFI		VFIPS
Method	SROCC	PLCC	2AFC	SROCC	PLCC	2AFC
LPIPS [23]	0.263	0.375	72.184	0.565	0.594	72.017
DISTS [61]	0.491	0.547	76.134	**0.609**	**0.618**	**75.378**
FloLPIPS [11]	0.471	0.570	71.092	0.553	0.575	72.268
VFIPS [10]	-	-	70.748	0.534	0.557	69.411
SimpleVQA [42]	0.463	0.402	68.933	0.524	0.557	67.772
FAST-VQA [43]	0.420	0.446	69.327	0.511	0.547	63.482
VFIVQA-FR	**0.675**	**0.671**	**77.815**	0.593	0.601	73.227
VFIVQA-NR	0.566	0.543	70.697	0.530	0.507	69.042

TABLE VIII
PARAMETERS AND RUNNING TIME OF THE PROPOSED METHODS AND COMPETING METHODS.

Method	**SROCC**	**Param(M)**	**Runtime(s)**
FloLPIPS [11]	0.733	2.47	0.0276
VFIPS [10]	0.743	4.60	0.0213
CVQA-FR [28]	0.708	24.51	0.0226
CVQA-NR [28]	0.608	30.75	0.0381
SimpleVQA [42]	0.750	24.72	0.0722
FAST-VQA [43]	0.744	28.13	0.2722
VFIVQA-FR	0.828	66.59	0.0433
VFIVQA-NR	0.810	33.30	0.0121

G. Computational Efficiency

The computational complexities of the proposed methods and competing VQA methods have been compared. Neural network-based methods have been selected for comparison, and the number of parameters and running time for each method are provided. Specifically, tests are conducted on 100 videos selected from the BVI-VFI dataset, with the average GPU running time reported in Table VIII. All the methods are tested on a computer with Intel Core i7-12700 CPU @2.10 GHz and NVIDIA GeForce RTX 3060 GPU. From Table VIII, it can be seen that the running times of the proposed methods are relatively short, particularly for VFIVQA-NR.

V. CONCLUSION

In this work, a novel objective VQA architecture for VFI videos is proposed, comprising three main components: feature extraction, feature fusion, and quality regression. The feature extraction extracts novel flow features from the triplet frames,

which are associated with temporal continuity, and motion features from the key frames in the video. Furthermore, two similarity computation methods are introduced in the respective feature extraction blocks to obtain effectively representative feature vectors. The architecture can be applied to both FR and NR VFIVQA. Both the flow feature and motion feature blocks are introduced in the FR method to obtain a more accurate result, while only the flow feature block is applied in the NR method for easier application. The VFIVQA-FR and VFIVQA-NR are compared with SOTA methods on three VFIVQA databases. The VFIVQA-FR achieves improvements of 11.44% and 4.18% in SROCC on the BVI-VFI and VFIIQA datasets, respectively. Moreover, the VFIVQA-NR is the first VQA method specifically designed for VFI videos, achieving SROCC improvements of 8.00% on BVI-VFI. Experimental results indicate that the proposed methods generalize well to videos with different spatial resolutions and frame rates. The results of ablation experiments indicate that the flow features and the corresponding similarity computations are crucial for the quality assessment of VFI.

REFERENCES

[1] B.-D. Choi, J.-W. Han, C.-S. Kim, and S.-J. Ko, "Motion-compensated frame interpolation using bilateral motion estimation and adaptive overlapped block motion compensation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 4, pp. 407–416, 2007.

[2] W. Bao, W.-S. Lai, C. Ma, X. Zhang, Z. Gao, and M.-H. Yang, "Depth-Aware video frame interpolation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3698–3707.

[3] M. Hu, J. Xiao, L. Liao, Z. Wang, C.-W. Lin, M. Wang, and S. Satoh, "Capturing small, fast-moving objects: Frame interpolation via recurrent motion enhancement," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 6, pp. 3390–3406, 2022.

[4] H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. Learned-Miller, and J. Kautz, "Super SloMo: High quality estimation of multiple intermediate frames for video interpolation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9000–9008.

[5] H. Men, V. Hosu, H. Lin, A. Bruhn, and D. Saupe, "Visual quality assessment for interpolated slow-motion videos based on a novel database," in *12th International Conference on Quality of Multimedia Experience(QoMEX)*. IEEE, 2020.

[6] G. Zhai and X. Min, "Perceptual image quality assessment: a survey," *Science China Information Sciences*, vol. 63, pp. 1–52, 2020.

[7] X. Min, K. Ma, K. Gu, G. Zhai, Z. Wang, and W. Lin, "Unified blind quality assessment of compressed natural, graphic, and screen content images," *IEEE Transactions on Image Processing*, vol. 26, no. 11, pp. 5462–5474, 2017.

[8] G. Wu, X. Tao, C. Li, W. Wang, X. Liu, and Q. Zheng, "Perception-oriented video frame interpolation via asymmetric blending," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2024, pp. 2753–2762.

[9] J. Han, X. Min, Y. Gao, J. Jia, L. Sun, Z. Cao, Y. Luo, and G. Zhai, "Perceptual quality assessment for video frame interpolation," in *IEEE International Conference on Visual Communications and Image Processing*, 2023, pp. 1–5.

[10] Q. Hou, A. Ghildyal, and F. Liu, "A perceptual quality metric for video frame interpolation," in *European Conference on Computer Vision*, 2022, pp. 234–253.

[11] D. Danier, F. Zhang, and D. Bull, "FloLPIPS: A bespoke video quality metric for frame interpolation," in *Picture Coding Symposium*, 2022, pp. 283–287.

[12] K.-C. Yang, A.-M. Huang, T. Q. Nguyen, C. C. Guest, and P. K. Das, "A new objective quality metric for frame interpolation used in video compression," *IEEE Transactions on Broadcasting*, vol. 54, no. 3, pp. 680–11, 2008.

[13] X. Min, H. Duan, W. Sun, Y. Zhu, and G. Zhai, "Perceptual video quality assessment: A survey," *Science China Information Sciences*, vol. 67, no. 11, p. 211301, 2024.

[14] H. Chen, F. Shao, X. Chai, Y. Gu, Q. Jiang, X. Meng, and Y.-S. Ho, "Quality evaluation of arbitrary style transfer: Subjective study and objective metric," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 7, pp. 3055–3070, 2023.

[15] L. Lin, Z. Wang, J. He, W. Chen, Y. Xu, and T. Zhao, "Deep quality assessment of compressed videos: A subjective and objective study," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 6, pp. 2616–2626, 2023.

[16] X. Min, G. Zhai, J. Zhou, M. C. Q. Farias, and A. C. Bovik, "Study of subjective and objective quality assessment of audio-visual signals," *IEEE Transactions on Image Processing*, vol. 29, pp. 6054–6068, 2020.

[17] M. Cheon and J.-S. Lee, "Subjective and objective quality assessment of compressed 4k uhd videos for immersive experience," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 7, pp. 1467–1480, 2018.

[18] D. Danier, F. Zhang, and D. Bull, "A subjective quality study for video frame interpolation," in *IEEE International Conference on Image Processing*, 2022, pp. 1361–1365.

[19] B. Li, W. Zhang, M. Tian, G. Zhai, and X. Wang, "Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 9, pp. 5944–5958, 2022.

[20] B. Chen, L. Zhu, G. Li, F. Lu, H. Fan, and S. Wang, "Learning generalized spatial-temporal deep feature representation for no-reference video quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 1903–1916, 2022.

[21] H. Zhu, B. Chen, L. Zhu, and S. Wang, "Learning spatiotemporal interactions for user-generated video quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 3, pp. 1031–1042, 2023.

[22] Y. Liu, J. Wu, L. Li, W. Dong, J. Zhang, and G. Shi, "Spatiotemporal representation learning for blind video quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 6, pp. 3500–3513, 2022.

[23] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[24] D. Danier, F. Zhang, and D. R. Bull, "BVI-VFI: A video quality database for video frame interpolation," *IEEE Transactions on Image Processing*, vol. 32, pp. 6004–6019, 2023.

[25] Q. Zheng, Y. Fan, L. Huang, T. Zhu, J. Liu, Z. Hao, S. Xing, C.-J. Chen, X. Min, A. C. Bovik, and Z. Tu, "Video quality assessment: A comprehensive survey," 2024. [Online]. Available: https://arxiv.org/abs/2412.04508

[26] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[27] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, M. Manohara *et al.*, "Toward a practical perceptual video quality metric," *The Netflix Tech Blog*.

[28] W. Sun, T. Wang, X. Min, F. Yi, and G. Zhai, "Deep learning based full-reference and no-reference quality assessment models for compressed ugc videos," in *IEEE International Conference on Multimedia and Expo Workshops*, 2021, pp. 1–6.

[29] M. Xu, J. Chen, H. Wang, S. Liu, G. Li, and Z. Bai, "C3DVQA: Full-reference video quality assessment with 3d convolutional neural network," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 4447–4451.

[30] T. Guan, C. Li, K. Gu, H. Liu, Y. Zheng, and X.-j. Wu, "Visibility and distortion measurement for no-reference dehazed image quality assessment via complex contourlet transform," *IEEE Transactions on Multimedia*, vol. 25, pp. 3934–3949, 2023.

[31] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[32] X. Min, Y. Gao, Y. Cao, G. Zhai, W. Zhang, H. Sun, and C. W. Chen, "Exploring rich subjective quality information for image quality assessment in the wild," *arXiv preprint arXiv:2409.05540*, 2024.

[33] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 36–47, 2020.

[34] X. Min, G. Zhai, K. Gu, X. Yang, and X. Guan, "Objective quality evaluation of dehazed images," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 8, pp. 2879–2892, 2019.

[35] H. Chen, F. Shao, B. Mu, and Q. Jiang, "Image aesthetics assessment with emotion-aware multibranch network," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–15, 2024.

[36] X. Min, G. Zhai, K. Gu, Y. Zhu, J. Zhou, G. Guo, X. Yang, X. Guan, and W. Zhang, "Quality evaluation of image dehazing methods using synthetic hazy images," *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2319–2333, 2019.

[37] H. Chen, X. Chai, F. Shao, X. Wang, Q. Jiang, X. Meng, and Y.-S. Ho, "Perceptual quality assessment of cartoon images," *IEEE Transactions on Multimedia*, vol. 25, pp. 140–153, 2023.

[38] X. Min, K. Gu, G. Zhai, J. Liu, X. Yang, and C. W. Chen, "Blind quality assessment based on pseudo-reference image," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2049–2062, 2018.

[39] J. Korhonen, "Two-level approach for no-reference consumer video quality assessment," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 5923–5938, 2019.

[40] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "UGC-VQA: Benchmarking blind video quality assessment for user generated content," *IEEE Transactions on Image Processing*, vol. 30, pp. 4449–4464, 2021.

[41] D. Li, T. Jiang, and M. Jiang, "Quality assessment of in-the-wild videos," in *ACM International Conference on Multimedia*. New York, NY, USA: Association for Computing Machinery, 2019, p. 2351–2359.

[42] W. Sun, X. Min, W. Lu, and G. Zhai, "A deep learning based no-reference quality assessment model for ugc videos," in *ACM International Conference on Multimedia*. New York, NY, USA: Association for Computing Machinery, 2022, p. 856–865.

[43] H. Wu, C. Chen, J. Hou, L. Liao, A. Wang, W. Sun, Q. Yan, and W. Lin, "FAST-VQA: Efficient end-to-end video quality assessment with fragment sampling," in *European Conference on Computer Vision*, 2022, pp. 538–554.

[44] Y. Cao, X. Min, W. Sun, and G. Zhai, "Attention-guided neural networks for full-reference and no-reference audio-visual quality assessment," *IEEE Transactions on Image Processing*, vol. 32, pp. 1882–1896, 2023.

[45] *Methodology for the subjective assessment of the quality of television pictures, International Telecommunication Union*, document Rec. ITU-R BT.500-13, 2012.

[46] Y. Fu, H. Liu, Y. Zou, S. Wang, Z. Li, and D. Zheng, "Category-level band learning based feature extraction for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–16, 2023.

[47] Y. Fu, T. Zhang, L. Wang, and H. Huang, "Coded hyperspectral image reconstruction using deep external and internal learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3404–3420, 2022.

[48] L. Chen, Y. Fu, L. Gu, C. Yan, T. Harada, and G. Huang, "Frequency-aware feature fusion for dense image prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10 763–10 780, 2024.

[49] Z. Lai, Y. Fu, and J. Zhang, "Hyperspectral image super resolution with real unaligned rgb guidance," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 1, no. 1, pp. 1–13, 2024.

[50] Z. Liu, R. A. Yeh, X. Tang, Y. Liu, and A. Agarwala, "Video frame synthesis using deep voxel flow," in *IEEE International Conference on Computer Vision*, Oct 2017.

[51] X. Xu, L. Siyao, W. Sun, Q. Yin, and M.-H. Yang, "Quadratic video interpolation," in *Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019.

[52] D. Danier, F. Zhang, and D. Bull, "ST-MFNet: A spatio-temporal multi-flow network for frame interpolation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3521–3531.

[53] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," *International Journal of Computer Vision*, vol. 92, pp. 1–31, 2011.

[54] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8934–8943.

[55] T. Guan, C. Li, Y. Zheng, X. Wu, and A. C. Bovik, "Dual-stream complex-valued convolutional network for authentic dehazed image quality assessment," *IEEE Transactions on Image Processing*, vol. 33, pp. 466–478, 2024.

[56] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *European Conference on Computer Vision*, 2020, pp. 402–419.

[57] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[58] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6201–6210.

[59] *Methodology for the subjective assessment of video quality in multimedia applications*, document Rec. ITU-R BT.1788, 2007.

[60] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[61] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2567–2581, 2022.

[62] S. Yang, T. Wu, S. Shi, S. Lao, Y. Gong, M. Cao, J. Wang, and Y. Yang, "MANIQA: Multi-dimension attention network for no-reference image quality assessment," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022, pp. 1190–1199.

[63] H. Wu, Z. Zhang, W. Zhang, C. Chen, L. Liao, C. Li, Y. Gao, A. Wang, E. Zhang, W. Sun, Q. Yan, X. Min, G. Zhai, and W. Lin, "Q-align: Teaching LMMs for visual scoring via discrete text-defined levels," in *Proceedings of the 41st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 235. PMLR, 21–27 Jul 2024, pp. 54 015–54 029.

[64] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1427–1441, 2010.

[65] J. Antkowiak, T. J. Baina, F. V. Baroncini, N. Chateau, F. FranceTelecom, A. C. F. Pessoa, F. S. Colonnese, I. L. Contin, J. Caviedes, and F. Philips, "Final report from the video quality experts group on the validation of objective models of video quality assessment," 2000.

[66] L. Krasula, P. Le Callet, K. Fliegel, and M. Klíma, "Quality assessment of sharpened images: Challenges, methodology, and objective metrics," *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1496–1508, 2017.

[67] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

Jinliang Han received the B.E. degree from the Xiamen University, Xiamen, China, in 2018. He is currently pursuing the Ph.D. degree with the Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai, China. His current research interest is in image/video quality assessment.

XiongKuo Min (Member, IEEE) received the B.E. degree from Wuhan University, Wuhan, China, in 2013, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2018, where he is currently a tenure-track Associate Professor with the Institute of Image Communication and Network Engineering. From Jun. 2018 to Sept. 2021, he was a Postdoc at Shanghai Jiao Tong University. From Jan. 2016 to Jan. 2017, he was a visiting student at University of Waterloo. From Jan. 2019 to Jan. 2021, he was a visiting scholar at The University of Texas at Austin and the University of Macau. He received the Best Paper Runner-up Award of IEEE Transactions on Multimedia in 2021, the Best Student Paper Award of IEEE International Conference on Multimedia and Expo (ICME) in 2016, the Best Paper Award of IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB) in 2022, and several first place awards of grand challenges held at IEEE ICME and ICIP. His research interests include image/video/audio quality, quality of experience, multimedia, image/video processing, computer vision.

Jun Jia received the B.S. degree in computer science and technology from Hunan University, Changsha, China, in 2018, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2024. He is currently a Post-Doctoral Fellow with Shanghai Jiao Tong University. His current research interests include computer vision and image processing.

Yixuan Gao received the B.E. degree from the Harbin Institute of Technology, Weihai, China, in 2020. She is currently pursuing the Ph.D. degree with the Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai, China. Her current research interests include image quality assessment.

Xiaohong Liu (Member, IEEE) received the B.E. degree in communication engineering from Southwest Jiaotong University, China, in 2014, the M.A.Sc. degree in electrical and computer engineering from the University of Ottawa, Canada, in 2016, and the Ph.D. degree in electrical and computer engineering from McMaster University, Canada, in 2021. Currently, he is a tenure-track Assistant Professor with the John Hopcroft Center, Shanghai Jiao Tong University. His research interests include computer vision and multimedia information processing. He received the Ontario Graduate Scholarship in 2019, the NSERC Alexander Graham Bell Canada Graduate Scholarship-Doctoral, and the Borealis AI Global Fellowship Award in 2020. He serves as a reviewer for several IEEE journals, including IEEE Transactions on Pattern Analysis And Machine Intelligence, IEEE Transactions on Image Processing, IEEE Transactions on Multimedia, IEEE Transactions on Circuits And Systems For Video Technology, and IEEE Transactions on Intelligent Transportation Systems.

Guangtao Zhai (Fellow, IEEE) received the B.E. and M.E. degrees from Shandong University, Shandong, China, in 2001 and 2004, respectively, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2009, where he is currently a Research Professor with the Institute of Image Communication and Information Processing. From 2008 to 2009, he was a Visiting Student with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON, Canada, where he was a Post-Doctoral Fellow from 2010 to 2012. From 2012 to 2013, he was a Humboldt Research Fellow with the Institute of Multimedia Communication and Signal Processing, Friedrich Alexander University of Erlangen-Nuremberg, Germany. He received the Award of National Excellent Ph.D. Thesis from the Ministry of Education of China in 2012. His research interests include multimedia signal processing and perceptual signal processing.