Video Quality Assessment Based on Swin Transformer with Spatio-Temporal Feature Fusion and Data Augmentation

Wei Wu[1], Shuming Hu[1], Pengxiang Xiao[1], Sibin Deng[1], Yilin Li[1*], Ying Chen[1], Kai Li[1]
[1]Department of Tao Technology, Alibaba Group
{guokui.ww, hushuming.hsm, xiaopengxiang.xpx, sibin.dsb}@alibaba-inc.com
{gustav.lyl, yingchen, kaishi.lk}@alibaba-inc.com

Abstract

While video enhancement has drawn significant interest and has been extensively studied by academia and industry, the corresponding research on video quality assessment (VQA) for enhanced video has not been widely addressed. Video enhancement methods normally change the relevant metrics like brightness, contrast, color, etc., leading to the fluctuation of perceptual quality and challenging the related VQA task. In this paper, we propose a novel approach for VQA task based on Swin Transformer with improved spatio-temporal feature fusion, which precisely mines the stage-wise feature concatenation and provides competitive assessment performance. In addition, we propose an efficient data augmentation strategy to improve data diversity and further enhance assessment accuracy. Experimental results demonstrate that the proposed approach achieves state-of-the-art performance on two benchmark VQA datasets, and ranks first in CVPR NTIRE 2023 Quality Assessment for Video Enhancement Challenge, which proves that the proposed approach is not only promising in VQA for enhanced video but also ubiquitous in general VQA tasks.

1. Introduction

The explosive growth of user-generated content (UGC), including live streaming and vlogs, has been witnessed by the world over the last decade. Unlike the pristine original version of the content provided by professional service providers, which rely on full-reference video quality assessment (FR-VQA) to achieve quality/bitrate tradeoff, UGC suffers from pre-existing distortions or compression artifacts [32], facing the assessment demands that FR-VQA are not coming close to meet.

Given this prevalence, understanding the perceptual subjective video quality of UGC is an imperative task for service providers. However, the biggest challenge in the quantitative assessment of UGC is its diversity including source video quality, ranging from 4K HDR to low-end shaky capturing, and processing, including crop, rescale, compression, etc. The combinations of these factors may significantly influence a viewer's expectation of video quality and their watching experience, which triggers the evolution in VQA for UGC — no-reference video quality assessment (NR-VQA) [4].

Classical NR-VQA methods employ handcrafted features to evaluate video quality. The underlying assumption of related studies is the observation that the variation of video quality can be comprehended with statistical characteristics, including pixel values of images/video [5, 30], optical flow [25], discrete cosine transformation coefficients [21], etc. However, these features are biased on content-related metrics and thus are less sensitive to subtle quality changes, while shallow feature aggregation does not help improve assessment accuracy but leads to extravagant computational complexity.

With these limitations, more attention is paid to learning-based features for NR-VQA. Driven by the remarkable performance delivered by convolutional neural networks (CNN) on a wide range of computer vision tasks, including image classification [12], detection [27], segmentation [11], etc., features extracted from pre-trained CNN networks for image quality assessment (IQA) tasks are exploited for NR-VQA in the context of insufficient labeled data. Representative works include V-CORNAIA [43], DeepBVQA [1], VSFA [20], and RIRNet [3]. The feature extractors behind, however, are not trained for NR-VQA and struggle to preserve spatio-temporal features that are crucial to videos [39]. To tackle with this issue, SimpleVQA [35] employs an image recognition based network to extract spatial features, which are further fused with the temporal features extracted by an action recognition-based network.

Recently, the success of the attention mechanism in natural language processing (NLP) tasks inspires researchers to integrate Transformers in vision tasks or employ it as a competitive alternative to CNN. Vision Transformer (ViT),

*Corresponding Author.

as a pure Transformer-based architecture, has outperformed its convolutional counterparts in many vision tasks [2,8,47]. Naturally, preliminary interest and discussions about employing ViT in NR-VQA have evolved into a full-fledged implementation, as addressed in some pioneering work like TRIQ [46], MUSIQ [16], where spatial and scale embedding mechanisms are utilized to help the Transformer capture features across spaces and scales.

In this work, we propose an improved NR-VQA model on top of SimpleVQA [35], which is composed of two key components: the spatial feature extraction module and the spatio-temporal feature fusion module. In the spatial feature extraction module, we employ Swin Transformer V2 [23] as the backbone of the spatial feature extraction network, as Swin Transformer V2 inherits the advantages of both CNN and ViT, which is an upgraded version as classical Swin Transformer [24]. In the spatio-temporal feature fusion module, we introduce a 1×1 convolutional layer, which deepens the spatial features extracted from the intermediate stages of the spatial feature extraction module to mitigate the gap between shallow and deep features. The spatial features from different stages are flattened and fused with the temporal features (originally from the motion feature extraction module in [35]) as the final features for video quality prediction. In addition, data augmentation strategies are performed in both spatial and temporal domains. Specifically, the input frames are resized and randomly cropped with a fixed resolution, and then randomly extracted from each video segment with a fixed sampling frequency to maintain temporal correlation.

The contributions of this paper are summarized as follows:

- We employ Swin Transformer V2 [23] as the backbone network to extract spatial features because due to its strong modeling capabilities and representative performance inherited from both CNN and ViT.
- We propose an efficient spatio-temporal feature fusion module that exploits features from different stages for better concatenation.
- We introduce data augmentation strategies in both spatial and temporal domains to improve the diversity of training samples.

The rest of this paper is organized as follows. In Section 2, we briefly review the existing NR-VQA metrics. The proposed method is detailed in Section 3, and experiments are presented in Section 4. Finally, Section 5 concludes this paper.

2. Related Work

2.1. Handcrafted Feature Based NR-VQA Metrics

Classical NR-VQA metrics exploit handcrafted features to evaluate video quality [37] [29] [36] [17]. Among these works, TLVQM [17] combines the spatial high-complexity and temporal low-complexity handcrafted features such as motion, jerkiness, blurriness, noise, etc. VIDEVAL [36] models diverse authentic distortions using different handcrafted features. However, video content also affects its quality, which cannot be well captured with these handcrafted features. Hence, some studies try to combine semantic features extracted by CNN with handcrafted features for NR-VQA task [37] [18]. CNN-TLVQM [18] combines the handcrafted features from TLVQM with spatial features extracted by a pre-trained CNN model. RAPIQUE [37] designs a model that can perceive video quality by statistical features and deep convolutional features.

2.2. Deep Learning Based NR-VQA Metrics

Deep learning based methods have recently drawn much attention for their superior performance. [22] proposes a video-based multi-task end-to-end optimized neural network (V-MEON) that can estimate video quality and classify the compression distortion. VSFA [20] first utilizes the semantic features extracted from a pre-trained CNN model and then uses a Gated Recurrent Unit (GRU) network to model the temporal memory effects. Further, the authors of VSFA propose MDVSFA, which is trained on multiple VQA datasets improving its performance. RIRNet [3] is proposed to fuse motion information extracted from different temporal frequencies. SIONR [41] is proposed to perceive video quality by considering the variations of semantic information, and the low-level features are combined to retain more detailed information about videos. Ying *et al.* [44] propose a local-to-global region-based method that combines the spatial and temporal features extracted by a 2D-CNN model and a 3D-CNN model, respectively. Wang *et al.* [38] propose a feature-rich VQA model for User Generated Content (UGC) videos. To achieve an accurate and reliable assessment of perceptual quality, it uses rich features that capture the quality information such as compress-based features, distortion-based features, and content-based features. Xu *et al.* [42] utilize the spatial features generated from a pre-trained IQA model and use the graph convolution to extract and enhance the features. After that, the motion features are extracted from the optical flow domain, and they finally used a bidirectional long short-term memory network to fuse the spatial and motion features.

Later, Transformer-based VQA methods have drawn more attention. LSCT [45] extracts features by a perceptual hierarchical network and then feeds the features into a long short-term convolutional Transformer to predict the video quality.

3. Proposed Method

The framework of the proposed model is depicted in Fig. 1, comprising the modules for spatial feature extraction, temporal feature extraction, and spatio-temporal feature fusion and regression. Specifically, quality-aware features are extracted from two aspects including the spatial and temporal aspects. Then the obtained multi-dimensional features are fused in spatio-temporal manners and mapped to quality scores via the quality regression module.

3.1. Feature Extraction

Given a video whose number of frames and frame rate is N and r, we split the video into $M = \frac{N}{r}$ video segments for feature extraction, and each segment lasts for 1 second. For each segment S_i (i represents the index of the segment), one frame is randomly sampled from each segment for spatial feature extraction while the whole segment is employed for temporal feature extraction.

3.1.1 Spatial Feature Extraction

According to Li *et al.* [20], the impact of distortions on human tolerance can vary based on the semantic content involved. For instance, humans are more likely to tolerate blur distortions on objects that lack texture or depth, such as clear skies and smooth walls. Conversely, objects with intricate textures, such as rough rocks and complex plants, may be considered unacceptable with similar distortions. Furthermore, researchers suggest that semantic information can play a vital role in identifying the extent and presence of perceived distortions [7].

Visual perception is a hierarchical process, in which input visual information is processed from low-level features to high-level features [40]. We use deep semantic information as a video quality representation by utilizing the features extracted from the last two Transformer blocks of Swin Transformer V2 [23]. Instead of using the pretrained model to extract the spatial features, we train an end-to-end spatial feature extraction network to learn quality-aware feature representation in the spatial domain, which allows us to fully utilize the various types of video content and distortion present in current VQA databases. Frame-level spatial feature is expressed as

$$SF_k^i = \mathrm{GAP}\left(L_1\left(F_k^i\right)\right) \oplus \mathrm{GAP}\left(\mathrm{Conv1}\left(L_2\left(F_k^i\right)\right)\right) \quad (1)$$

where SF_k^i indicates the extracted spatial features from the k-th sampled frame F_k^i of segment S_i, $\oplus$ stands for the concatenation operation, $\mathrm{GAP}(\cdot)$ represents the global average pooling operation, $L_j\left(F_k^i\right)$ stands for the feature maps obtained from j-th last transformer block of Swin Transformer V2, and Conv1 denotes 1×1 convolution operation.

3.1.2 Temporal Feature Extraction

Motion distortions caused by an unstable shooting environment often affect the quality of UGC videos. However, these distortions, including video shaking and motion blur, are not easily detected based solely on spatial features. To address this issue and improve the model's comprehension of temporal information, we utilize a pretrained 3D-CNN backbone called SlowFast [9] to capture segment-level temporal distortions:

$$TF^i = \Phi\left(S_i\right) \quad (2)$$

where TF^i indicates the extracted temporal features from the segment S_i, and $\Phi(\cdot)$ denotes the temporal feature extraction operation.

In summary, for the i-th segment S_i of the video, we can extract spatial features $SF^i \in \mathbb{R}^{M \times N_s}$ and temporal features $TF^i \in \mathbb{R}^{M \times N_t}$ at the segment-level. The number of channels for the spatial and temporal features are represented by N_s and N_t respectively.

3.2. Spatio-temporal Feature Fusion for Quality Prediction

Studies in neuroscience have revealed the presence of a hierarchical mechanism in visual perception [13, 28]. Based on this characteristic, we propose to integrate features from different levels. Rather than simply merging features from different layers, we introduce a 1×1 convolutional layer as shown in Fig. 2 to deepen the spatial features extracted from the intermediate stages of the pretrained network. This mitigates the gap between shallow and deep features. Depending only on spatial quality may not be enough as it overlooks the important temporal factors that play a crucial role in VQA. Several researchers have emphasized the importance of considering quality across the temporal axis [15, 31]. Therefore, it is logical for us to consider the contribution of both spatial and temporal factors in determining VQA.

After the concatenation of spatial and temporal features, the dimension of fused features is gradually reduced to 1 through FC1 and FC2, and the output dimension of FC1 is 64. After FC1, Rectified Linear Unit (ReLU) activation is employed, followed using the sigmoid function after FC2. For the segment S_i, we can obtain its segment-level quality score q_i via the quality regression module. Then, temporal average pooling is applied to obtain the video-level quality Q.

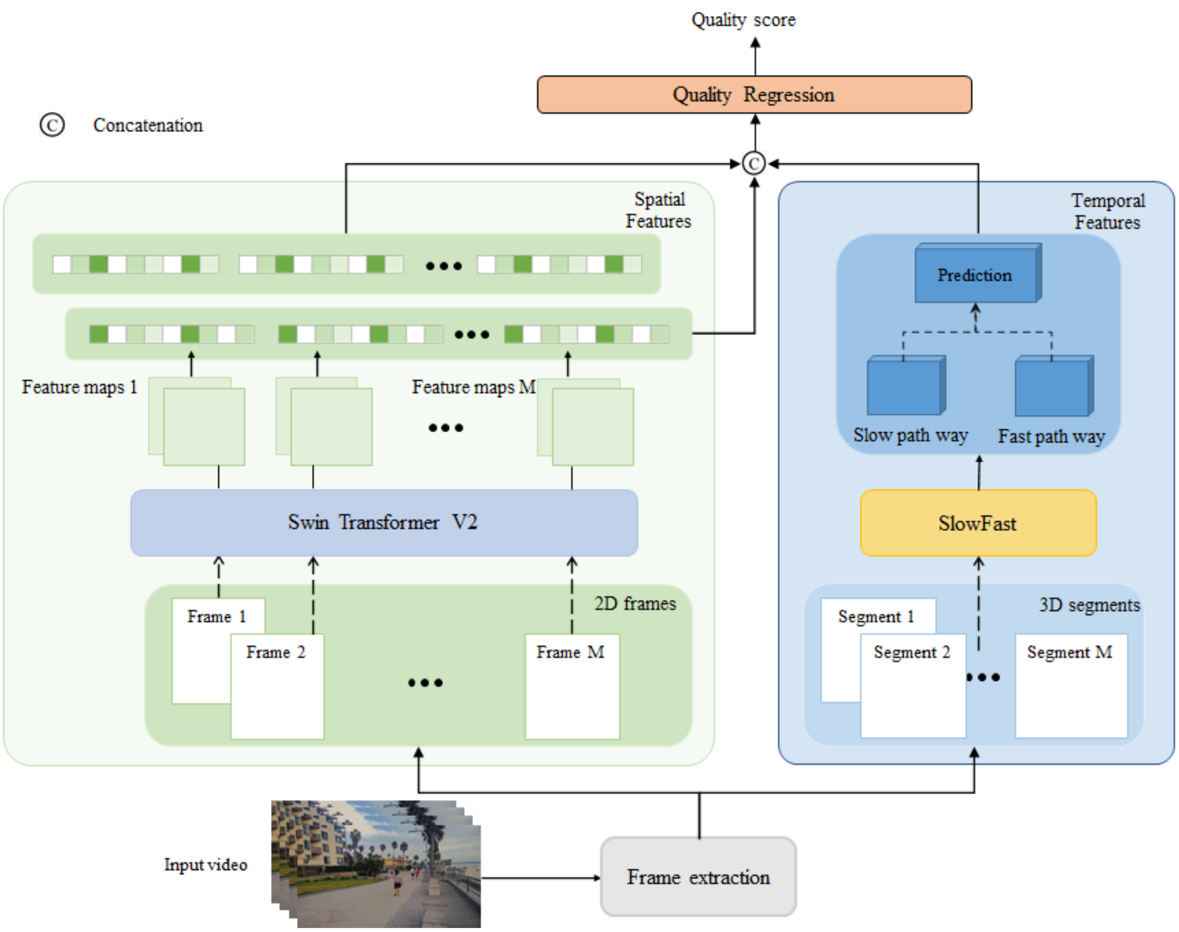

Figure 1. The framework of the proposed method, where the spatial and temporal features are extracted by Swin Transformer V2 [23]and pretrained SlowFast [9] respectively. Finally, spatial and temporal features are spatio-temporally fused and regressed into quality values.

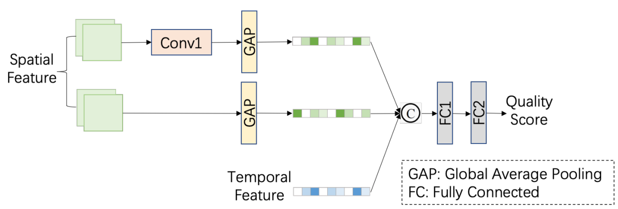

Figure 2. Spatio-temporal Feature Fusion and Regression Module. After a 1×1 convolution operation, the spatial features from the last two transformer blocks are combined. The resulting features are concatenated with temporal features and fed into fully connected layers to form a score.

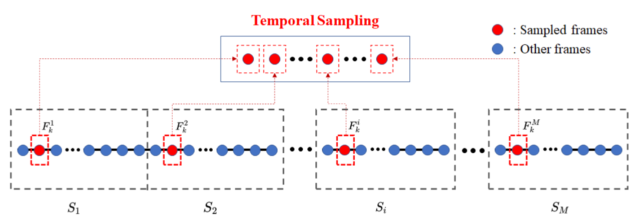

Figure 3. Temporal sampling at equal time intervals.

3.3. Data Augmentation

We leverage various data augmentation techniques, both spatially and temporally, to expand the number of videos in the training dataset and enhance the robustness of our model.

3.3.1 Training Data Augmentation

Data augmentation techniques are used in the spatial feature extraction. In spatial domain, each input frame is resized to 320×320 and randomly cropped a patch with a resolution of 256×256. In the temporal domain, the input video is divided into M segments. Then we randomly sample frame F_k^i from segment S_i, and constrain the position of sampled frames to align across segments as shown in Fig. 3. These tricks bring significantly improvements to our model performance.

3.3.2 Testing Data Augmentation

In the testing stage, the input frames are resized to 320 × 320, and the ”torchvision.transforms.TenCrop” function is used to crop 10 image patches with a resolution 256 × 256, which are located at the four corners and the center, respectively, as well as the horizontally flipped version of the previous crops. Moreover, we evenly sample 4 frames for each video segment in temporal domain.

4. Experiments

The comparison experiments are implemented to demonstrate the effectiveness of our VQA model. Two public datasets are used to train and test for evaluating the proposed model. Ablation studies are conducted to analyze the effectiveness of the proposed model. Through numerical and experimental verification, we demonstrate the effectiveness, performance, and advantages of our proposed method in this section.

4.1. Datasets and Evaluation Metrics

To evaluate the proposed method, we utilize two relevant NR-VQA databases: KoNViD-1k [14] and LIVE-VQC [34]. KoNViD-1k consists of 1200 public-domain video sequences while LIVE-VQC includes 585 videos.

Another video dataset is VDPVE [10], which is released by NTIRE 2023 Quality Assessment of Video Enhancement Challenge. Distortions of VDPVE videos are quite different from the aforementioned ones, which can be induced by various video enhancement algorithms.

Two commonly used evaluation metrics are used for performance comparison of different metrics: Spearman's Rank-order Correlation Coefficient (SROCC) and Pearson's Linear Correlation Coefficient (PLCC). SROCC represents the monotonic relationship between the predicted scores and the ground truths, which is computed as:

$$\mathrm{SROCC} = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N\left(N^2-1\right)} \tag{3}$$

where d_i is the distance between rank orders in predictions and the ground truths of the same video, N is number of videos. Slightly different from SROCC, PLCC measures prediction accuracy between predictions and ground truths. Before calculating the PLCC value, a four-parameter logistic regression function [33] is utilized to map the predicted scores to the scale of MOSs. The value range for SROCC and PLCC is $[0, 1]$ and better metrics should yield higher SROCC and PLCC values.

$$\mathrm{PLCC} = \frac{\sum_{i=1}^{N}\left(s_i-\bar{s}\right)\left(p_i-\bar{p}\right)}{\sqrt{\sum_{i=1}^{N}\left(s_i-\bar{s}\right)^2}\sqrt{\sum_{i=1}^{N}\left(p_i-\bar{p}\right)^2}} \tag{4}$$

where s_i and p_i are the subjective MOS and predictive score of each video respectively.

4.2. Implementation details

In the training stage, we used a batch size of 16 and employed the MSE loss as loss function. We employed the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, a weight decay of 10^{-7}. The learning rate is initialized as 10^{-5} and decayed by $\gamma = 0.95$ every 2 epochs. Language and other implementation details (including platform, memory, parallelization requirements) are shown as:

- Platform: PyTorch
- Language: Python 3.9
- Linux version 4.19.91-011.ali4000.alios7.x86_64
- CUDA Version 11.6
- Dependencies: PyTorch >= 1.13.1, NVIDIA GPU + CUDA
- GPU: 32G V100

4.3. Experimental Results

In order to conduct a comprehensive assessment of the proposed method's performance, we compare it with several popularly quality assessment models, namely BRISQUE [26], TLVQM [17], VIDEVAL [36], RAPIQUE [37], VSFA [20], PVQ [44], BVQA [19], and SimpleVQA [35]. It should be noted that BRISQUE [26] is categorized as a NR-IQA method, and we obtain the video quality features by taking the average of the features extracted from each frame using BRISQUE [26].

The experimental performances on the two UGC VQA databases are shown in Table1, from which we can draw several conclusions. Firstly, our proposed method achieves first place and outperforms the second place (SimpleVQA [35]) by approximately 0.0361, 0.0267 in terms of SROCC values on the KoNViD-1k [14] and LIVE-VQC [34] databases, respectively, thus demonstrating its effectiveness in predicting the quality scores of UGC videos. Secondly, except for the method VSFA [20], most of the deep learning-based methods (RAPIQUE [37], PVQ [44], BVQA [19], SimpleVQA [35] and the proposed method) significantly outperform handcraft-based methods (BRISQUE [26], TLVQM [17], VIDEVAL [36]). This can be attributed to the fact that handcrafted-based methods rely on prior experience of video distortions, which is based on

Table 1. Experimental performance comparison on KoNViD-1k [14] and LIVE-VQC [34]. ‘Hand’ denotes using handcrafted-based features while ‘Deep’ denotes using deep learning-based features.

Model	Hand	Deep	KoNViD-1k		LIVE-VQC	
			SROCC	PLCC	SROCC	PLCC
BRISQUE (TIP, 2012) [26]	✓		0.6567	0.6576	0.5925	0.6380
TLVQM (TIP, 2019) [17]	✓		0.7729	0.7688	0.7988	0.8025
VIDEVAL (TIP, 2021) [36]	✓		0.7832	0.7803	0.7522	0.7514
RAPIQUE (OJSP, 2021) [37]	✓	✓	0.8031	0.8175	0.7548	0.7863
VSFA (ACM MM, 2019) [20]		✓	0.7728	0.7754	0.6978	0.7426
PVQ (CVPR, 2021) [44]		✓	0.791	0.795	0.770	0.807
BVQA (TCSVT, 2022) [19]		✓	0.8362	0.8335	0.8412	0.8415
SimpleVQA (ACM MM, 2022) [35]		✓	0.850	0.860	0.845	0.859
Ours		✓	**0.8861**	**0.8931**	**0.8717**	**0.8830**

pristine videos, whereas the characteristics of UGC videos are far more complex and do not fit the regularities of artificial distortions.

4.4. Ablation Studies

In this section, we analyze the effectiveness of the proposed network by conducting ablation studies on the KoNViD-1k [14] and LIVE-VQC [34]. With different configuration and implementation strategies, we evaluate four major components: spatial feature extraction module, spatio-temporal fusion module, data augmentation and pre-training strategy. Table 2 shows the results of ablation studies. Model 1 (M1) only uses temporal features extracted by the pretrained SlowFast [9] for quality score regression. Model 2 (M2) only uses spatial features extracted by the transformer-based backbone Swin Transformer V2 [23] for quality score regression. Model 3 (M3) uses a transformer-based backbone Swin Transformer V2 [23] to replace the CNN-based backbone ResNet50 [12] of the SimpleVQA [35] model, which is equivalent to using both spatial and temporal features. In contrast to M3, Model 4 (M4) uses a 1×1 convolutional layer, which deepens the spatial features extracted from the intermediate stages of the pre-trained network, to mitigate the gap between shallow and deep features. Not only spatial data augmentation, Model 5 (M5) also considers temporal data augmentation. Compared to M5, Model 6 (M6) use the model pre-trained on LSVQ [44].

Effectiveness of Spatial Features (SF). Spatial feature is directly conscious of quality from the video frames. A comparison between M1 and M2 clearly indicates the critical role played by spatial features in the process of perceiving video quality. When comparing M1 with M3, it can be observed that fusing spatial and temporal features enables the model to perceive video quality more effectively.

Effectiveness of Temporal Features (TF). Comparing M2 with M3, in terms of the values of SROCC, M3 has achieved higher results on both datasets. This demonstrates that temporal features are capable of quantifying temporal distortions that are manifested in the motion of video frames and are often consistent within local regions of the frames. These distortions cannot be modeled by spatial features [35], which demonstrates that the introduction of temporal features effectively enhances the performance of the model.

Effectiveness of Swin Transformer V2 (Swin). We use Swin Transformer V2 [23] with swinv2-tiny-patch4-window8-256 weights as the backbone of the spatial feature extraction module. The weights of Swin Transformer V2 are initialized by the ImageNet-1K dataset [6]. Comparing the performance of M3 in Table 2 and the performance of SimpleVQA in Table 1, the SROCC value increases by 0.0147 on the KoNViD-1k database, but decreases by 0.0525 on the LIVE-VQC database. These results show that the CNN-based backbone network is easier to train on the small dataset LIVE-VQC (585) than the transformer-based backbone network. But on the larger dataset KoNViD-1k (1200), the transformer-based backbone network has more advantages.

Effectiveness of Convolution (Conv). In the spatio-temporal feature fusion module, we use a 1×1 convolutional layer, which deepens the spatial features extracted from the intermediate stages of the pre-trained network, to mitigate the gap between shallow and deep features. Comparing M4 with M5, in terms of the values of SROCC, M4 has achieved higher results on both KoNViD-1k and LIVE-VQC databases. These results suggest that 1×1 convolution operation before feature concatenation is effective.

Effectiveness of Data Augmentation (DA). In addition to the commonly used randomly crop to augment video data, we propose a new data augmentation method in the temporal domain. Our experimental results demonstrate the effectiveness of this temporal data enhancement, particularly for the small-scale LIVE-VQC database, where the

Table 2. Ablation studies on KoNViD-1k [14] and LIVE-VQC [34]

Model	TF	SF(Swin)	Conv	DA	Pre	KoNViD-1k		LIVE-VQC	
						SROCC	PLCC	SROCC	PLCC
M1	✓					0.6382	0.6752	0.6133	0.6473
M2		✓				0.8365	0.8500	0.7859	0.8070
M3	✓	✓				0.8647	0.8595	0.7925	0.8118
M4	✓	✓	✓			0.8679	0.8673	0.8025	0.8204
M5	✓	✓	✓	✓		0.8733	0.8810	0.8259	0.8220
M6	✓	✓	✓	✓	✓	**0.8861**	**0.8931**	**0.8717**	**0.8830**

Table 3. Quantitative results for the NTIRE 2023 Quality Assessment of Video Enhancement Challenge. This table only shows part of the participants and the best scores are bolded.

Team	Main Score	SROCC	PLCC
TB-VQA(ours)	**0.8576**	**0.8493**	**0.8659**
2nd	0.8396	0.8408	0.8383
3rd	0.8289	0.8261	0.8317
4th	0.8199	0.8163	0.8236
5th	0.7994	0.7962	0.8026
6th	0.7859	0.7896	0.7822
7th	0.7850	0.7879	0.7821
8th	0.7727	0.7756	0.7698

SROCC performance improvement reaches 0.024.

Effectiveness of Pre-traning (Pre). By pretraining with large VQA dataset LSVQ [44], we can learn quality-related features in an end-to-end manner, transfer them to specific VQA scenarios with small datasets, and improve their performance. The proposed method (M6) achieves the best performance with these video-quality-related features, which steadily improves model performance. These results suggest that pretraining strategy can serve as a solid backbone to enhance downstream tasks related to video quality.

4.5. NTIRE 2023 Quality Assessment of Video Enhancement Challenge

This work is proposed to participate in the NTIRE 2023 Quality Assessment of Video Enhancement Challenge, the objective of which is to propose an algorithm to estimate the quality of enhanced videos consistent with human perception. The final results of the challenge in the testing phase are shown in Tabel 3, our team (TB-VQA) won the first place in terms of PLCC, SROCC and main score.

5. Conclusion

In this paper, we propose a novel network based on Swin Transformer V2 with spatio-temporal feature fusion and data augmentation, for the quality assessment of video enhancement task. Specifically, we replace the CNN based backbone ResNet50 with a transformer-based backbone Swin Transformer V2. In addition, we propose a spatio-temporal feature fusion network that deepens the spatial feature extracted by the intermediate layer of the backbone network for better feature concatenation. Furthermore, a data augmentation strategy is applied in both spatial and temporal domain to improve data diversity. Experiments show that the proposed method outperforms the state-of-the-art methods on two standard VQA datasets. Additionally, we ranked first place on the NTIRE 2023 Quality Assessment of Video Enhancement Challenge.

References

[1] Sewoong Ahn and Sanghoon Lee. Deep blind video quality assessment based on temporal human perception. In *IEEE international Conference on Image Processing*, pages 619–623, 2018. 1

[2] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021. 2

[3] Pengfei Chen, Leida Li, Lei Ma, Jinjian Wu, and Guangming Shi. RIRNet: Recurrent-in-recurrent network for video quality assessment. In *Proceedings of the ACM International Conference on Multimedia*, pages 834–842, 2020. 1, 2

[4] Shyamprasad Chikkerur, Vijay Sundaram, Martin Reisslein, and Lina J Karam. Objective video quality assessment methods: A classification, review, and performance comparison. *IEEE Transactions on Broadcasting*, 57(2):165–182, 2011. 1

[5] Sathya Veera Reddy Dendi and Sumohana S Channappayya. No-reference video quality assessment using natural spatiotemporal scene statistics. *IEEE Transactions on Image Processing*, 29:5612–5624, 2020. 1

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 6

[7] Samuel Dodge and Lina Karam. Understanding how image quality affects deep neural networks. In *International Conference on Quality of Multimedia Experience*, pages 1–6, 2016. 3

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2

[9] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6202–6211, 2019. 3, 4, 6

[10] Yixuan Gao, Yuqin Cao, Tengchuan Kou, Wei Sun, Yunlong Dong, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. VDPVE: VQA dataset for perceptual video enhancement. *arXiv preprint arXiv:2303.09290*, 2023. 5

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 1, 6

[13] Shaul Hochstein and Merav Ahissar. View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, 36(5):791–804, 2002. 3

[14] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe. The konstanz natural video database (KoNViD-1k). In *International Conference on Quality of Multimedia Experience*, pages 1–6, 2017. 5, 6, 7

[15] Quan Huynh-Thu and Mohammed Ghanbari. Modelling of spatio–temporal interaction for video quality assessment. *Signal Processing: Image Communication*, 25(7):535–546, 2010. 3

[16] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5148–5157, 2021. 2

[17] Jari Korhonen. Two-level approach for no-reference consumer video quality assessment. *IEEE Transactions on Image Processing*, 28:5923–5938, 2019. 2, 5, 6

[18] Jari Korhonen, Yicheng Su, and Junyong You. Blind natural video quality prediction via statistical temporal features and deep spatial features. In *Proceedings of the ACM International Conference on Multimedia*, pages 3311–3319, 2020. 2

[19] Bowen Li, Weixia Zhang, Meng Tian, Guangtao Zhai, and Xianpei Wang. Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9):5944–5958, 2022. 5, 6

[20] Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *Proceedings of the ACM International Conference on Multimedia*, pages 2351–2359, 2019. 1, 2, 3, 5, 6

[21] Xuelong Li, Qun Guo, and Xiaoqiang Lu. Spatiotemporal statistics for video quality assessment. *IEEE Transactions on Image Processing*, 25(7):3329–3342, 2016. 1

[22] Wentao Liu, Zhengfang Duanmu, and Zhou Wang. End-to-end blind quality assessment of compressed videos using deep neural networks. In *Proceedings of the ACM International Conference on Multimedia*, pages 546–554, 2018. 2

[23] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12009–12019, 2022. 2, 3, 4, 6

[24] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2

[25] K Manasa and Sumohana S Channappayya. An optical flow-based no-reference video quality assessment algorithm. In *IEEE International Conference on Image Processing*, pages 2400–2404, 2016. 1

[26] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. 5, 6

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 2015. 1

[28] Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025, 1999. 3

[29] M. A. Saad, A. C. Bovik, and C. Charrier. Blind image quality assessment: A natural scene statistics approach in the dct domain. *IEEE Transactions on Image Processing*, 21:3339–3352, 2012. 2

[30] Michele A Saad, Alan C Bovik, and Christophe Charrier. Blind prediction of natural video quality. *IEEE Transactions on Image Processing*, 23(3):1352–1365, 2014. 1

[31] Kalpana Seshadrinathan and Alan Conrad Bovik. Motion tuned spatio-temporal quality assessment of natural videos. *IEEE Transactions on Image Processing*, 19(2):335–350, 2009. 3

[32] Hamid R Sheikh and Alan C Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2):430–444, 2006. 1

[33] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing*, 15(11):3440–3451, 2006. 5

[34] Zeina Sinno and Alan Conrad Bovik. Large-scale study of perceptual video quality. *IEEE Transactions on Image Processing*, 28(2):612–627, 2018. 5, 6, 7

[35] Wei Sun, Xiongkuo Min, Wei Lu, and Guangtao Zhai. A deep learning based no-reference quality assessment model for UGC videos. In *Proceedings of the ACM International Conference on Multimedia*, pages 856–865, 2022. 1, 2, 5, 6

[36] Zhengzhong Tu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C. Bovik. UGC-VQA: Benchmarking blind video

quality assessment for user generated content. *IEEE Transactions on Image Processing*, 30:4449–4464, 2021. 2, 5, 6

[37] Zhengzhong Tu, Xiangxu Yu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik. RAPIQUE: Rapid and accurate video quality prediction of user generated content. *IEEE Open Journal of Signal Processing*, 2:425–440, 2021. 2, 5, 6

[38] Yilin Wang, Junjie Ke, Hossein Talebi, Joong Gon Yim, Neil Birkbeck, Balu Adsumilli, Peyman Milanfar, and Feng Yang. Rich features for perceptual quality assessment of UGC videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13435–13444, 2021. 2

[39] Jinjian Wu, Yongxu Liu, Weisheng Dong, Guangming Shi, and Weisi Lin. Quality assessment for video with degradation along salient trajectories. *IEEE Transactions on Multimedia*, 21(11):2738–2749, 2019. 1

[40] Jinjian Wu, Jupo Ma, Fuhu Liang, Weisheng Dong, Guangming Shi, and Weisi Lin. End-to-end blind image quality prediction with cascaded deep neural network. *IEEE Transactions on Image Processing*, 29:7414–7426, 2020. 3

[41] Wei Wu, Qinyao Li, Zhenzhong Chen, and Shan Liu. Semantic information oriented no-reference video quality assessment. *IEEE Signal Processing Letters*, 28:204–208, 2021. 2

[42] Jiahua Xu, Jing Li, Xingguang Zhou, Wei Zhou, Baichao Wang, and Zhibo Chen. Perceptual quality assessment of internet videos. In *Proceedings of the ACM International Conference on Multimedia*, pages 1248–1257, 2021. 2

[43] Jingtao Xu, Peng Ye, Yong Liu, and David Doermann. No-reference video quality assessment via feature learning. In *IEEE International Conference on Image Processing*, pages 491–495, 2014. 1

[44] Zhenqiang Ying, Maniratnam Mandal, Deepti Ghadiyaram, and Alan Bovik. Patch-VQ:'patching up'the video quality problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14019–14029, 2021. 2, 5, 6, 7

[45] Junyong You. Long short-term convolutional transformer for no-reference video quality assessment. In *Proceedings of the ACM International Conference on Multimedia*, pages 2112–2120, 2021. 2

[46] Junyong You and Jari Korhonen. Transformer for image quality assessment. In *IEEE International Conference on Image Processing*, pages 1389–1393, 2021. 2

[47] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021. 2