On the Importance of Spatio-Temporal Learning for Video Quality Assessment

Dario Fontanel*
Politecnico di Torino, Italy
dario.fontanel@polito.it

David Higham
Amazon Prime Video, UK
highamdh@amazon.co.uk

Benoit Quentin Arthur Vallade
Amazon Prime Video, UK
valladeb@amazon.co.uk

Abstract

Video quality assessment (VQA) has sparked a lot of interest in the computer vision community, as it plays a critical role in services that provide customers with high quality video content. Due to the lack of high quality reference videos and the difficulties in collecting subjective evaluations, assessing video quality is a challenging and still unsolved problem. Moreover, most of the public research efforts focus only on user-generated content (UGC), making it unclear if reliable solutions can be adopted for assessing the quality of production-related videos. The goal of this work is to assess the importance of spatial and temporal learning for production-related VQA. In particular, it assesses state-of-the-art UGC video quality assessment perspectives on LIVE-APV dataset, demonstrating the importance of learning contextual characteristics from each video frame, as well as capturing temporal correlations between them.

1. Introduction

With the constant growth of services that offer video content as a form of entertainment for customers, video quality assessment (VQA) has attracted more and more interest recently [24, 22, 9, 26, 21, 5, 10]. Automated algorithms capable of predicting the quality of a never-before-seen video become then essential in services that aim at constantly improving their user experience. These methods are commonly referred to as No-Reference (NR) methods, since a high quality counterpart of the video is not available. However, most of state-of-the-art NR methods [8, 22, 21, 9, 26] mainly focus on user-generated content (UGC), due to the vast availability of public benchmarks. Moreover, only recently [22, 21, 9, 26, 14] deep learning models have been adopted in this field, due to their outstanding performances in various computer vision tasks [27, 12, 13, 2]. Starting from these premises, this work selects [9] among the NR leading approaches in MSU Video Quality Metrics Benchmark 2022 [1], to investigate the impact of spatio-temporal learning for accurate production-related video quality assessment. Our intuition is that production-related content (PRC) is perceived as low quality when customers identify unexpected spatial and temporal patterns in videos. Indeed, a scene represented with an unusual set of colours, a character behaving unexpectedly across time, defects introducing visual and temporal artifacts such as aliasing, video bars, and block corruption are just a few examples of unexpected patterns that contribute to a customer perception of low quality. To validate the aforementioned hypothesis, this work investigates the importance of spatial-temporal learning for PRC adopting LIVE-APV dataset [19, 18] as benchmark. Moreover, it highlights the diversity in the challenges between user-generated content and production-related content (see Figure 1) comparing respective state-of-the-art models.

To summarize, the contributions are as follows:

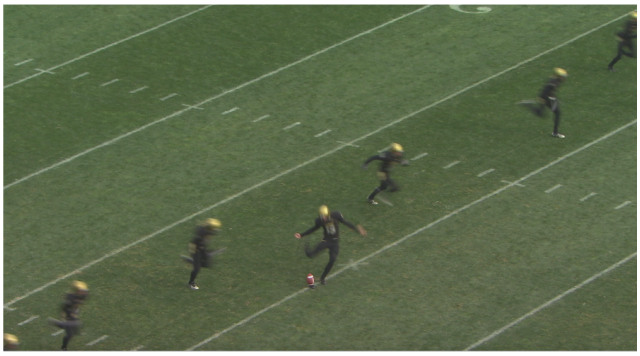
1. An investigation into the contribution of spatio-temporal learning for production-related video quality assessment. The investigation identified that both aspects are essential to properly learn models able to predict subjective quality of video content.
2. The assessment of VSFA [9] on the production-related LIVE-APV [19] benchmark. The comparison with ChipQA [5] highlights the diversity between UGC and PRC challenges.
3. An investigation into the impact of training frame rates in the perceptual quality that a model can learn. The study showed that a consistent frame rate is crucial to capture the pace of natural movements and the performance is positively correlated with the amount of frames analysed per second.

2. Related work

Historically, VQA approaches have focused on capturing natural video statistics [20, 16, 17] to address VQA

*Work done during an internship at Amazon Prime Video, UK.

(a) Low quality UGC video



(b) Low quality PRC video

(c) High quality UGC video

(d) High quality PRC video

Figure 1: Challenges in video quality assessment domain. UGC is more prone to be recorded under unusual fields of view and with lots of camera motion which might cut portions of the scene, while PRC recordings are usually more stable, and focused on the subjects of the scene. (a) shows a low perceived quality UGC video frame due to the recording setting and the presence of distractions (spectators, hand, hand rest); (b) shows a low quality PRC video frame characterized by a stable camera but with a poor motion-capturing system. (c) shows a high quality UGC, recorded from an unusual field of view; while (d) shows high quality PRC which focuses the field of view on the subject. UGC taken from [25], PRC taken from [19].

challenges. While other attempts have been made exploiting the structural information [23], motion [15], energy [11] and saliency [28] in videos, only recently pre-defined features have been replaced by deep learning approaches [7, 30, 29, 9].

Motivated by the intuition that PRC is perceived as low quality when customers identify unexpected spatial and temporal patterns in videos, this work starts from analyzing the state-of-the-art UGC approach [9], which captures spatio-temporal features adopting two separated modules. Firstly, it pre-trains an image classification model on ImageNet [3] to learn discriminatory boundaries on various contexts and images. Then, a Gated Recurrent Unit (GRU) module computes long-term dependencies on the extracted features and a final subjectively-inspired temporal pooling is applied. On the other hand, [5] takes a slightly different direction and introduces the concept of Space-Time Chips, which are localized and oriented portions of a video volume extracted along the direction of the motion. In pristine videos, they have been proved to follow certain regular statistics, which motivates their adoption in video quality assessment.

3. Method

In this section we first formalize the video quality assessment problem (Section 3.1), we then describe how [9] addresses the problem in Section 3.2 and then we analyse the spatio-temporal learning contributions in Section 3.3.

3.1. Problem formulation

The goal of video quality assessment is to correlate video content to the perceived quality by humans. At training time we are provided with a dataset $\mathcal{D} = \{(x, y)\}$ where $x \in \mathcal{X}$ is a video and $y \in \mathcal{Y}$ is its corresponding mean opinion score (MOS). MOSs represent the ground-truth perceived quality scores on a scale between 0 and 100, *i.e.*

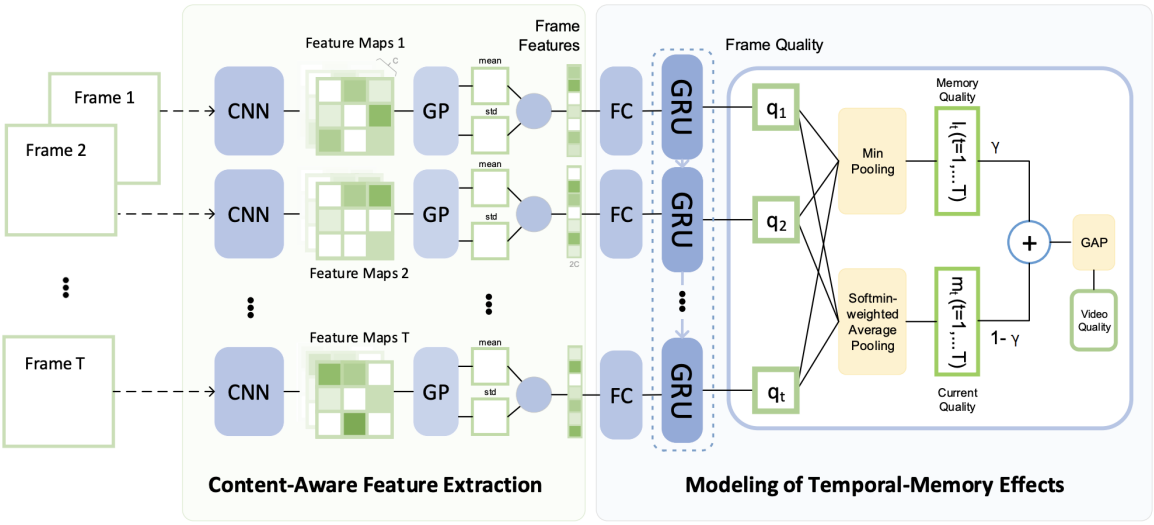


Figure 2: Illustration of VSFA. Visual content originally in [9].

$\mathcal{Y} = \{y \mid y \in \mathbb{R}, 0 \leq x \leq 100\}$. Given $\mathcal{D}$, VQA's goal is to learn a function $\mathcal{M}$ mapping a video to its corresponding MOS, *i.e.* $\mathcal{M} : \mathcal{X} \rightarrow \mathbb{R}^{|Y|}$. We consider $\mathcal{M}$ built on three components. The first is a feature extractor $\mathcal{M}_{FE}$ mapping video frames into a feature space. The second is a temporal learning function $\mathcal{M}_T$ mapping the spatial features learned by $\mathcal{M}_{FE}$ to temporal features. The third is a quality prediction function $\mathcal{M}_{VQA}$ which maps the temporal features to the final quality scores.

3.2. PRC Video Quality Assessment

In [9], $\mathcal{M}$ is built adopting i) a CNN module as $\mathcal{M}_{FE}$, ii) a GRU network as $\mathcal{M}_T$ and iii) a subjectively-inspired temporal pooling module as $\mathcal{M}_{VQA}$. $\mathcal{M}_{FE}$ exploits the capabilities of CNNs to capture contextual features. More precisely, each video frame of x is fed into a pre-trained CNN model which outputs the semantic feature maps from its deepest convolutional layers. To reduce the dimensionality of the extracted features, a spatial global pooling operation is then applied. As this operation discards spatial information, [9] also applies the spatial global standard deviation pooling operation to capture the variance. The results are concatenated, fed into a fully connected layer and then given as input to $\mathcal{M}_T$. The output of the GRU module is then fed into $\mathcal{M}_{VQA}$ to take into account the hysteresis effect (we advise referring to [9] for further explanation). An overview of the method is shown in Figure. 2.

3.3. Spatio-temporal learning

This section analyses the contribution of both spatial and temporal features learning for production-related VQA. Content-aware features play a key role in predicting the perceived quality of a video due to the high correlation between subjective human judgements and the content of the video itself. Towards this end, pre-trained image classification CNNs find applications in VQA as they are capable of extracting discriminative features based on the content of the video frames themselves. Also, as deep CNN features are distortion-sensitive [4] they can correlate the subjective quality with defects in videos as well. This work firstly evaluates the importance of contextual learning in VQA by investigating the contributions of the pre-trained image classification network. Section 4 details evaluations of multiple ResNet [6] instances while varying their depth.

Secondly, this work reports analysis on the temporal learning aspect. Capturing temporal features is essential to detect defects along temporal dimensions, as they will impact in the overall quality of videos. [9] adopts GRU network to catch the temporal relations between frames, and a subjectively-inspired temporal pooling layer to account the hysteresis effect. In the following, the importance of temporal learning is experimentally investigated by evaluating the contribution of the GRU network coupled with the subjectively-inspired temporal pooling (sub-GRU) against the LSTM network coupled with the same pooling operation (sub-LSTM) and with the GRU network coupled with an average pooling approach (avg-GRU). Finally, this work evaluates the performance of a model which does not take into account temporal aspects (no-time) to assess the quality of a video.

Table 1: Performance evaluation of temporal learning considering different types of temporal modules and pooling operations. Best results in bold.

Method	$\mathcal{M}_T$	$\mathcal{M}_{VQA}$	SROCC	KROCC	LCC
no-time	✗	✗	0.659 ± 0.049	0.482 ± 0.042	0.680 ± 0.071
avg-GRU	GRU	Average	0.746 ± 0.067	0.558 ± 0.066	0.769 ± 0.060
sub-LSTM	LSTM	Subjective	0.797 ± 0.052	0.610 ± 0.056	0.800 ± 0.056
sub-GRU	GRU	Subjective	$\mathbf{0.804 \pm 0.060}$	$\mathbf{0.619 \pm 0.066}$	$\mathbf{0.818 \pm 0.055}$

Table 2: Performance evaluation of spatial learning considering different instances of ResNet [6] architectures. Best results in bold.

	SROCC	KROCC	LCC
ResNet-18	0.759 ± 0.061	0.575 ± 0.068	0.776 ± 0.052
ResNet-50	$\mathbf{0.804 \pm 0.060}$	$\mathbf{0.619 \pm 0.066}$	$\mathbf{0.818 \pm 0.055}$
ResNet-101	0.801 ± 0.072	0.611 ± 0.074	0.810 ± 0.066

Table 3: Comparison between VSFA [9] and ChipQA [5] under SROCC. Best results in bold.

	SROCC
VSFA [9]	0.80 ± 0.06
ChipQA [5]	$\mathbf{0.83 \pm 0.04}$

4. Experiments

4.1. Evaluation protocol

This work evaluates the temporal and spatial contributions on LIVE-APV [19] dataset. It contains 45 high-resolution professional-grade pristine videos of production-related content and 6 different distortions synthetically applied to each of them, resulting in a total of 315 synthetically distorted videos. The applied distortions are compression, aliasing, interlacing, judder, flicker, and frame drop. Videos are then selected randomly to be part of the training, validation and test sets, making sure that each pristine video and its distorted versions fall into the same set. To fairly investigate the experiments, this work computes the mean and the standard deviation among five different runs. An extract of LIVE-APV dataset is shown in Figure 3.

Metrics. In LIVE-APV dataset subjective quality scores come in the form of mean opinion score (MOS). The standard performance criteria for evaluating VQA methods include the Spearman's rank-order correlation coefficient (SROCC), the Kendall's rank-order correlation coefficient (KROCC) and the Pearson's linear correlation coefficient (LCC). While SROCC and KROCC suggest the prediction monotonicity, LCC estimates the prediction accuracy.

4.2. Quantitative results

Table 1 reports the results assessing the temporal learning for production-related VQA. For each evaluation criteria, both the mean and the standard deviation are reported. sub-GRU, which couples the GRU network with the subjectively-inspired temporal pooling, achieves the best results, reaching up to 0.80 under SROCC, 0.62 under KROCC and 0.82 under LCC. Adopting a standard average pooling operation (avg-GRU) instead of the subjective temporal one is shown to be less effective, decreasing the performance under SROCC by 0.06. Moreover, adopting a LSTM module (sub-LSTM) is shown to achieve comparable results with sub-GRU under all the metrics, showing that beside the architectural choices, a module able to capture temporal dependencies is essential for assessing the quality of production-related videos. This is also supported by the first row of Table 1 which reports a model deprived of the temporal modules achieving 0.66 under SROCC. The overall results drop significantly under all metrics, showing that only the spatial learning is not sufficient for high video quality assessment performance.

Similarly, Table 2 reports the results assessing the spatial learning for production-related VQA. ResNet architectures are evaluated with various levels of depth. Among all, ResNet-50 performs best, reaching up to 0.80 under SROCC, 0.62 under KROCC and 0.82 under LCC. ResNet-18 is shown to be less effective, decreasing by 0.04 under SROCC. The reason is that the content-aware extracted features are less discriminative due to the lack of depth, and the temporal module struggles to learn dependencies. On the other hand, ResNet-101 achieves performance comparable to ResNet-50. This work attributes the lack of improvements to the fixed capacity of GRU which needs the input features to be re-projected into a lower dimensional

Figure 3: Pristine and defective samples taken from LIVE-APV [19] dataset. From top-left to bottom-right: pristine, pristine, defective (compressions), defective (interlacing). Best viewed in color.

features space.

Finally, Table 3 compares UGC and PRC state-of-the-art approaches under SROCC metric. Despite being created for UGC only, [9] achieves almost comparable results with [5], highlighting the early-staged yet promising approach of spatio-temporal learning for production-related content. This finding motivates us to further investigate the immense possibilities of adopting an end-to-end deep learning model for production-related video quality assessment.

4.2.1 On the importance of training frame rates

We begin the investigation into the impact of training frame rates on the perceptual PRC quality that a model can learn by mentioning that in the original LIVE-APV dataset the frame rates are not consistent across videos, ranging from 25 to 30 fps. In Figure. 4, we compare the original results to those obtained by feeding VSFA with LIVE-APV videos sampled at lower and fixed frame rates.

Adopting 25 and 20 fps surpasses the performance achieved using the original frame rates, reaching up to 0.810 SROCC, 0.820 LCC and 0.626 KROCC in the former scenario, and 0.806 SROCC, 0.820 LCC and 0.622 KROCC in the latter. We attribute the cause to the model learning the expected pace of moving subjects. As an explanatory

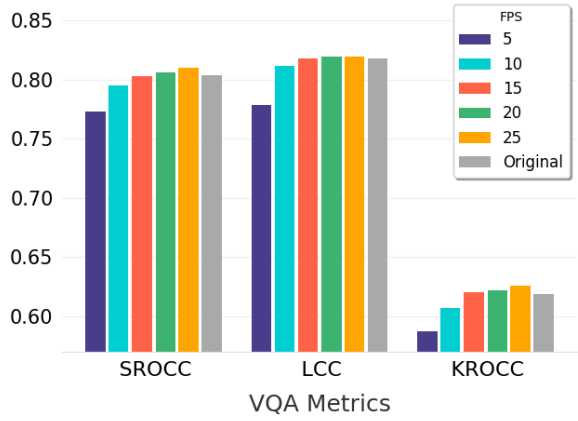


Figure 4: Impact of training frame rates for perceptual PRC quality prediction.

example, let us take the instance of humans jogging. Setting the training frame rate to a constant value, teaches the model a consistent pace for human runs. On contrary, if the model learns the pattern of a human run at - for example - 25 fps and then observes a person jogging at 30 fps, it will interpret the person's slower movement as a drop in quality

caused by the unusual temporal pattern.

However, downsampling considerably the training frames will have a negative impact on the model. In fact, with respect to the results achieved using 25 fps, the model's performance decreases by 0.015 SROCC, 0.019 KROCC and 0.008 LCC when adopting 10 fps, and respectively by 0.037, 0.039 and 0.041 when adopting 5 fps only.

To achieve high quality assessment performance on PRC video, we therefore find it necessary to i) analyse videos at a consistent frame rate, which allow the model to capture the pace of natural movements and ii) adopt the maximum frame rate possible, to prevent misinterpreting abrupt transition between two consecutive frames for an uncommon temporal pattern, which leads to low perceived quality.

5. Conclusions and future work

This paper investigates the importance of spatio-temporal learning for production-related video quality assessment. It assesses state-of-the-art UGC video quality assessment perspectives on LIVE-APV benchmark and demonstrate the contribution of context-aware features is essential, due to the high correlation between the content of a video and its perceived subjective quality. Moreover, this work shows that capturing temporal correlations becomes fundamental for production-related content, as the overall performance drops drastically when a model is deprived of this capability.

However, despite being created for UGC, this work finds the approach of spatio-temporal learning promising. The deep end-to-end training, indeed, opens up unexplored future directions. In particular, we are interested in exploiting unlabeled production-related datasets for additional supervision in a self- and semi- supervised manner. In the first approach, a contrastive learning strategy is planned to be adopted to learn a richer and sharper feature space. Inspired by [14], our intuition is that specific augmentation strategies such as flipping do not interfere with the perceived subjective quality of a video. On the other hand, the second approach adopts a two-stage iterative pipeline. We aim at using the main model to produce pseudo-labels for the large unlabeled dataset, and then use the pseudo-annotated samples to re-train the main model. This pattern is planned to be iterated until the extra pseudo-supervision stops providing additional improvements.

References

[1] MSU Video Quality Metrics Benchmark 2022. `https://videoprocessing.ai/benchmarks/video-quality-metrics_nrm.html`. Accessed: 2022-07-15.

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[4] Samuel Dodge and Lina Karam. Understanding how image quality affects deep neural networks. In *2016 eighth international conference on quality of multimedia experience (QoMEX)*, pages 1–6. IEEE, 2016.

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

[6] 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.

[7] Woojae Kim, Jongyoo Kim, Sewoong Ahn, Jinwoo Kim, and Sanghoon Lee. Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 219–234, 2018.

[8] Jari Korhonen. Two-level approach for no-reference consumer video quality assessment. *IEEE Transactions on Image Processing*, 28(12):5923–5938, 2019.

[9] Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2351–2359, 2019.

[10] Dingquan Li, Tingting Jiang, and Ming Jiang. Unified quality assessment of in-the-wild videos with mixed datasets training. *International Journal of Computer Vision*, 129(4):1238–1257, 2021.

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

[12] 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.

[13] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

[14] Pavan C Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C Bovik. Conviqt: Contrastive video quality estimator. *arXiv preprint arXiv:2206.14713*, 2022.

[15] K Manasa and Sumohana S Channappayya. An optical flow-based full reference video quality assessment algorithm. *IEEE Transactions on Image Processing*, 25(6):2480–2492, 2016.

[16] Anish Mittal, Michele A Saad, and Alan C Bovik. A completely blind video integrity oracle. *IEEE Transactions on Image Processing*, 25(1):289–300, 2015.

[17] 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.

[18] Zaixi Shang, Joshua Ebenezer, Yongjun Wu, Hai Wei, Sriram Sethuraman, and Alan Bovik. Study of the subjective and objective quality of high motion live streaming videos. *IEEE Transactions on Image Processing*, PP:1–1, 12 2021.

[19] Zaixi Shang, Joshua P Ebenezer, Alan C Bovik, Yongjun Wu, Hai Wei, and Sriram Sethuraman. Assessment of subjective and objective quality of live streaming sports videos. In *2021 Picture Coding Symposium (PCS)*, pages 1–5. IEEE, 2021.

[20] Zeina Sinno and Alan C Bovik. Spatio-temporal measures of naturalness. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1750–1754. IEEE, 2019.

[21] 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.

[22] 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.

[23] Yue Wang, Tingting Jiang, Siwei Ma, and Wen Gao. Novel spatio-temporal structural information based video quality metric. *IEEE transactions on circuits and systems for video technology*, 22(7):989–998, 2012.

[24] Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling. *arXiv preprint arXiv:2207.02595*, 2022.

[25] Joong Gon Yim, Yilin Wang, Neil Birkbeck, and Balu Adsumilli. Subjective quality assessment for youtube ugc dataset. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 131–135. IEEE, 2020.

[26] 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.

[27] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

[28] Wei Zhang and Hantao Liu. Study of saliency in objective video quality assessment. *IEEE Transactions on Image Processing*, 26(3):1275–1288, 2017.

[29] Yu Zhang, Xinbo Gao, Lihuo He, Wen Lu, and Ran He. Blind video quality assessment with weakly supervised learning and resampling strategy. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(8):2244–2255, 2018.

[30] Yu Zhang, Xinbo Gao, Lihuo He, Wen Lu, and Ran He. Objective video quality assessment combining transfer learning with cnn. *IEEE transactions on neural networks and learning systems*, 31(8):2716–2730, 2019.